Contents

Preface ix

PART ONE: An Introduction to Criminal Justice Inquiry  1

Chapter 1: Crime, Criminal Justice, and Scientific Inquiry  2

Introduction 3

Home Detention 4

What Is This Book About? 4

Two Realities 4

The Role of Science 6

Personal Human Inquiry 7

Tradition 7

Arrest and Domestic Violence 8

Authority 8

Errors in Personal Human Inquiry 9

Inaccurate Observation 9

Overgeneralization 10

Selective Observation 10

Illogical Reasoning 11

Ideology and Politics 11

To Err Is Human 11

Foundations of Social Science 12

Theory, Not Philosophy or Belief 12

Regularities 14

What About Exceptions? 14

Aggregates, Not Individuals 14

A Variable Language 15

Variables and Attributes 15

Variables and Relationships 18

Purposes of Research 19

Exploration 19

Description 20

Explanation 20

Application 20

Differing Avenues for Inquiry 21

Idiographic and Nomothetic Explanations 21

Inductive and Deductive Reasoning 23

Quantitative and Qualitative Data 23

Putting It All Together: “Driving While Black” 24

Knowing Through Experience: Summing Up and Looking Ahead 27

Main Points 27

Chapter 2: Theory and Criminal Justice Research  30

Introduction 31

The Creation of Social Science Theory 31

Terms Used in Theory Construction 31

The Traditional Model of Science 33

Two Logical Systems 37

A Deductive Illustration 37

An Inductive Illustration 39

Grounded Theory and Community Prosecution 40

Inductive Theory and Criminal Justice Policy 42

Putting It All Together: Theories of Discretion and Discrimination 44

Theory, Research, and Public Policy 46

Criminological Evolution: Ecological Theory and Crime Prevention Policy 47

Main Points 48

Chapter 3: Ethics and Criminal Justice Research  51

Introduction 52

Ethical Issues in Criminal Justice Research 52

No Harm to Participants 53

Ethics and Extreme Field Research 54

Voluntary Participation 57

Anonymity and Confidentiality 57

Deceiving Subjects 59

Analysis and Reporting 59

Legal Liability 60

Special Problems 61

Promoting Compliance with Ethical Principles 63

Putting It All Together: Ethics and Racial Profiling Research 64

Codes of Professional Ethics 64

Institutional Review Boards 66
Institutional Review Board Requirements and Researcher Rights  69

ETHICS AND JUVENILE GANG MEMBERS  70

Two Ethical Controversies  72
Trouble in the Tearoom  72
The Stanford Prison Experiment  73

Discussion Examples  75
Main Points  76

PART TWO: Structuring Criminal Justice Inquiry  79

Chapter 4: General Issues in Research Design  80

Introduction  81
Causation in the Social Sciences  82
Criteria for Causality  82
Necessary and Sufficient Causes  83

Validity and Causal Inference  84
Statistical Conclusion Validity  85
Internal Validity  86
External Validity  86
Construct Validity  86

Validity and Causal Inference Summarized  88
Does Drug Use Cause Crime?  88
Introducing Scientific Realism  89

Units of Analysis  90
Individuals  91
Groups  91
Organizations  92
Social Artifacts  92

Units of Analysis in Review  93
The Ecological Fallacy  94
Reductionism  95

The Time Dimension  95

UNITS OF ANALYSIS IN THE NATIONAL YOUTH GANG SURVEY  96

Cross-Sectional Studies  96
Longitudinal Studies  97
Approximating Longitudinal Studies  98

Retrospective Studies  99
The Time Dimension Summarized  101

PUTTING IT ALL TOGETHER: CAUSE, UNITS, AND TIME  102

How to Design a Research Project  104
The Research Process  104
Getting Started  106
Conceptualization  107
Choice of Research Method  107
Operationalization  108
Population and Sampling  108
Observations  108
Data Processing and Analysis  108
Application  109
Research Design in Review  109

The Research Proposal  110
Elements of a Research Proposal  110

Answers to the Units-of-Analysis Exercise  112
Main Points  112

Chapter 5: Concepts, Operationalization, and Measurement  115

Introduction  116
Conceptions and Concepts  116
Conceptualization  118
Indicators and Dimensions  118
Confusion over Definitions and Reality  119
Creating Conceptual Order  119

WHAT IS RECIDIVISM?  122

Operationalization Choices  122

JAIL STAY  124

Measurement as “Scoring”  124
Exhaustive and Exclusive Measurement  126
Levels of Measurement  127
Implications of Levels of Measurement  128

Criteria for Measurement Quality  130
Reliability  130
Validity  133

PUTTING IT ALL TOGETHER: MEASURING RACE  136
Chapter 6: Measuring Crime 144

Introduction 145
General Issues in Measuring Crime 145
What Offenses? 145
What Units of Analysis? 146
What Purpose? 147

Measuring Crime Summary 170
Main Points 172

Chapter 7: Experimental and Quasi-Experimental Designs 176

Introduction 177
The Classical Experiment 178
Independent and Dependent Variables 178
Pretesting and Posttesting 179
Experimental and Control Groups 179
Double-Blind Experiments 181
Selecting Subjects 182
Randomization 182

Experiments and Causal Inference 183
Experiments and Threats to Validity 183
Threats to Internal Validity 184
Ruling Out Threats to Internal Validity 187
Generalizability and Threats to Validity 189
Threats to Construct Validity 189
Threats to External Validity 190
Threats to Statistical Conclusion Validity 191

Variations in the Classical Experimental Design 191

Quasi-Experimental Designs 193
Nonequivalent-Groups Designs 193
Cohort Designs 196
Time-Series Designs 197
Variations in Time-Series Designs 201
Variable-Oriented Research and Scientific Realism 202

Putting It All Together: Evaluating Explanations for Disproportionality in Traffic Stops 204

Experimental and Quasi-Experimental Designs Summarized 206
Main Points 207

PART THREE: Modes of Observation 211

Chapter 8: Sampling 212

Introduction 213
The Logic of Probability Sampling 213
Conscious and Unconscious Sampling Bias 215
Representativeness and Probability of Selection 216
Probability Theory and Sampling Distribution 217
The Sampling Distribution of 10 Cases 218
From Sampling Distribution to Parameter Estimate 221
Estimating Sampling Error 222
Confidence Levels and Confidence Intervals 223
Probability Theory and Sampling Distribution Summed Up 224
Probability Sampling 225
Populations and Sampling Frames 225
Simple Random Sampling 226
Systematic Sampling 227
Stratified Sampling 227
Disproportionate Stratified Sampling 228
Multistage Cluster Sampling 229
Multistage Cluster Sampling with Stratification 230
Illustration: Two National Crime Surveys 232
The National Crime Victimization Survey 232
The British Crime Survey 233
Probability Sampling in Review 234
Nonprobability Sampling 234
Purposive or Judgmental Sampling 235
Quota Sampling 236
Reliance on Available Subjects 236
PUTTING IT ALL TOGETHER: SAMPLING DRIVERS, PLACES, AND "KNOWN SPEEDERS" 238
Snowball Sampling 238
Nonprobability Sampling in Review 240
Main Points 240

Chapter 9: Survey Research and Other Ways of Asking Questions 244
Introduction 245
Topics Appropriate to Survey Research 246
Counting Crime 246
Self-Reports 246
Perceptions and Attitudes 247
Targeted Victim Surveys 247
Other Evaluation Uses 248
General-Purpose Crime Surveys 248
Guidelines for Asking Questions 248
Open-Ended and Closed-Ended Questions 249
Questions and Statements 249
Make Items Clear 249
Short Items Are Best 250
Avoid Negative Items 250
Biased Items and Terms 250
Designing Self-Report Items 251
Questionnaire Construction 253
General Questionnaire Format 253
Contingency Questions 253
Matrix Questions 254
Ordering Questions in a Questionnaire 256

DON’T START FROM SCRATCH! 257
Self-Administered Questionnaires 257
Mail Distribution and Return 258
Warning Mailings and Cover Letters 258
Follow-Up Mailings 259
Acceptable Response Rates 259
Computer-Based Self-Administration 260
In-Person Interview Surveys 262
The Role of the Interviewer 262
General Rules for Interviewing 262
Coordination and Control 263
Computer-Assisted In-Person Interviews 264
Telephone Surveys 266
Computer-Assisted Telephone Interviewing 268
Comparison of the Three Methods 269
Strengths and Weaknesses of Survey Research 270
Other Ways of Asking Questions 272
Specialized Interviewing 272
Focus Groups 273
Should You Do It Yourself? 275
PUTTING IT ALL TOGETHER: ASKING QUESTIONS 276
Main Points 279
Chapter 10: Field Research  282

Introduction  283
Topics Appropriate to Field Research  284
The Various Roles of the Observer  286
Asking Questions  288
Gaining Access to Subjects  290
  Gaining Access to Formal Organizations  290
  Gaining Access to Subcultures  293
  Selecting Cases for Observation  294
  Purposive Sampling in Field Research  295
Recording Observations  297
  Cameras and Sound Recorders  298
  Field Notes  299
  Structured Observations  300
Linking Field Observations and Other Data  302
Illustrations of Field Research  304
  Shoplifting  305
  How Many People Wear Seat Belts?  307
  Bars and Violence  309
Strengths and Weaknesses of Field Research  310
  Validity  311
  Reliability  312
  Generalizability  313
Conducting a Safety Audit  304

Chapter 11: Agency Records, Content Analysis, and Secondary Data  320

Introduction  321
Topics Appropriate for Agency Records and Content Analysis  322
Types of Agency Records  323
  Published Statistics  323
  Nonpublic Agency Records  326
  New Data Collected by Agency Staff  330
Improving Police Records of Domestic Violence  332
Units of Analysis and Sampling  332
  Units of Analysis  333
  Sampling  334
Reliability and Validity  335
  Sources of Reliability and Validity Problems  336
How Many Parole Violators Were There Last Month?  338
Content Analysis  340
  Units of Analysis and Sampling in Content Analysis  341
  Coding in Content Analysis  343
  Illustrations of Content Analysis  344
Secondary Analysis  347
Putting It All Together: Agency Records and Content Analysis  348
  Sources of Secondary Data  350
  Advantages and Disadvantages of Secondary Data  351
Main Points  352

PART FOUR: Application and Analysis  355

Chapter 12: Evaluation Research and Problem Analysis  356

Introduction  357
Topics Appropriate for Evaluation Research and Problem Analysis  358
The Policy Process  358
  Linking the Process to Evaluation  360
Getting Started  362
  Evaluability Assessment  363
  Problem Formulation  364
  Measurement  366
Designs for Program Evaluation  370
  Randomized Evaluation Designs  370
  Home Detention: Two Randomized Studies  373
  Quasi-Experimental Designs  375
  Other Types of Evaluation Studies  379
Problem Analysis and Scientific Realism  379
  Problem-Oriented Policing  380
  Auto Theft in Chula Vista  381
  Other Applications of Problem Analysis  382
  Scientific Realism and Applied Research  384
Putting It All Together: Applied Research on Racial Profiling  386
The Political Context of Applied Research 388
Evaluation and Stakeholders 389
Politics and Objectivity 390
WHEN POLITICS ACCOMMODATES FACTS 391
Main Points 392

Chapter 13: Interpreting Data 395
Introduction 396
Univariate Description 396
Distributions 396
Measures of Central Tendency 397
Measures of Dispersion 398
Comparing Measures of Dispersion and Central Tendency 401
Computing Rates 403
Detail Versus Manageability 404
Describing Two or More Variables 405
Bivariate Analysis 405
Multivariate Analysis 408
MURDER ON THE JOB 410
Measures of Association 413
Inferential Statistics 418
Univariate Inferences 419
Tests of Statistical Significance 420
Visualizing Statistical Significance 421
Chi Square 422
PUTTING IT ALL TOGETHER: STOPS, DRIVERS, SPEEDERS, AND CITATION ZONES 424
Cautions in Interpreting Statistical Significance 426
Visualizing Discernible Differences 428
Main Points 429
Glossary 433
References 441
Indexes 456
Preface

One of my* most oddly rewarding teaching experiences took place not in the classroom but on the streets of downtown Indianapolis. On my way to a meeting with staff from the Indiana Department of Correction, I recognized a student from the previous semester’s research methods class. Ryan was seated on a shaded bench, clipboard in hand, watching pedestrians make their way down the sidewalk. After we had exchanged greetings, I learned that Ryan had landed a summer internship with the city’s planning department and was currently at work conducting a study of pedestrian traffic.

“Ha!” I exclaimed, recalling student complaints about how research methods are not relevant (what I have since referred to as “Ryan’s lament”). “And you whined about how you were never going to use the stuff we talked about in class.” Ryan responded that the systematic study of pedestrians was interesting, and he admitted that some course topics did, in fact, relate to his work as an intern. He also said something about not really knowing what actual research involved until he began his current project. Ryan remained attentive to people passing by while we chatted for a few minutes. I was pleased to see that he was a careful observer, applying some of the skills he had learned in my course only a few weeks after the semester’s end.

Later, thinking more about the encounter, I recognized the need to change my approach to teaching the course. Ryan clearly enjoyed his experience in doing research, but he had not recognized how much fun research could be until after he left the classroom. As a result, I restructured the course to involve students more actively in the research process. I resolved to be more diligent in linking fundamental concepts of research methods to a broad spectrum of examples, and to show students how they, like Ryan, could apply systematic inquiry and observation techniques to a wide variety of situations in criminal justice and other policy areas.

Collaborating with Earl Babbie to produce this textbook, I joined a colleague whose writing embodied my efforts to engage students in the learning process. Earl’s classic text, The Practice of Social Research, has always been an enviable model of clarity—generating student interest while still presenting a rigorous treatment of social science research methods.

As has always been the case with Practice, our text illustrates principles of doing research with examples specifically selected to appeal to students. We have sought to convey something of the excitement of doing research that Ryan discovered as he observed pedestrians in downtown Indianapolis.

A Familiar, Comfortable Approach

This text has several distinctive features. Anyone who has taught with or learned from The Practice of Social Research will recognize much in our collaborative effort. This enables instructors of criminal justice research methods to organize their course around a familiar approach, capitalizing on the strengths and popularity of Earl Babbie’s superb text. At the same time, we have designed our book to address the particular needs of research methods for criminal justice and criminology.

Features of the New Edition

All editions of our text Research Methods for Criminal Justice and Criminology have retained a good deal of raw material from Practice, albeit revised and otherwise adapted for students in criminology and criminal justice. In preparing the Fifth Edition, we stayed with what has proved to be a popular formula. But we have

*Note: In this Preface, the first-person singular refers to Michael Maxfield, while the first-person plural refers to Maxfield and Earl Babbie.
also responded to suggestions from several people—reviewers, colleagues, and instructors—who used earlier editions.

Putting It All Together: “Driving while Black”

The running example of declining crime in New York has been a popular feature of the text since it was first introduced in the third edition. Its primary purpose was to illustrate elements of each chapter’s content with one example that is carried through the text. While we have received many comments praising this feature, we decided to replace it with a new running example, racial profiling, also known as “driving while black.” Two reasons underlie this change. First, one of the appeals of the declining crime example was that it was current and therefore familiar to students. That’s less true in the year 2007, when downward trends in crime have been reversed in many cities (though the decline continues in New York). Second, reviewers have pointed out that while New York is the country’s most high-profile city it is also unique in many respects and remote from students’ experience.

Our new running example is more current and familiar. Racial profiling began to attract attention from mass media and public officials in the late 1990s and has continued to some extent. New Jersey, where one of us lives and teaches, has probably been most widely known among areas where racially biased policing has been a problem, but it has also emerged in states and cities throughout the country. We draw primarily on major studies conducted in New Jersey, North Carolina, and Pennsylvania to examine the problem of racial profiling in each chapter. Students and instructors in other states should recognize similar problems closer to home. Instructors should also be able to supplement the new running example with similar material of local interest.

The new example retains two important features. First, it illustrates a wide range of questions in research methods, from ethics through data analysis. Second, Maxfield and a Rutgers University student were involved in research on racial profiling in New Jersey. Installments in many chapters draw on the PhD dissertation by Carsten Andresen (2005) and a report prepared by Maxfield and colleague George Kelling (2005). For those who found the New York example especially useful, it has been retained on the text’s companion website. So instructors may continue to use and update this example as appropriate.

Online Sessions

Each chapter now includes a series of exercises under the heading “Online Sessions.” This new feature resembles the Internet exercises from previous editions, but it is based more on my own experiences in the classroom.

The Online Sessions reflect my increased use of Internet-based resources to illustrate material covered in the text. I almost always now teach with a computer projector and Internet connection. That makes it possible to pull up reports, data tables, examples of questionnaires, and an endless variety of other items.

I have also recently used online statistics sites to illustrate some basic data analysis applications. Displaying such websites on a projector really enhances classroom discussion.

A variation on this is to have students locate an online resource as part of an assignment. They locate websites, write up descriptions relating the site to chapter content, then e-mail the assignment and web address to me. During class sessions I can then link to sites submitted by students, project these along with what students have written, and launch the discussion by having students talk about their work. This technique makes the course more interactive and can even be used in large classes. It really engages students in the discussion.

In most cases, I prepared the online sessions for each chapter as if I were preparing a class session on the chapter’s material. As much as possible I try to supplement chapter content with current reports, data, or even current hot
topics. It’s especially fun to include topics I may be currently studying myself. For example, Chapter 10 directs students to “hot spot” locations for car crime in New Jersey. Using an online mapping link, students can observe the features of parking lots that have been found to be high-risk locations.

Even if you do not use the examples listed for each chapter, we strongly encourage you to draw on the vast array of web resources to supplement class lectures. In any event, the Online Sessions section for each chapter should offer a number of ideas for expanding what can be covered in the classroom.

Student Research Examples
With each edition, we have been able to incorporate more examples from students at Rutgers University. By now, in the Fifth Edition, this feature has become even more prominent. It is especially satisfying to cite research completed by students as they pursue advanced degrees. Many of the examples will not be well known, but they embody the kind of supervision typical between graduate students and faculty advisers. I have great familiarity with the details of work by my students. Such details are rarely described in published articles, and being able to report them adds behind-the-scenes information not readily available elsewhere.

Problem Analysis
Chapter 12 has been substantially revised. Most important, the section on policy analysis is reshaped to focus on problem analysis. This is consistent with evolving practice in applied research, especially problem-oriented policing. The chapter, and other sections of the text, draw on new problem-solving tools available on websites sponsored by agencies in the United States and England. Chapter 12 also includes a new section on space-based analysis in applied research. This includes basic information on crime mapping and hot spots. The Internet is the best resource for illustrations of crime mapping, as more law enforcement and even regional planning agencies grant public access to maps. We refer students to these in several chapters.

Theory and Research
Chapter 2, Theory and Criminal Justice Research, has been extensively revised. The most important change is its focus on theory and research. Sections summarizing criminal justice theories have been deleted. Other chapter contents have been rearranged to make the chapter flow more smoothly. It is now substantially shorter, and should be more accessible for more students.

Notable Features Retained and Updated
Qualitative and Quantitative
Reflecting on our own graduate training more than a few years past, memories of the quantitative/qualitative debate remain fresh. Some reviewers of this text have objected to what they perceive as a quantitative bias, pointing to the lack of a chapter titled something along the lines of “Qualitative Field Research.” However, a quick review of the contents in this and previous editions also fails to locate a chapter with a title like “Quantitative Designs.” To be even more precise, an index search of all text files for this edition finds the word “qualitative” appearing in 29 paragraphs, headings, and citations while the word “quantitative” appears in 26. Lest we be charged with conducting a quantitative analysis to illustrate a qualitative point, we cheerfully plead “guilty”!

Chapter 1 addresses the traditional qualitative/quantitative distinction directly, concluding with a position carried over into all other chapters: (1) each approach has advantages; (2) researchers are best advised to draw on elements of each approach as appropriate; and (3) the best advice is to do the best you can. In Chapter 4 we describe threats to validity as illustrations of how research findings might be wrong, an approach that applies to all
methods. Chapters on measurement (5 and 6) and making observations (9, 10, 11) examine a range of approaches. Design is also discussed in several chapters, though most thoroughly in Chapters 7 and 12.

In all cases we present a range of methods that will be familiar to instructors who may view themselves as either qualitative or quantitative researchers. This is entirely consistent with the most widely respected book on research design in the social sciences (Shadish, Cook, and Campbell, 2002).

**Survey Technology**
The Internet continues to develop as an increasingly viable conduit for survey research. Chapter 9 discusses how sampling, item presentation, and analysis capabilities continue to expand. The chapter also cautions that web-based surveys have their own problems, not the least of which is a higher-tech version of junk mail and telemarketing. Nonetheless, we expanded our discussion of web-facilitated surveys and sampling.

**The Research Writer for Criminal Justice CD-ROM**
A new version of the Research Writer for Criminal Justice is included as a CD-ROM. The Research Writer for Criminal Justice helps students tackle the task of writing research reports by taking some of the mystery out of the endeavor; it provides them with a template that they can use as they write their research projects. The template guides students through the process of writing a research report, from the beginning idea to the documentation of the last source. Moreover, students can access helpful hints and tips as they write with just a click of a button. They can also e-mail their work at any stage of the process to their instructor, and they can even export their work to Microsoft Word where the document will be formatted in a style consistent with standard research projects. We hope students find this CD-ROM exciting and helpful.

**Measurement**
Many people believe that measurement presents the greatest challenge in doing social science research. We feel this is especially true for criminal justice. In response to suggestions, the two chapters on measurement (5 and 6) include additional examples and more in-depth discussion to illustrate the importance of conceptualization and measurement. As in earlier editions, throughout the text we remind students to be careful but creative when it comes to measurement. This edition adds information on a new survey of self-reported offending and behavior initiated in England and Wales.

**Design Building Blocks**
Research methods seems to be one of those courses in which students feel driven to memorize the book, lectures, practice questions, last semester’s notes purchased from a friend, or whatever. The strange vocabulary of the subject probably encourages this tendency, which seems to be especially troublesome when teaching research design. All those Xs and Os must be faithfully recorded, memorized, and associated with the right label.

Although the XO diagrams in Chapter 7 will be familiar to many instructors, we encourage students to become engaged in learning research methods by describing how common research designs represent creative uses of design building blocks. In Chapter 7 and elsewhere, we describe how the fundamentals of subject selection, making observations, and administering or withholding a treatment represent basic building blocks of design. Then we offer examples to illustrate different ways to combine those building blocks. Rather than trying to shoehorn a research problem into some pattern of Xs and Os that appears in their books, we urge students to learn what these building
blocks represent and what different building blocks can be expected to accomplish. Happily, this is similar to the position taken by Shadish, Cook, and Campbell (2002) in their revision of the classic on experiments and quasi-experiments.

**Survey Research and Sampling**

One of the strengths of *The Practice of Social Research* has always been a comprehensive but eminently readable treatment of sampling and survey methods. Our criminal justice text (in Chapters 8 and 9) retains Earl’s general approach, but it also points to some of the more specialized criminal justice applications. These range from victim and self-report surveys to specialized interviews with nonprobability samples. Again, our approach focuses on arming students with the principles of survey methods so that they can adapt these general tools to a variety of uses. The chapter also includes more material on special-purpose crime surveys, such as those seeking to measure family violence. The chapter on sampling has been streamlined to make this material more accessible to students.

**Applied Research and Policy Experiments**

This edition retains a chapter devoted to applied criminal justice research; Chapter 12 examines program evaluation and problem analysis, though the latter is revised from previous editions. In addition, we link policy and management applications to virtually every stage of the research process, from theory in Chapter 2 through data interpretation in Chapter 13. We feel this approach is crucial for two complementary reasons. First, students whose interests center on criminal justice policy must understand that applied research is as dependent on theory and reasoned expectations as is basic research. Second, basic research in criminology or criminal justice is usually conducted in some applied context, so the researcher interested in a causal proposition about, say, drug use and violence must recognize that these are not simply abstract constructs. Most measures of drug use and violence will be operationalized with legal or policy definitions of those constructs in mind.

Randomized field experiments have become the designs of choice for many applied studies. Our chapter on experimental and quasi-experimental designs (Chapter 7) and applied research describe the advantages of randomized designs. However, we also caution students that all designs have weaknesses and no design is well suited for all research purposes. Unfortunately, the weaknesses of randomized experiments are sometimes overlooked by their champions and by many textbooks. Our treatment of this topic, encouraging students to think carefully and creatively, is more balanced and will enable students to better recognize appropriate and inappropriate uses of experiments. Expanded material on scientific realism reinforces this perspective.

**Data Analysis**

Chapter 13 guides students through fundamental principles of descriptive and inferential statistics. Our coverage here is conceptual and brief, reflecting our view that criminal justice research design, measurement, and data collection require the concerted attention of students for a full semester. We also believe that understanding these issues is a necessary foundation for doing meaningful statistical analysis. Future producers and consumers of criminal justice research must understand how concepts become observations and how observations become data before they learn the details of data analysis.

We have deleted sections on multiple regression since the linear families of that technique are used less often in criminal justice, and nonlinear models are too complex for the level
of discussion we have chosen. This change also makes the chapter more accessible.

At the same time, our approach to statistics is both thorough and conceptually sound. As a result, instructors who wish to cover data analysis in more detail (perhaps in a second semester course) will find Chapter 13 an excellent point of departure.

Appendices
Now that we have a companion website to accompany this text, we’ve moved the appendices from the CD-ROM to the website for greater access anywhere, anytime. Each appendix has been updated to reflect new web-based resources.

Appendix A includes examples and general information about evolving library and information technology. Depending on the availability of tools and resources on your campus, we suggest that you supplement Appendix A with guides or manuals that document campus facilities and routines.

Appendix B describes the National Criminal Justice Reference Service (NCJRS), a specialized library and information tool for criminal justice research and policy development. The NCJRS was rather late in developing a useful website, but it now provides ready access to publications issued by Justice Department agencies. Because more agencies are issuing reports in electronic format, virtually all students have access to documents previously available only in a small number of specialized library collections. Appendix B leads students to several sources.

Appendix C presents guidelines on writing research reports. If your course will require a proposal or research report, we recommend that students review this appendix early in the term. See also Chapter 4 for further information on proposals. You may wish either to supplement or to modify our suggestions to reflect your own preferences. To complement this appendix, Wadsworth has replaced the CD-ROM that accompanied the third edition with the Research Writer for Criminal Justice CD-ROM.

Appendix D describes major sources of secondary data, most notably the National Archive of Criminal Justice Data (NACJD) and the Interuniversity Consortium for Political and Social Research (ICPSR). If adequate computing resources are available, students can retrieve information about NACJD and ICPSR holdings over the Internet. If not, online data analysis of selected collections is available at the NACJD.

Ancillary Materials
An extensive package of supplemental aids accompanies this edition of Research Methods for Criminal Justice and Criminology:

For the Instructor
- Instructor’s Resource Manual  The fully updated and revised Instructor’s Resource Manual for this edition includes learning objectives, detailed chapter outlines, key terms and figures, class discussion exercises, lecture suggestions, and a complete test bank. Each chapter’s test bank contains multiple-choice, true-false, fill-in-the-blank, and essay questions, and a full answer key.
- ExamView® Assessment Suite  The comprehensive Instructor’s Resource Manual described above is backed up by ExamView, a computerized test bank available for IBM-PC compatibles and Macintosh computers. With ExamView, an easy-to-use assessment and tutorial system, instructors can create, deliver, and customize tests and study guides (both print and online) in minutes. You can easily edit and import your own questions and graphics, change test layouts, and reorganize questions. And using ExamView’s complete word processing capabili-
ties, you can enter an unlimited number of new questions or edit existing questions.

• **JoinIn on Turning Point®** Spark discussion and assess your students’ comprehension of chapter concepts with interactive classroom quizzes and background polls developed specifically for use with this edition of *Research Methods for Criminal Justice and Criminology*. Thomson Wadsworth’s exclusive agreement with TurningPoint software lets you run our tailor-made Microsoft® PowerPoint® slides in conjunction with the “clicker” hardware of your choice. Enhance how your students interact with you, your lecture, and each other. *For college and university adopters only. Contact your local Thomson representative to learn more.*

• **The Wadsworth Criminal Justice Video Library** So many exciting new videos—so many great ways to enrich your lectures and spark discussion of the material in this text! A list of our unique and expansive video program follows. Or, visit [www.cj.wadsworth.com/videos](http://www.cj.wadsworth.com/videos) for a complete, up-to-the-minute list of all of Wadsworth’s video offerings—many of which are also available in DVD format—as well as clip lists and running times. The library includes these selections and many others:

  - **ABC Videos:** Featuring short, high-interest clips from current news events specially developed for courses including Introduction to Criminal Justice, Criminology, Corrections, Terrorism, and White Collar Crime, these videos are perfect for use as discussion starters or lecture launchers to spark student interest. The brief video clips provide students with a new lens through which to view the past and present, one that will greatly enhance their knowledge and understanding of significant events and open up to them new dimensions in learning. Clips are drawn from such programs as *World News Tonight, Good Morning America, This Week, PrimeTime Live, 20/20, and Nightline,* as well as numerous ABC News specials and material from the Associated Press Television News and British Movietone News collections.

  - **The Wadsworth Custom Videos for Criminal Justice:** Produced by Wadsworth and Films for the Humanities, these videos include short (5- to 10-minute) segments that encourage classroom discussion. Topics include white-collar crime, domestic violence, forensics, suicide and the police officer, the court process, the history of corrections, prison society, and juvenile justice.

  - **COURTTV Videos:** One-hour videos presenting seminal and high-profile cases, such as the interrogation of Michael Crowe and serial killer Ted Bundy, as well as crucial and current issues such as cybercrime, double jeopardy, and the management of the prison on Riker’s Island.

  - **A&E American Justice:** 40 videos to choose from, on topics such as deadly force, women on death row, juvenile justice, strange defenses, and Alcatraz.

  - **Films for the Humanities:** Nearly 200 videos to choose from on a variety of topics such as elder abuse, supermax prisons, suicide and the police officer, the making of an FBI agent, domestic violence, and more.

  - **Oral History Project:** Developed in association with the American Society of Criminology, the Academy of Criminal Justice Society, and the National Institute of Justice, these videos will help you introduce your students to the scholars who have developed the criminal justice discipline. Compiled over the last several years, each video features a set of Guest Lecturers—scholars whose thinking has helped to build the foundation of present ideas in the discipline.

• **Classroom Activities for Criminal Justice** This valuable booklet, available to adopters of any Wadsworth criminal justice text, offers instructors the best of the
best in criminal justice classroom activities. Containing both tried-and-true favorites and exciting new projects, its activities are drawn from across the spectrum of criminal justice subjects, including introduction to criminal justice, criminology, corrections, criminal law, policing, and juvenile justice, and can be customized to fit any course. Novice and seasoned instructors alike will find it a powerful tool to stimulate classroom engagement.

- **Internet Activities for Criminal Justice, Third Edition** This is the resource that no introductory criminal justice instructor should be without! The user-friendly booklet allows instructors to send their students far beyond the classroom, guiding them online to conduct research and retrieve information. Its URLs and virtual projects, drawn from all foundational criminal justice areas, have been completely revised and expanded for 2008.

- **The Wadsworth Criminal Justice Resource Center, www.thomsonedu.com/criminaljustice** Designed with the instructor in mind, this website features information about Thomson Wadsworth’s technology and teaching solutions, as well as several features created specifically for today’s criminal justice student. Supreme Court updates, timelines, and hot-topic polling can all be used to supplement in-class assignments and discussions. You’ll also find a wealth of links to careers and news in criminal justice, book-specific sites, and much more.

**For the Student**

- **Study Guide** The already extensive student study guide that accompanies Research Methods for Criminal Justice and Criminology has been thoroughly revised and updated for the latest edition. Because students learn in different ways, the guide includes a variety of pedagogical aids. Each chapter is outlined and summarized, major terms and figures are defined, and worksheets and self-tests are provided.

- **Online Sessions** Online study sessions for each chapter provide specific assignments that use interesting, real-world research. These sessions will help students to better understand and apply chapter concepts.

- **Criminology: A Workbook Using MicroCase® Explorit, Fifth Edition** This workbook introduces students to the real world of criminological research and data. The goal of this workbook is to provide students with the opportunity to participate in the criminological enterprise. In each chapter, the author introduces a particular topic of criminology and demonstrates how data are used to support, augment, and test the ideas proposed. He also provides worksheets that include review questions on the material covered, as well as additional data analysis exercises that enable students to further explore the topics under investigation in each chapter. This unique text and software package teaches students how to use data to gain a better understanding of crime and of public opinions about crime.

- **Careers in Criminal Justice Website, www.thomsonedu.com/login** This unique website helps students investigate the criminal justice career choices that are right for them with the help of several important tools:
  - Career Profiles—video testimonials from a variety of practicing professionals in the field as well as information on many criminal justice careers, including job descriptions, requirements, training, salary and benefits, and the application process.
  - Interest Assessment—self-assessment tool to help students decide which careers suit their personalities and interests.
  - Career Planner—résumé-writing tips and worksheets, interviewing techniques, and successful job search strategies.
Links for Reference—direct links to federal, state, and local agencies where students can get contact information and learn more about current job opportunities.

Wadsworth’s Guide to Careers in Criminal Justice, Third Edition This handy guide, compiled by Caridad Sanchez-Leguelinel, of John Jay College of Criminal Justice, give students information on a wide variety of career paths, including requirements, salaries, training, contact information for key agencies, and employment outlooks.

Writing for Criminal Justice This book contains articles on writing skills—along with a basic grammar review and a survey of verbal communication on the job—that will give students an introduction to academic, professional, and research writing in criminal justice. The voices of professionals who have used these techniques on the job will help students see the relevance of these skills to their future careers.

Handbook of Selected Supreme Court Cases, Third Edition This supplementary handbook covers almost 40 landmark cases, each of which includes a full case citation, an introduction, a summary from WestLaw, excerpts from the case, and the decision. The updated edition includes Hamdi v. Rumsfeld, Roper v. Simmons, Ring v. Arizona, Atkins v. Virginia, Illinois v. Caballes, and much more.

Current Perspectives: Readings from InfoTrac College Edition These readers, designed to give students a deeper taste of special topics in criminal justice, include free access to InfoTrac College Edition. The timely articles are selected by experts in each topic from within InfoTrac. They are available for free when bundled with the text.

Terrorism and Homeland Security
Cyber Crime
Juvenile Justice
Public Policy and Criminal Justice
Crisis Management and National Emergency Response
Racial Profiling

New Technologies and Criminal Justice
White-Collar Crime

Terrorism: An Interdisciplinary Perspective Available for bundling with each copy of The American System of Criminal Justice, Eleventh Edition, this 80-page booklet discusses terrorism in general and the issues surrounding the events of September 11, 2001. This information-packed booklet examines the origins of terrorism in the Middle East, focusing on Osama bin Laden in particular, as well as issues involving bioterrorism, the specific role played by religion in Middle Eastern terrorism, globalization as it relates to terrorism, and the reactions and repercussions of terrorist attacks.

Crime Scenes 2.0: An Interactive Criminal Justice CD-ROM Recipient of several New Media Magazine Invision Awards, this interactive CD-ROM allows your students to take on the roles of investigating officer, lawyer, parole officer, and judge in excitingly realistic scenarios. Available FREE when bundled with every copy of Criminal Justice in America, Fifth Edition. An instructor’s manual is also available for the CD-ROM online.

Mind of a Killer CD-ROM (bundle version) Voted one of the top 100 CD-ROMs by an annual PC Magazine survey, Mind of a Killer gives students a chilling glimpse into the realm of serial killers, with over 80 minutes of video and 3D simulations, an extensive mapping system, a library, and much more.

Internet Guide for Criminal Justice, Second Edition Intended for the novice user, this guide provides students with the background and vocabulary necessary to navigate and understand the web, and then it offers a wealth of criminal justice websites and Internet project ideas.

Acknowledgments
Several reviewers made perceptive and useful comments on various drafts of the book. We
are especially grateful to them for their insights and suggestions: Shaun L. Gabbidon, Penn State Capital College; David A. Jenks, California State University, Los Angeles; J. Mitchell Miller, University of South Carolina; Michael J. Sabath, San Diego State University; and Theodore P. Skotnicki, Niagara County Community College.

Other colleagues offered suggestions, comments, and advice: Patricia Brantingham, Simon Fraser University; Judith Collins, Michigan State University; Robyn Dawes, Carnegie Mellon University; Brian Forst, American University; Mangai Natarajan and her Fall 2003 class in research methods, John Jay College of Criminal Justice; and Cathy Spatz Widom, John Jay College of Criminal Justice.

A growing number of students at the Rutgers University School of Criminal Justice have offered advice, feedback, and contributions for this and earlier editions. I thank Dr. Gisela Bichler (now at California State University, San Bernardino), Dr. Sharon Chamard (now at University of Alaska), Shuryo Fujita, Galma Jahic (now at Istanbul Bilgi University, Turkey), Dr. Jarret Lovell (now at California State University, Fullerton), Ellen McCann, Dr. Marie Mele (now at Monmouth University), Melanie-Angela Neuilly (now at the University of Idaho), Dina Perrone, Dr. William Sousa (now at the University of Nevada, Las Vegas), and Dr. Christopher Sullivan (now at the University of South Florida). Dr. Carsten Andresen (now at the Travis County Department of Community Corrections and Supervision) merits special thanks for generously sharing information about his own research.

Finally, Earl and I are very grateful for the patient, professional assistance from people at Wadsworth: Chris Caldeira, Amanda Kaufmann, Christina Ho, and the entire book team.
An Introduction to Criminal Justice Inquiry

What comes to mind when you encounter the word *science*? What do you think of when we describe criminal justice as a social science? For some people, science is mathematics; for others, it is white coats and laboratories. Some confuse it with technology or equate it with difficult high school or college courses.

Science is, of course, none of these things per se, but it is difficult to specify exactly what science is. Scientists, in fact, disagree on the proper definition. Some object to the whole idea of social science; others question more specifically whether criminal justice can be a social science.

For the purposes of this book, we view science as a method of inquiry—a way of learning and knowing things about the world around us. Like other ways of learning and knowing about the world, science has some special characteristics. We’ll examine these traits in this opening set of chapters. We’ll also see how the scientific method of inquiry can be applied to the study of crime and criminal justice.

Part One of the book lays the groundwork for the rest of the book by examining the fundamental characteristics and issues that make science different from other ways of knowing things. Chapter 1 begins with a look at native human inquiry, the sort of thing all of us have been doing all our lives. Because people sometimes go astray in trying to understand the world around them, we’ll consider the primary characteristics of scientific inquiry that guard against those errors.

Chapter 2 deals specifically with the social scientific approach to criminal justice inquiry and the links between theory and research. The lessons of Chapter 1 are applied in the study of crime and criminal justice. Although special considerations arise in studying people and organizations, the basic logic of all science is the same.

Ethics is one of these special considerations we face in studying people. In Chapter 3, we’ll see that most ethical questions are rooted in two fundamental principles: (1) Research subjects should not be harmed, and (2) their participation must be voluntary.

The overall purpose of Part One, therefore, is to construct a backdrop against which to view more specific aspects of research design and execution. By the time you complete the chapters in Part One, you’ll be ready to look at some of the more concrete aspects of criminal justice research.
People learn about their world in a variety of ways, and they often make mistakes along the way. Science is different from other ways of learning and knowing. We’ll consider the foundations of social science, different purposes of research, and different general approaches to social science.
Introduction

Criminal justice professionals are both consumers and producers of research.

Spending a semester studying criminal justice research methodology may not be high on your list of “Fun Things to Do.” Perhaps you are or plan to be a criminal justice professional and are thinking, “Why do I have to study research methods? When I graduate, I’ll be working in probation (or law enforcement, or corrections, or court services), not conducting research! I would benefit more from learning about probation counseling (or police management, or corrections policy, or court administration).” Fair enough. But as a criminal justice professional, you will need to be a consumer of research. One objective of this book is to help you become an informed consumer of research.

For example, in the section “Two Realities,” we will see how findings from one of the first experimental studies of policing appeared to contradict a traditional tenet of law enforcement—that a visible patrol force prevents crime. Acting as a consumer of research findings, a police officer, supervisor, or executive should be able to understand how the research was conducted and how the study’s findings might apply in his or her department. Because police practices vary from city to city, a police executive would benefit from being aware of research methods and knowing how to interpret findings.

Most criminal justice professionals, especially those in supervisory roles, routinely review various performance reports and statistical tabulations. In the past 30 years or so, thousands of criminal justice research and evaluation studies have been conducted. The National Criminal Justice Reference Service (http://www.ncjrs.gov) was established to archive and distribute research reports to criminal justice professionals and researchers around the world. Many such reports are prepared specifically to keep the criminal justice community informed about new research developments. More recently, the Center for Problem-Oriented Policing (POP Center, http://www.popcenter.org) was created to share applied research on varieties of law enforcement problems. By understanding research methods decision makers are better equipped to critically evaluate research reports and recognize when methods are properly and improperly applied. The box titled “Home Detention” describes an example of how knowledge of research methods can help policy makers avoid mistakes.
Another objective of this book is to help you produce research. In other courses you take or in your job, you may become a producer of research. For example, probation officers sometimes test new approaches to supervising or counseling clients, and police officers try new methods of working with the community, or with witnesses and suspects. Many cities and states have a compelling need to assess how changes in sentencing policy might affect jail and prison populations. Determining whether such changes are effective is an example of applied research. A problem-solving approach, rooted in systematic research, is being used in more and more police departments and in many other criminal justice agencies as well. Many items on the POP Center website are the product of applied research conducted by police departments. Therefore, criminal justice professionals need to know not only how to interpret research accurately but also how to produce accurate research.

**What Is This Book About?**

*This book focuses on how we know what we know.*

This book focuses on how we learn and know things, not on what we know. Although you will come away from the book knowing some things you don’t know right now, our primary purpose is to help you look at how you know things, not what you know.

**Two Realities**

Ultimately, we live in a world of two realities. Part of what we know could be called our “experiential reality”—the things we know from direct experience. For example, if you dive into
a glacial stream flowing down through the Canadian Rockies, you don’t need anyone to tell you that the water is cold; you notice that all by yourself. And the first time you step on a thorn, you know it hurts even before anyone tells you. The other part of what we know could be called our “agreement reality”—the things we consider real because we’ve been told they’re real and everyone else seems to agree they are real. A big part of growing up in any society, in fact, is learning to accept what everybody around us “knows” to be true. If we don’t know those same things, we can’t really be a part of the group. If you were to seriously question a geography professor as to whether the sun really sets in the west, you’d quickly find yourself set apart from other people. The first reality is a product of our own experience; the second is a product of what people have told us.

To illustrate the difference between agreement and experiential realities, consider preventive police patrol. The term preventive implies that when police patrol their assigned beats they prevent crime. Police do not prevent all crime, of course, but it is a commonsense belief that a visible, mobile police force will prevent some crimes. In fact, the value of patrol in preventing crime was a fundamental tenet of police operations for many years. O.W. Wilson, a legendary police chief in Chicago and the author of an influential book on police administration, wrote that patrol was indispensable in preventing crime by eliminating incentives and opportunities for misconduct (Wilson and McLaren, 1963:320). A 1967 report on policing by President Lyndon Johnson’s Commission on Law Enforcement and Administration of Justice (1967:1) stated that “the heart of the police

<table>
<thead>
<tr>
<th>Convicted</th>
<th>Pretrial</th>
<th>Juveniles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adults</td>
<td>5%</td>
<td>1%</td>
</tr>
<tr>
<td>New arrest</td>
<td>53</td>
<td>52</td>
</tr>
</tbody>
</table>

As the table shows, many more juveniles were arrested, and juveniles successfully answered a much lower percentage of telephone calls to their homes. What happened?

The simple answer is that the staff responsible for administering the juvenile program were not keeping track of offenders. The ELMO equipment was not maintained properly, and police were not visiting the homes of juveniles as planned. Because staff were not keeping track of program participants, they were not aware that many juveniles were violating the conditions of home detention. And because they did not detect violations, they naturally reported that the vast majority of young burglars and thieves completed their home detention successfully.

A county official who relied on only agency reports of program success would have made a big mistake in judging the juvenile program to be 99 percent successful. In contrast, an informed consumer of such reports would have been skeptical of a 99 percent success rate and searched for more information.

Source: Adapted from Maxfield and Baumer (1991) and Baumer, Maxfield, and Mendelsohn (1993).
effort against crime is patrol. . . . The object of patrol is to disperse policemen in a way that will eliminate or reduce the opportunity for misconduct and to increase the probability that a criminal will be apprehended while he is committing a crime or immediately thereafter."

Seven years later, the Police Foundation, a private research organization, published results from an experimental study that presented a dramatic challenge to conventional wisdom. Known as the “Kansas City Preventive Patrol Experiment,” this classic study compared police beats with three levels of preventive patrol: (1) control beats, with one car per beat; (2) proactive beats, with two or three cars per beat; and (3) reactive beats, with no routine preventive patrol. After almost 1 year, researchers examined data from the three types of beats and found no differences in crime rates, citizen satisfaction with police, fear of crime, or other measures of police performance (Kelling, Pate, Dieckman, and Brown, 1974).

Researchers and law enforcement professionals alike were surprised by these findings. For the record, the Kansas City researchers never claimed to have proved that preventive patrol had no impact on crime. Instead, they argued that police should work more closely with community members and that routine patrol might be more effective if combined with other strategies that used police resources in a more thoughtful way.

Additional studies conducted in the 1970s cast doubt on other fundamental assumptions about police practices. A quick response to crime reports made no difference in arrests, according to a research study in Kansas City (Van Kirk, 1977). And criminal investigation by police detectives rarely resulted in an arrest (Greenwood, 1975).

We mention these examples not to attack routine law enforcement practices. Rather, we want to show that systematic research on policing has illustrated how traditional beliefs—agreement reality—can be misleading. Simply increasing the number of police officers on patrol does not reduce crime because police patrol often lacks direction. Faster response time to calls for police assistance does not increase arrests because there is often a long delay between the time when a crime occurs and when it is reported to police. Clever detective work seldom solves crimes because investigators get most of their information from reports prepared by patrol officers, who in turn get their information from victims and witnesses.

Traditional beliefs about patrol effectiveness, response time, and detective work are examples of agreement reality. In contrast, the research projects that produced alternative views about each law enforcement practice represent experiential reality. These studies are examples of empirical research, the production of knowledge based on experience or observation. In each case, researchers conducted studies of police practices and based their conclusions on observations and experience. Empirical research is a way of learning about crime and criminal justice, and explaining how to conduct empirical research is the purpose of this book.

In focusing on empirical research, we do not intend to downplay the importance of other ways of knowing things. Law students, for example, are trained in how to interpret statutes and judicial opinions. Historians take courses on methods of historical interpretation, mathematics majors learn numerical analysis, and students of philosophy study logic. If you are a criminal justice major, many of the other courses you take—say, a course on theories of crime and deviance—will add to your agreement reality.

The Role of Science

Science offers an approach to both agreement reality and experiential reality. Scientists have certain criteria that must be met before they will agree on the reality of something they haven’t personally experienced. In general, an
assertion must have both *logical* and *empirical* support: It must make sense, and it must agree with actual observations. For example, why do earthbound scientists accept the assertion that it’s cold on the dark side of the moon? First, it makes sense because the surface heat of the moon comes from the sun’s rays. Second, scientific measurements made on the moon’s dark side confirm the assertion. Therefore, scientists accept the reality of things they don’t personally experience—they accept an agreement reality—but they have special standards for doing so.

More to the point of this book, however, science offers a special approach to the discovery of reality through personal experience. Epistemology is the science of knowing; methodology (a subfield of epistemology) might be called “the science of finding out.” This book focuses on criminal justice methodology—how social science methods can be used to better understand crime and criminal justice policy. To understand scientific inquiry, let’s first look at the kinds of inquiry we all do each day.

**Personal Human Inquiry**

*Everyday human inquiry draws on personal experience and secondhand authority.*

Most of us feel more comfortable if we understand what’s going on around us, and are able to predict our future circumstances. We seem quite willing, moreover, to undertake this task using causal and probabilistic reasoning. First, we generally recognize that future circumstances are somehow caused or conditioned by present ones. For example, we learn that getting an education will affect how much money we earn later in life and that running stoplights may result in an unhappy encounter with an alert traffic officer. As students, we learn that studying hard will result in better examination grades.

Second, we recognize that such patterns of cause and effect are probabilistic in nature: The effects occur more often when the causes occur than when the causes are absent—but not always. Thus, as students, we learn that studying hard produces good grades in most instances, but not every time. We recognize the danger of ignoring stoplights without believing that every such violation will produce a traffic ticket.

The concepts of causality and probability play a prominent role in this book. Science makes causality and probability more explicit and provides techniques for dealing with them more rigorously than does casual human inquiry.

As we look at our own personal inquiry, it is useful to distinguish between prediction and understanding. Attempts to predict often occur in a context of knowledge and understanding. If we can understand why things are related to one another, why certain regular patterns occur, we can predict better than if we simply observe and remember those patterns. For example, college dormitories are often targets of burglary during semester breaks. If you understand something about the opportunity structure of burglary (it’s more likely when a suitably motivated offender knows that a dwelling will be unoccupied), you will be able to predict that hanging a sign on your room that says, “Off to Jamaica! See you in two weeks!” invites trouble.

Thus, human inquiry attempts to answer both *what* and *why* questions, and we pursue these goals by observing phenomena and figuring them out.

However, our attempts to learn about the world are only partly linked to personal inquiry and direct experience. Another, much larger, part comes from the agreed-on knowledge that others give us. This agreement reality both assists and hinders our attempts to find out things for ourselves. Two important sources of secondhand knowledge—tradition and authority—deserve brief consideration here.

**Tradition**

Each of us is born into and inherits a culture made up, in part, of firmly accepted knowledge about the workings of the world. We may learn
from others that planting corn in the spring will result in the greatest assistance from the gods, that the circumference of a circle is approximately $3\frac{1}{7}$ times its diameter, or that driving on the left side of the road (in the United States) is dangerous. We may test a few of these “truths” on our own, but we simply accept the great majority of them. These are the things that “everybody knows.”

Tradition, in this sense, has some clear advantages for human inquiry. By accepting what everybody knows, we are spared the overwhelming task of starting from scratch in our search for regularities and understanding. Knowledge is cumulative, and an inherited body of information and understanding is the jumping-off point for the development of more knowledge.

**Authority**

Despite the power of tradition, new knowledge appears every day. In addition to our own personal inquiries, throughout life we learn about the new discoveries and understandings of others. Our acceptance of this new knowledge often depends on the status of the discoverer. For example, you are more likely to believe a judge who declares that your next traffic violation will result in a suspension of your driver’s license than to believe your parents when they say the same thing.

Like tradition, authority can both help and hinder human inquiry. We do well to trust the judgment of individuals who have special training, expertise, and credentials in a matter, especially in the face of contradictory arguments on a given question. At the same time, inquiry can be greatly hindered by the legitimate authorities who err within their own special province. Biologists, after all, do make mistakes in the field of biology, and biological knowledge changes over time. Most of us assume that over-the-counter medications are safe when taken as directed, trusting the authority of drug manufacturers and government agencies.

---

**ARREST AND DOMESTIC VIOLENCE**

In 1983, preliminary results were released from a study on the deterrent effects of arrest in cases of domestic violence. The study reported that male abusers who were arrested were less likely to commit future assaults than offenders who were not arrested. Conducted by researchers from the Police Foundation, the study used rigorous experimental methods adapted from the natural sciences. Criminal justice scholars generally agreed that the research was well designed and executed. Public officials were quick to embrace the study’s findings that arresting domestic violence offenders deterred them from future violence.

Here, at last, was empirical evidence to support an effective policy in combating domestic assaults. Results of the Minneapolis Domestic Violence Experiment were widely disseminated, in part due to aggressive efforts by the researchers to publicize their findings (Sherman and Cohn, 1989). The attorney general of the United States recommended that police departments make arrests in all cases of misdemeanor domestic violence. Within 5 years, more than 80 percent of law enforcement agencies in U.S. cities adopted arrest as the preferred way of responding to domestic assaults (Sherman, 1992:2).

Several things contributed to the rapid adoption of arrest policies to deter domestic violence. First, the experimental study was conducted carefully by highly respected researchers. Second, results were widely publicized in newspapers, in professional journals, and on television programs. Third, officials could understand the study, and most believed that its findings made sense. Finally, mandating arrest in less serious cases of domestic violence was a straightforward
However, in the late nineteenth century, our trust might have led us to buy a bottle of Bayer Heroin, then available as an over-the-counter pain relief medication (Inciardi, 1986). The box titled “Arrest and Domestic Violence” illustrates the difficult problems that can result when criminal justice policy makers accept too quickly the results from criminal justice research.

Inquiry is also hindered when we depend on the authority of experts speaking outside their realm of expertise. For example, consider the political or religious leader, lacking any biochemical expertise, who declares marijuana to be a dangerous drug. The advertising industry plays heavily on this misuse of authority by having popular athletes endorse various consumer products.

Both tradition and authority, then, are double-edged swords in the search for knowledge about the world. Simply put, they provide us with a starting point for our own inquiry, but they may lead us to start at the wrong point or push us in the wrong direction.

**Errors in Personal Human Inquiry**

*Everyday personal human inquiry reveals a number of potential biases.*

Aside from the potential dangers of relying on tradition and authority, we often stumble when we set out to learn for ourselves. Let’s consider some of the common errors we make in our own casual inquiries and then look at the ways science provides safeguards against those errors.

**Inaccurate Observation**

The keystone of inquiry is observation. We can never understand the way things are without first having something to understand. We have
to know what before we can explain why. On the whole, however, people are rather sloppy observers of the flow of events in life. We fail to observe things right in front of us and mistakenly observe things that aren’t so. Do you recall, for example, what your instructor was wearing on the first day of this class? If you had to guess now, what are the chances you would be right?

In contrast to casual human inquiry, scientific observation is a conscious activity. Simply making observations in a more deliberate way helps to reduce error. If you had gone to the first class meeting with a conscious plan to observe and record what your instructor was wearing, you’d have been more accurate.

In many cases, using both simple and complex measurement devices helps to guard against inaccurate observations. Suppose, for example, that you had taken a photograph of your instructor on the first day. The photo would have added a degree of precision well beyond that provided by unassisted human senses.

Overgeneralization

When we look for patterns among the specific things we observe around us, we often assume that a few similar events are evidence of a general pattern. The tendency to overgeneralize is probably greatest when there is pressure to reach a general understanding, yet overgeneralization also occurs in the absence of pressure. Whenever overgeneralization does occur, it can misdirect or impede inquiry.

Imagine you are a rookie police officer newly assigned to foot patrol in an urban neighborhood. Your sergeant wants to meet with you at the end of your shift to discuss what you think are the major law enforcement problems on the beat. Eager to earn favor with your supervisor, you interview the manager of a popular store in a small shopping area. If the manager mentions vandalism as the biggest concern, you might report that vandalism is the main problem on your beat, even though other store managers believe that drug dealing is the main problem, and that it contributes to local burglary, car break-ins, robbery, and vandalism. Overgeneralization would lead to misrepresentation and simplification of the problems on your beat.

Criminal justice researchers guard against overgeneralization by committing themselves in advance to a sufficiently large sample of observations and by being attentive to how representative those observations are. The replication of inquiry provides another safeguard. Replication means repeating a study, checking to see whether similar results are obtained each time. The study may also be repeated under slightly different conditions or in different locations. The box titled “Arrest and Domestic Violence” describes an example of why replication can be especially important in applied research.

Selective Observation

One danger of overgeneralization is that it may lead to selective observation. Once we have concluded that a particular pattern exists and have developed a general understanding of why, we will be tempted to pay attention to future events and situations that correspond with the pattern and to ignore those that don’t. Racial, ethnic, and other prejudices are reinforced by selective observation.

Researchers often specify in advance the number and kind of observations to be made before marking a conclusion to a particular project. For example, if we wanted to learn whether women were more likely than men to support long prison sentences for sex offenders, we would have to make a specified number of observations on that question. We might select a thousand people to be interviewed. Even if the first 10 women supported long sentences and the first 10 men opposed them, we would continue to interview everyone selected for the study and record each observation. We would base our conclusion on an analysis of all the observations, not just those first 20.
Illogical Reasoning

People have various ways of handling observations that contradict their judgments about the way things are. Surely one of the most remarkable creations of the human mind is “the exception that proves the rule,” an idea that makes no sense at all. An exception can draw attention to a rule or to a supposed rule, but in no system of logic can it prove the rule it contradicts. Yet we often use this pithy saying to brush away contradictions with a simple stroke of illogic.

What statisticians call the “gambler’s fallacy” is another illustration of illogic in day-to-day reasoning. According to this fallacy, a consistent run of either good or bad luck is presumed to foreshadow its opposite. Thus, an evening of bad luck at poker may kindle the belief that a winning hand is just around the corner; many a poker player has stayed in a game too long because of that mistaken belief. Conversely, an extended period of good weather may lead us to worry that it is certain to rain on our weekend picnic.

Although we all sometimes use embarrassingly illogical reasoning, scientists avoid this pitfall by using systems of logic consciously and explicitly. Chapters 2 and 4 examine the logic(s) of science in more depth.

Ideology and Politics

Crime is, of course, an important social problem, and a great deal of controversy surrounds policies for dealing with crime. Many people feel strongly one way or another about the death penalty, gun control, and long prison terms for drug users as approaches to reducing crime. There is ongoing concern about racial bias in police practices and sentencing policies. Ideological or political views on such issues can undermine objectivity in the research process. Criminal justice professionals may have particular difficulty separating ideology and politics from a more detached, scientific study of crime.

Criminologist Samuel Walker (1994:16) compares ideological bias in criminal justice research to theology: “The basic problem . . . is that faith triumphs over facts. For both liberals and conservatives, certain ideas are unchallenged articles of faith, almost like religious beliefs that remain unshaken by empirical facts.”

Most of us have our own beliefs about public policy, including policies for dealing with crime. The danger lies in allowing such beliefs to distort how research problems are defined and how research results are interpreted. The scientific approach to the study of crime and criminal justice policy guards against, but does not prevent, ideology and theology from coloring the research process. In empirical research, so-called articles of faith are compared with experience.

To Err Is Human

We have seen some of the ways that we can go astray in our attempts to know and understand the world and some of the ways that science protects its inquiries from these pitfalls. Social science differs from our casual, day-to-day inquiry in two important respects. First, social scientific inquiry is a conscious activity. Although we engage in continuous observation in daily life, much of it is unconscious or semi-conscious. In social scientific inquiry, in contrast, we make a conscious decision to observe, and we stay alert while we do it. Second, social scientific inquiry is a more careful process than our casual efforts; we are more wary of making mistakes and take special precautions to avoid doing so.

Do social science research methods offer total protection against the errors that people commit in personal inquiry? No. Not only do individuals make every kind of error we’ve looked at, but social scientists as a group also succumb to the pitfalls and stay trapped for long periods of time.
Foundations of Social Science

Social scientific inquiry generates knowledge through logic and observation.

Science is sometimes characterized as “logico-empirical.” This ungainly term carries an important message: The two pillars of science are (1) logic or rationality and (2) observation. A scientific understanding of the world must make sense and must agree with what we observe. Both of these elements are essential to social science and relate to three key aspects of the overall scientific enterprise: theory, data collection, and data analysis.

As a broad generalization, scientific theory deals with the logical aspect of science, data collection deals with the observational aspect, and data analysis looks for patterns in what is observed. This book focuses mainly on issues related to data collection—demonstrating how to conduct empirical research—but social science involves all three elements. With this in mind, Chapter 2 discusses the theoretical context of designing and executing research, and Chapter 13 presents a conceptual introduction to the statistical analysis of data. Figure 1.1 offers a schematic view of how the book addresses these three aspects of social science.

Let’s turn now to some of the fundamental issues that distinguish social science from other ways of looking at social phenomena.

Theory, Not Philosophy or Belief

Social scientific theory has to do with what is, not what should be. This means that scientific theory—and, more broadly, science itself—cannot settle debates on value or worth. Social science cannot determine, for example, whether prosecutors who are elected (as in most states) are “better” than prosecutors who are appointed by a state official (as in New Jersey) except in terms of some agreed-on criteria. We can determine scientifically whether elected or appointed prosecutors are more respected by the citizens they serve only if we agree on some measures of citizen respect, and our conclusion will depend totally on the measures we agree on.

By the same token, if we agree that, say, conviction rate or average sentence length is a good measure of a prosecutor’s quality, then we will be in a position to determine scientifically whether a prosecutor in one city is better or worse than a prosecutor in another city. Again, however, our conclusion will be inextricably tied to the agreed-on criteria. As a practical matter, however, people are seldom able to agree on criteria for determining issues of value, so science is seldom of any use in settling such debates.

We return to this point in Chapter 12 when we examine evaluation research. Criminal justice research often involves studying whether programs achieve their intended goals. One of the biggest problems faced by evaluation researchers is getting people to agree on the criteria for success and failure. Yet such criteria are essential if social scientific research is to tell us anything useful about matters of value.

As an example, consider the dilemma of how to identify a good parole officer. On the one hand, a parole officer whose clients are rarely cited for violations and returned to prison might be considered a good officer. However, parole officers might attain low violation rates by ignoring misbehavior among clients they are supposed to supervise—in effect, by not supervising them at all. Therefore, we might view a parole officer who frequently cites parolees for violations as being especially attentive to his or her job. We might also consider other factors in judging parole officers. Someone who routinely cites parolees for trivial rule infractions would not necessarily be considered a good officer, especially if such actions swell already crowded prison populations.

Thus, social science can assist us in knowing only what is and why. It can be used to address the question of what ought to be only when people agree on the criteria for deciding
Chapter 2

DATA COLLECTION

Planning to do research

Chapters 3–7

Sampling

Chapter 8

Observation

Chapters 9–11

DATA ANALYSIS

\[ Y = a + x_1 + x_2 + x_3 + x_4 + \theta \]

<table>
<thead>
<tr>
<th>x</th>
<th>( \bar{x} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>34%</td>
</tr>
<tr>
<td>( \bar{y} )</td>
<td>78%</td>
</tr>
</tbody>
</table>

| y  | 66%         |
| \( \bar{y} \) | 22%         |

Figure 1.1 Social Science = Theory + Data Collection + Data Analysis
what makes one thing better than another. But this agreement seldom occurs. With that understanding, let’s turn now to some of the fundamentals that social scientists use to develop theories about what is and why.

Regularities

Ultimately, social science aims to find patterns of regularity in social life. This assumes, of course, that life is regular, not chaotic or random. That assumption applies to all science, but it is sometimes a barrier for people when they first approach social science.

Certainly, at first glance, the subject matter of the physical sciences appears to be more regular than that of the social sciences. A heavy object, after all, falls to earth every time we drop it. In contrast, a judge may sentence one person to prison and give another probation even though each is convicted of the same offense.

A vast number of norms and rules in society create regularity. For example, only persons who have reached a certain age may obtain a driver’s license. In the National Hockey League, only men participate on the ice. Such informal and formal prescriptions, then, regulate, or regularize, social behavior.

In addition to regularities produced by norms and rules, social science deals with other types of regularities based on observation. For example, teenagers commit more crimes than middle-aged people. When males commit murder, they usually kill another male, but female murderers more often kill a male. On average, white urban residents view police more favorably than nonwhites do. Judges receive higher salaries than police officers. And probation officers have more empathy for the people they supervise than prison guards do.

What About Exceptions?

The objection that there are always exceptions to any social regularity does not mean that the regularity itself is unreal or unimportant. A police officer in a large city might earn more than a judge in a small town, but overall judges earn more than police officers. The pattern still exists. Social regularities represent probabilistic patterns, and a general pattern does not have to be reflected in 100 percent of the observable cases to be a pattern.

This rule applies in the physical as well as the social sciences. In genetics, for example, the mating of a blue-eyed person with a brown-eyed person will probably result in a brown-eyed offspring. The birth of a blue-eyed child does not challenge the observed regularity, however. Rather, the geneticist states only that the brown-eyed offspring is more likely and, furthermore, that a brown-eyed offspring will be born in only a certain percentage of cases. The social scientist makes a similar, probabilistic prediction: that women overall are less likely to murder anybody, but when they do, their victims are most often males.

Aggregates, Not Individuals

Social scientists primarily study social patterns rather than individual ones. All regular patterns reflect the aggregate or combined actions and situations of many individuals. Although social scientists study motivations that affect individuals, aggregates are more often the subject of social science research.

A focus on aggregate patterns rather than on individuals distinguishes the activities of criminal justice researchers from the daily routines of many criminal justice practitioners. Consider, for example, the task of processing and classifying individuals newly admitted to a correctional facility. Prison staff administer psychological tests and review the prior record of each new inmate to determine security risks, program needs, and job options. A researcher who is studying whether white inmates tended to be assigned to more desirable jobs than nonwhite inmates would be more interested in patterns of job assignment. The focus would be on aggregates of white and nonwhite persons rather than the assignment for any particular individual.
Social scientific theories, then, typically deal with aggregate, not individual, behavior. Their purpose is to explain why aggregate patterns of behavior are so regular even when the individuals who perform them change over time. In another important sense, social science doesn’t even seek to explain people. Rather, it seeks to understand the systems within which people operate, the systems that explain why people do what they do. The elements in such systems are not people but variables.

A Variable Language
Our natural attempts at understanding usually take place at the concrete, idiosyncratic level. That’s just the way we think. Suppose someone says to you, “Women are too soft-hearted and weak to be police officers.” You are likely to “hear” that comment in terms of what you know about the speaker. If it’s your old Uncle Albert, who, you recall, is also strongly opposed to daylight savings time, zip codes, and computers, you are likely to think his latest pronouncement simply fits into his dated views about things in general. If, in contrast, the statement comes from a candidate for sheriff who is trailing a female challenger and who has begun making other statements about women being unfit for public office, you may “hear” his latest comment in the context of this political challenge.

In both of these examples, we try to understand the thoughts of a particular, concrete individual. In social science, however, we go beyond that level of understanding to seek insights into classes or types of individuals. In the two preceding examples, we might use terms like old-fashioned or bigoted to describe the person who made the comment. In other words, we try to identify the actual individual with some set of similar individuals, and that identification operates on the basis of abstract concepts.

One implication of this approach is that it enables us to make sense out of more than one person. In understanding what makes the bigoted candidate think the way he does, we can also learn about other people who are “like him.” This is possible because we have not been studying bigots as much as we have been studying bigotry.

Bigotry is considered a variable in this case because the level of bigotry varies. Some people in an observed group are more bigoted than others. Social scientists may be interested in understanding other variables that might cause bigotry to be high in one instance and low in another. However, bigotry is not the only variable here. Gender, age, and economic status also vary among members of the observed group.

Here’s another example: Consider the problem of whether police should make arrests in cases of domestic violence. The object of a police officer’s attention in handling a domestic assault is the individual case. Of course, each case includes a victim and an offender, and police are concerned with preventing further harm to the victim. The officer must decide whether to arrest an assailant or to take some other action. The criminal justice researcher’s subject matter is different: Does arrest as a general policy prevent future assaults? The researcher may study an individual case (victim and offender), but that case is relevant only as a situation in which an arrest policy might be invoked, which is what the researcher is really studying.

This is not to say that criminal justice researchers don’t care about real people. They certainly do. Their ultimate purpose in studying domestic violence cases is to identify ways to protect potential victims from future assaults. But in this example, victims and offenders are most relevant for what they reveal about the effectiveness of the arrest policy. As researchers, our interest centers on variables and aggregates, not on individuals.

Variables and Attributes
Social scientists study variables and the attributes that compose them. Social scientific theories are written in a variable language,
and people get involved mostly as the carriers of those variables. Here's what social scientists mean by variables and attributes.

**Attributes** are characteristics or qualities that describe some object such as a person. Examples are “bigoted,” “old-fashioned,” “married,” “unemployed,” and “intoxicated.” Any quality we might use to describe ourselves or someone else is an attribute.

**Variables** are logical groupings of attributes. Thus, for example, “male” and “female” are attributes, and “gender” is the variable composed of the logical grouping of those two attributes. The variable “occupation” is composed of attributes like “dentist,” “professor,” and “truck driver.” “Prior record” is a variable composed of a set of attributes such as “prior convictions,” “prior arrests without convictions,” and “no prior arrests.” It’s helpful to think of attributes as the categories that make up a variable. See Figure 1.2 for a schematic view of what social scientists mean by variables and attributes.

The relationship between attributes and variables lies at the heart of both description and explanation in science. For example, we might describe a prosecutor’s office in terms of the variable “gender” by reporting the observed frequencies of the attributes “male” and “female”: “The office staff is 60 percent men and 40 percent women.” An incarceration rate can be thought of as a description of the variable “incarceration status” of a state’s population in terms of the attributes “incarcerated” and “not incarcerated.” Even the report of family income for a city is a summary of attributes composing

---

**Figure 1.2** Variables and Attributes

<table>
<thead>
<tr>
<th>A. Some Common Criminal Justice Concepts</th>
<th>B. Two Different Kinds of Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td><strong>Variables</strong></td>
</tr>
<tr>
<td>Female</td>
<td><strong>Gender</strong></td>
</tr>
<tr>
<td>Probation</td>
<td>Sentence</td>
</tr>
<tr>
<td>Thief</td>
<td>Property crime</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td>Age</td>
</tr>
<tr>
<td>Sentence</td>
<td>Occupation</td>
</tr>
<tr>
<td>Property crime</td>
<td></td>
</tr>
<tr>
<td>Middle-aged</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>Auto theft</td>
<td></td>
</tr>
<tr>
<td>Occupation</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C. The Relationship Between Variables and Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variables</strong></td>
</tr>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Sentence</td>
</tr>
<tr>
<td>Property crime</td>
</tr>
<tr>
<td>Occupation</td>
</tr>
</tbody>
</table>
the income variable: $27,124, $44,980, $76,000, and so forth.

The relationship between attributes and variables becomes more complicated as we try to explain things. Here’s a simple example involving two variables: type of defense attorney and sentence. For the sake of simplicity, let’s assume that the variable “defense attorney” has only two attributes: “private attorney” and “public defender.” Similarly, let’s give the variable “sentence” two attributes: “probation” and “prison.” Now let’s suppose that 90 percent of people represented by public defenders are sentenced to prison and the other 10 percent are sentenced to probation. And let’s suppose that 30 percent of people with private attorneys go to prison and the other 70 percent receive probation. This is shown visually in Figure 1.3A.

Figure 1.3A illustrates a relationship between the variables “defense attorney” and “sentence.” This relationship can be seen by the pairings of attributes on the two variables. There are two predominant pairings: (1) persons represented by private attorneys who are sentenced to probation and (2) persons represented by public defenders who are sentenced to prison. But there are two other useful ways of viewing that relationship.
First, imagine that we play a game in which we bet on your ability to guess whether a person is sentenced to prison or probation. We’ll pick the people one at a time (not telling you which ones we’ve picked), and you have to guess which sentence each person receives. We’ll do it for all 20 people in Figure 1.3A. Your best strategy in this case is to always guess prison because 12 out of the 20 people are categorized that way. Thus, you’ll get 12 right and 8 wrong, for a net success score of 4.

Now suppose that we pick a person from Figure 1.3A and we have to tell you whether the person has a private attorney or a public defender. Your best strategy now is to guess prison for each person with a public defender and probation for each person represented by a private attorney. If you follow that strategy, you will get 16 right and 4 wrong. Your improvement in guessing the sentence based on knowing the type of defense attorney illustrates what it means to say that the variables are related. You would have made a probabilistic statement based on some empirical observations about the relationship between type of lawyer and type of sentence.

Second, let’s consider how the 20 people would be distributed if type of defense attorney and sentence were unrelated. This is illustrated in Figure 1.3B. Notice that half the people have private attorneys and half have public defenders. Also notice that 12 of the 20 (60 percent) are sentenced to prison—6 who have private attorneys and 6 who have public defenders. The equal distribution of those sentenced to probation and those sentenced to prison no matter what type of defense attorney each person had allows us to conclude that the two variables are unrelated. Here, knowing what type of attorney a person had would not be of any value to you in guessing whether that person was sentenced to prison or probation.

Variables and Relationships

We will look more closely at the nature of the relationships between variables later in this book. For now, let’s consider some basic observations about variables and relationships that illustrate the logic of social scientific theories and their use in criminal justice research.

Theories describe relationships that might logically be expected among variables. This expectation often involves the notion of causation: A person’s attributes on one variable are expected to cause or encourage a particular attribute on another variable. In the example just given, having a private attorney or a public defender seemed to cause a person to be sentenced to probation or prison, respectively. Apparently, there is something about having a public defender that leads people to be sentenced to prison more often than if they are represented by a private attorney.

Type of defense attorney and sentence are examples of independent and dependent variables, respectively. These two concepts are implicit in causation, which is the focus of explanatory research. In this example, we assume that criminal sentences are determined or caused by something; the type of sentence depends on something and so is called the dependent variable. The dependent variable depends on an independent variable; in this case, sentence depends on type of defense attorney.

Notice, at the same time, that type of defense attorney might be found to depend on something else—our subjects’ employment status, for example. People who have full-time jobs are more likely to be represented by private attorneys than those who are unemployed. In this latter relationship, the type of attorney is the dependent variable, and the subject’s employment status is the independent variable. In cause-and-effect terms, the independent variable is the cause, and the dependent variable is the effect.

How does this relate to theory? Our discussion of Figure 1.3 involved the interpretation of data. We looked at the distribution of the 20 people in terms of the two variables. In constructing a theory, we form an expectation about the relationship between the two variables based on what we know about each. For example, we know that private attorneys tend
to be more experienced than public defenders. Many people fresh out of law school gain a few years of experience as public defenders before they enter private practice. Logically, then, we would expect the more experienced private attorneys to be better able to get more lenient sentences for their clients. We might explore this question directly by examining the relationship between attorney experience and sentence, perhaps comparing inexperienced public defenders with public defenders who have been working for a few years. Pursuing this line of reasoning, we could also compare experienced private attorneys with private attorneys fresh out of law school.

Notice that the theory has to do with the variables “defense attorney,” “sentence,” and “years of experience,” not with individual people per se. People are the carriers of those variables. We study the relationship between the variables by observing people. Ultimately, however, the theory is constructed in terms of variables. It describes the associations that might logically be expected to exist between particular attributes of different variables.

**Purposes of Research**

*We conduct criminal justice research to serve different purposes.*

Criminal justice research, of course, serves many purposes. Explaining associations between two or more variables is one of those purposes; others include exploration, description, and application. Although a given study can have several purposes, it is useful to examine them separately because each has different implications for other aspects of research design.

**Exploration**

Much research in criminal justice is conducted to explore a specific problem. A researcher or official may be interested in some crime or criminal justice policy issue about which little is known. Or perhaps an innovative approach to policing, court management, or corrections has been tried in some jurisdiction, and the researcher wishes to determine how common such practices are in other cities or states. An exploratory project might collect data on some measure to establish a baseline with which future changes will be compared.

For example, heightened concern with drug use might prompt efforts to estimate the level of drug abuse in the United States. How many people are arrested for drug sales or possession each year? How many high school seniors report using marijuana in the past week or the past month? How many hours per day do drug dealers work, and how much money do they make? These are examples of research questions intended to explore different aspects of the problem of drug abuse. Exploratory questions may also be formulated in connection with criminal justice responses to drug problems. How many cities have created special police or prosecutor task forces to crack down on drug sales? What sentences are imposed on major dealers or on casual users? How much money is spent on treatment for drug users? What options exist for treating different types of addiction?

Exploratory studies are also appropriate when a policy change is being considered. Stricter enforcement of laws and longer prison sentences were common policy responses to drug abuse in the 1980s and 1990s, and jails and prisons were soon filled with newly arrested and sentenced drug offenders. This prompted a search for alternatives to incarceration, such as home detention. One of the first questions public officials typically ask when they consider some new policy is, “How have other cities (or states) handled this problem?” Exploratory research in criminal justice can be simple or complex, using a variety of methods.

For example, a mayor seeking to learn about drug arrests in his or her city might simply phone the police chief and request a report. In contrast, estimating how many high school seniors have used marijuana requires more sophisticated survey methods. Since the early 1970s, the National Institute on Drug Abuse has conducted nationwide surveys of students regarding drug use.
Description

A key purpose of many criminal justice studies is to describe the scope of the crime problem or policy responses to the problem. A researcher or public official observes and then describes what was observed. Criminal justice observation and description, methods grounded in the social sciences, tend to be more accurate than the casual observations people may make about how much crime there is or how violent teenagers are today. Descriptive studies are often concerned with counting or documenting observations; exploratory studies focus more on developing a preliminary understanding about a new or unusual problem.

Descriptive studies are frequently conducted in criminal justice. The FBI has compiled Uniform Crime Reports (UCR) since the 1930s. UCR data are routinely reported in newspapers and widely interpreted as accurately describing crime in the United States. For example, 2004 UCR figures (Federal Bureau of Investigation, 2005) showed that Nevada had the highest rate of auto theft (969.5 per 100,000 residents) in the nation, and Vermont had the lowest (92.5 per 100,000 residents).

Because criminal justice policy in the United States is almost exclusively under the control of state and local governments, many descriptive studies collect and summarize information from local governments. For example, since 1850 the federal government has conducted an annual census of prisoners in state and local correctional facilities. Like the decennial U.S. census, it gathers basic characteristics of a population—in this case, the population of people in detention (jail or prison) and on probation or parole.

Descriptive studies in criminal justice have other uses. A researcher may attend meetings of neighborhood anticrime groups and observe their efforts to organize block watch committees. These observations form the basis for a case study that describes the activities of neighborhood anticrime groups. Such a descriptive study might present information that officials and residents of other cities can use to promote such organizations themselves. Or consider research by Bruce Jacobs and associates (Jacobs et al., 2003) in which they describe the motivations for carjacking, how offenders select targets, and how carjacking differs from other types of robbery.

Explanation

A third general purpose of criminal justice research is to explain things. Recall our earlier example, in which we sought to explain the relationship between type of attorney and sentence length. Reporting that urban residents have generally favorable attitudes toward police is a descriptive activity, but reporting why some people believe that police are doing a good job while other people do not is an explanatory activity. Similarly, reporting why Nevada has the highest auto theft rate in the nation is explanation; simply reporting auto theft rates for different states is description. A researcher has an explanatory purpose if he or she wishes to know why the number of 14-year-olds involved in gangs has increased, as opposed to simply describing changes in gang membership.

Application

Researchers also conduct criminal justice studies of an applied nature. Applied research stems from a need for specific facts and findings with policy implications. Another purpose of criminal justice research, therefore, is its application to public policy. We can distinguish two types of applied research: evaluation and policy analysis.

First, applied research is often used to evaluate the effects of specific criminal justice programs. Determining whether a program designed to reduce burglary actually had that intended effect is an example of evaluation. In its most basic form, evaluation involves comparing the goals of a program with the results. For example, if one goal of increased police foot patrol is to reduce fear of crime, then an evaluation of foot patrol might compare levels
of fear before and after increasing the number of police officers on the beat on foot. In most cases, evaluation research uses social science methods to test the results of some program or policy change. Because crime problems persist and seem to change frequently, officials are constantly seeking new approaches, and it is becoming more common for public officials or researchers to conduct evaluations of new programs.

The second type of applied research is policy analysis. What would happen to court backlogs if we designated a judge and prosecutor who would handle only drug-dealing cases? How many new police officers would have to be hired if a department shifted to community policing? These are examples of what if questions addressed by policy analysis. Answering such questions is sort of a counterpart to program evaluation. Policy analysis is different from other forms of criminal justice research primarily in its focus on future events. Rather than observing and analyzing current or past behavior, policy analysis tries to anticipate the future consequences of alternative actions.

Similarly, justice organizations are increasingly using techniques of problem analysis to study patterns of cases and devise appropriate responses. Problem-oriented policing is perhaps the best-known example, in which crime analysts work with police and other organizations to examine recurring problems. Ron Clarke and John Eck (2005) have prepared a comprehensive guide for this type of applied research.

Our brief discussion of distinct research purposes is not intended to imply that research purposes are mutually exclusive. Many criminal justice studies have elements of more than one purpose. Suppose you want to evaluate a new program to reduce bicycle theft at your university. First, you need some information that describes the problem of bicycle theft on campus. Let’s assume your evaluation finds that thefts from some campus locations have declined but that there was an increase in bikes stolen from racks outside dormitories. You might explain these findings by noting that bicycles parked outside dorms tend to be unused for longer periods and that there is more coming and going among bikes parked near classrooms. One option to further reduce thefts would be to install more secure bicycle racks. A policy analysis might compare the costs of installing the racks with the predicted savings resulting from a reduction in bike theft.

**Differing Avenues for Inquiry**

*Social science research is conducted in a variety of ways.*

There is no one way of doing criminal justice research. If there were, this would be a much shorter book. In fact, much of the power and potential of social science research lies in the many valid approaches it comprises.

Three broad and interrelated distinctions underlie many of the variations of social science research: (1) idiographic and nomothetic explanations, (2) inductive and deductive reasoning, and (3) quantitative and qualitative data. Although it is possible to see them as competing choices, a good researcher masters each of these orientations.

**Idiographic and Nomothetic Explanations**

All of us go through life explaining things; we do it every day. You explain why you did poorly or well on an exam, why your favorite team is winning or losing, and why you keep getting speeding tickets. In our everyday explanations, we engage in two distinct forms of causal reasoning—idiographic and nomothetic explanation—although we do not ordinarily distinguish them.

Sometimes, we attempt to explain a single situation exhaustively. Thus, for example, you may have done poorly on an exam because (1) you had forgotten there was an exam that day, (2) it was in your worst subject, (3) a traffic
jam caused you to be late to class, and (4) your roommate kept you up the night before with loud music. Given all these circumstances, it is no wonder that you did poorly on the exam.

This type of causal reasoning is idiographic explanation. *Idio* in this context means “unique, separate, peculiar, or distinct,” as in the word *idiosyncrasy*. When we complete an idiographic explanation, we feel that we fully understand the many causes of what happened in a particular instance. At the same time, the scope of our explanation is limited to the case at hand. Although parts of the idiographic explanation might apply to other situations, our intention is to explain one case fully.

Now consider a different kind of explanation. For example, every time you study with a group, you do better on an exam than if you study alone; or your favorite team does better at home than on the road; or you get more speeding tickets on weekends than during the week. This type of explanation—called nomothetic—seeks to explain a class of situations or events rather than a single one. Moreover, it seeks to explain “efficiently,” using only one or just a few explanatory factors. Finally, it settles for a partial rather than a full explanation of a type of situation.

In each of the preceding nomothetic examples, you might qualify your causal statements with phrases like “on the whole” or “usually.” Thus, you usually do better on exams when you’ve studied in a group, but there have been exceptions. Similarly, your team has won some games on the road and lost some at home. And last week you got a speeding ticket on the way to Tuesday’s chemistry class, but you did not get one over the weekend. Such exceptions are an acceptable price to pay for a broader range of overall explanation.

Both the idiographic and the nomothetic approaches to understanding can be useful in daily life. They are also powerful tools for criminal justice research. Thus, the researcher who seeks an exhaustive understanding of the inner workings of a particular juvenile gang or the rulings of a specific judge is engaging in idiographic research. The aim is to understand that particular group or individual as fully as possible.

Rick Brown and Ron Clarke (2004) sought to understand thefts of a particular model of Nissan trucks in the south of England. Most stolen trucks were never recovered. Their research led Brown and Clarke to a shipping yard where trucks were taken apart and shipped to ports in France and Nigeria as scrap metal. They later learned that trucks were reassembled and sold to individuals and small companies. In the course of their research, they linked most thieves in England, and most resellers abroad to legitimate shipping and scrap metal businesses. Even though Brown and Clarke sought answers to the idiosyncratic problem of stolen trucks in one region of England, they came to some tentative conclusions about loosely organized international theft rings.

Sometimes, however, the aim is a more generalized understanding, across a class of events, even though the level of understanding is inevitably more superficial. For example, researchers who seek to uncover the chief factors that lead to juvenile delinquency are pursuing a nomothetic inquiry. They might discover that children who frequently skip school are more likely to have records of delinquency than those who attend school regularly. This explanation would extend well beyond any single juvenile, but it would do so at the expense of a complete explanation.

In contrast to the idiographic study of Nissan truck theft, Pierre Tremblay and associates (2001) explored how a theory of offending helped explain different types of offender networks. Examining auto thefts over 25 years in Quebec, the authors concluded that different types of relationships were involved in different types of professional car theft. However, Tremblay and associates found that persons involved in legitimate car sales and repair businesses were key members in all networks. The researchers showed how complex relation-
ships among people involved in legitimate and illegitimate activities helped explain patterns of car theft over a quarter-century. This is an illustration of the nomothetic approach to understanding.

Thus, social scientists have access to two distinct logics of explanation. We can alternate between searching for broad, albeit less detailed, universals (nomothetic) and probing more deeply into more specific cases (idiographic). Both are good science, both are rewarding, and both can be fun.

**Inductive and Deductive Reasoning**

The distinction between inductive and deductive reasoning exists in daily life, as well as in criminal justice research. You might take two different routes to reach the conclusion that you do better on exams if you study with others. Suppose you find yourself puzzling, halfway through your college career, about why you do so well on exams sometimes and more poorly at other times. You list all the exams you’ve taken, noting how well you did on each. Then you try to recall any circumstances shared by all the good exams and by all the poor ones. Do you do better on multiple-choice exams or essay exams? Morning exams or afternoon exams? Exams in the natural sciences, the humanities, or the social sciences? After you studied alone or in a group? It suddenly occurs to you that you almost always do better on exams when you studied with others than when you studied alone. This is known as the inductive mode of inquiry.

**Inductive reasoning** (induction) moves from the specific to the general, from a set of particular observations to the discovery of a pattern that represents some degree of order among the varied events under examination. Notice, incidentally, that your discovery doesn’t necessarily tell you why the pattern exists—merely that it does.

There is the second, and very different, way you might reach the same conclusion about studying for exams. As you approach your first set of exams in college, you might wonder about the best ways to study. You might consider how much you should review the readings and how much you should focus on your class notes. Should you study at a measured pace over time or pull an all-nighter just before the exam? Among these musings, you might ask whether you should get together with other students in the class or study on your own. You decide to evaluate the pros and cons of both options. On the one hand, studying with others might not be as efficient because a lot of time might be spent on material you already know. Or the group might get distracted from studying. On the other hand, you can understand something even better when you’ve explained it to someone else. And other students might understand material that you’ve been having trouble with and reveal perspectives that might have escaped you.

So, you add up the pros and the cons and conclude, logically, that you’d benefit from studying with others. This seems reasonable to you in theory. To see whether it is true in practice, you test your idea by studying alone for half your exams and studying with others for half. This second approach is known as the deductive mode of inquiry.

**Deductive reasoning** (deduction) moves from the general to the specific. It moves from a pattern that might be logically or theoretically expected to observations that test whether the expected pattern actually occurs in the real world. Notice that deduction begins with “why” and moves to “whether,” while induction moves in the opposite direction.

Both inductive and deductive reasoning are valid avenues for criminal justice and other social science research. Moreover, they work together to provide ever more powerful and complete understandings.

**Quantitative and Qualitative Data**

Simply put, the distinction between quantitative and qualitative data is the distinction between numerical and non-numerical data.
When we say that someone is witty, we are making a qualitative assertion. When we say that that person has appeared three times in a local comedy club, we are adding quantitative evidence to our assertion.

Most observations are qualitative at the outset, whether it is our experience of someone’s sense of humor, the location of a pointer on a measuring scale, or a check mark entered in a questionnaire. None of these things is inherently numerical. But it is often useful to convert observations to a numerical form. Quantification often makes our observations more explicit, makes it easier to aggregate and summarize data, and opens up the possibility of statistical analyses, ranging from simple de-

PUTTING IT ALL TOGETHER

“Driving While Black”

Introduction

Racial profiling of drivers on the nation’s streets and highways became a prominent issue in the late 1990s. Concern was fueled by compelling stories of minority motorists stopped by police for minor traffic violations, then subjected to aggressive questioning, searches, and even arrest. One of the most highly publicized examples involved:

. . . Robert Wilkins, a Harvard Law School graduate and a public defender in Washington, D.C., who went to a family funeral in Ohio in May 1992. On the return trip, he and his aunt, uncle, and 29-year-old cousin rented a Cadillac for the trip home. His cousin was stopped for speeding in western Maryland while driving 60 miles per hour in a 55-mile-per-hour zone of the interstate. The group was forced to stand on the side of the interstate in the rain for an extended period while officers and drug-sniffing dogs searched their car. Nothing was found. Wilkins, represented by the ACLU, filed suit and received a settlement from the state of Maryland. (Ramirez, McDevitt, and Farrell, 2000:6)

What came to be known as “driving while black” generated a number of lawsuits and eventually legislation at the state and federal level.

The underlying question was whether police traffic stops targeted African American and other minority drivers in a discriminatory way. Because the U.S. Constitution prohibits law enforcement officers from discriminatory behavior based on race, allegations of racial profiling generated a number of legal challenges and court cases.

By their nature, lawsuits revolve around individual cases, such as that of Robert Wilkins, and whether evidence of discrimination was present in Wilkins’s encounter with the Maryland State Police. In this sense, court cases tend to seek idio
dographic explanations of what happened in an individual case. Eventually, however, social scientists became involved in trying to assess the scope of racial profiling and what sorts of things might be associated with the practice. And social scientists focused more on nomothetic explanations of what kinds of things accounted for more general patterns of police actions in traffic stops.

Racial profiling also offers examples of different types of errors in traditional human inquiry. You might recognize the role overgeneralization plays in most forms of prejudice. Selective observation is another example. If police believe minorities are more often involved in drug or weapons smuggling, they will selectively stop cars driven by minorities. Assuming the extreme case where police stop only minorities, they will only be able to detect possible weapons or drug offenses
among minority motorists. In that case race profiling might become a self-fulfilling prophecy.

This box is the first of a running example that appears throughout the book. We examine racial profiling for several reasons. First, it was a highly publicized issue that promoted policy action and research throughout the United States. Second, for most people a traffic stop is their most common experience of coming under suspicion of police. A much larger proportion of people have a contact with police through a traffic stop than a criminal arrest. Further, in the words of David Harris (1999) “. . . almost any black person any place in the country . . .” could describe a personal experience of what was believed to be a discriminatory traffic stop. Third, police and other public officials challenged accusations of discrimination, claiming any disproportionate traffic stops of black motorists reflected different rates of traffic violations. This raised questions about how to determine whether individual traffic stops or patterns of traffic stops were based on the race of drivers or on something else. In this way, measuring patterns of traffic stops and the reasons underlying those patterns became an important research topic.

Finally, researchers at Rutgers University conducted their own research on race profiling in New Jersey, a state that came to symbolize race profiling for many people. Maxfield collaborated with colleagues George Kelling and Carsten Andresen to conduct research on the New Jersey State Police (Maxfield and Andresen, 2002; Maxfield and Kelling, 2005). And Carsten Andresen completed his own dissertation research on this topic (Andresen, 2005). Thus, the running example reflects some direct experiences in the messy realities of criminal justice research, realities that are not usually depicted in published studies.

How Do We Know?
Let’s consider what we have covered so far and examine how these general issues in research are reflected in the questions about racial profiling on the nation’s highways.

■ What percentage of cars stopped have minority drivers (descriptive)? What affects police decisions to stop particular vehicles (explanatory)? What changes should be made in police practices regarding traffic stops (applied)?
■ How do we come to believe that discriminatory practices underlie patterns of police traffic stops? What evidence supports that belief? And what evidence supports claims that police actions are not affected by race or ethnicity?
■ In addition to those mentioned above, what errors of traditional human inquiry might be involved? What about ideology and politics?
■ Are particular theories available to guide our inquiry?
■ Quite a lot of research has been conducted on municipal policing and police actions with respect to crime, but state police and traffic enforcement have hardly been studied at all. Can findings from other police actions help understand traffic enforcement?

How might the problem of racial profiling illustrate other topics described in this chapter?

Quantification requires focusing our attention and specifying meaning. Suppose someone asks whether your friends tend to be older or younger than you. A quantitative answer seems easy. You think about how old each of your friends is, calculate an average, and see whether it is higher or lower than your own age. Case closed.

Or is it? Although we focused our attention on “older or younger” in terms of the number of years people have been alive, we might mean something different with that idea—for example, “maturity” or “worldliness.” Thus, your friends may tend to be a little younger
than you in age but to act more mature. Or we might have been thinking of how young or old your friends look or of the variation in their life experiences, their worldliness. All these other meanings are lost in the numerical calculation of average age.

In addition to greater detail, non-numerical observations seem to convey a greater richness of meaning than do quantified data. Think of the cliché “He is older than his years.” The meaning of that expression is lost in attempts to specify how much older. In this sense, the richness of meaning is partly a function of ambiguity. If the expression meant something to you when you read it, that meaning came from your own experiences, from people you have known who might fit the description of being “older than their years.” Two things are certain: (1) Your understanding of the expression is different from ours, and (2) you don’t know exactly what either of us means by the expression.

Earl Babbie has a young friend, Ray Zhang, who was responsible for communications at the 1989 freedom demonstrations at Tiananmen Square in Beijing. Following the army crackdown on the demonstrators, Ray fled south, was arrested, and then was released with orders to return to Beijing. Instead, he escaped from China and made his way to Paris. Eventually, he came to the United States, where he resumed the graduate studies he had to abandon in fleeing his homeland. Ray had to deal with the difficulties of enrolling in school without any transcripts from China, studying in a foreign language, and meeting his financial needs—on his own, thousands of miles from his family and friends. Ray still speaks of one day returning to China to build a system of democracy.

We would probably agree that Ray sounds like someone who is both worldly and “older than his years.” However, the brief description of Ray’s experiences, while it fleshes out the meaning of the phrase, still does not equip us to say how much older or even to compare Ray with someone else without the risk of disagreeing as to which one is more “worldly.”

This concept can be quantified to a certain extent, however. For example, we could make a list of life experiences that contribute to what we mean by “worldliness”:

- Getting married
- Getting divorced
- Having a parent die
- Being arrested
- Being fired from a job

We could quantify people’s worldliness by counting how many of these experiences they have had: the more such experiences, the more worldly we say they are. If we think that some experiences are more powerful than others, we can give those experiences more points than others. Once we decide on the specific experiences to be considered and the number of points each warrants, scoring people and comparing their worldliness is fairly straightforward.

To quantify a concept like “worldliness,” we must be explicit about what we mean. By focusing specifically on what we will include in our measurement of the concept, as we did here, we also exclude the other possible meanings. Inevitably, then, quantitative measures will be more superficial than qualitative descriptions. This is the trade-off.

What a dilemma! Which approach should we choose? Which is better? Which is more appropriate to criminal justice research?

The good news is that we don’t have to choose. In fact, by choosing to undertake a qualitative or quantitative study, researchers run the risk of artificially limiting the scope of their inquiry. Both qualitative and quantitative methods are useful and legitimate. And some research situations and topics require elements of both approaches.
Knowing Through Experience: Summing Up and Looking Ahead

Empirical research involves measurement and interpretation.

This chapter introduced the foundation of criminal justice research: empirical research, or learning through experience. Each avenue for inquiry—nomothetic or idiographic description, inductive or deductive reasoning, qualitative or quantitative data—is fundamentally empirical. It’s worth keeping that in mind as we examine the various forms criminal justice research can take.

It is also helpful to think of criminal justice research as organized around two basic activities: measurement and interpretation. Researchers measure aspects of reality and then draw conclusions about the meaning of what they have measured. All of us are observing all the time, but measurement refers to something more deliberate and rigorous. Part Two of this book describes ways of structuring observations to produce more deliberate, rigorous measures. Our ability to interpret observations in criminal justice research depends crucially on how those observations are structured. After deciding how to structure observations, we have to actually measure them. Whereas physical scientists sometimes use tensiometers, spectrophotographs, and other such equipment for measurement, criminal justice researchers use a variety of techniques, examined in Part Three.

The other key to criminal justice research is interpretation. Much of interpretation is based on data analysis, which is introduced in Part Four. More generally, however, interpretation is very much dependent on how observations are structured, a point we will encounter repeatedly.

When we put the pieces together—measurement and interpretation—we are in a position to describe, explain, or predict something. And that is what social science is all about.

✪ Main Points

- Knowledge of research methods is valuable to criminal justice professionals as consumers and producers of research.
- The study of research methods is the study of how we know what we know.
- Inquiry is a natural human activity for gaining an understanding of the world around us.
- Much of our knowledge is based on agreement rather than on direct experience.
- Tradition and authority are important sources of knowledge.
- Empirical research is based on experience and produces knowledge through systematic observation.
- In day-to-day inquiry, we often make mistakes. Science offers protection against such mistakes.
- Whereas people often observe inaccurately, science avoids such errors by making observation a careful and deliberate activity.
- Sometimes we jump to general conclusions on the basis of only a few observations. Scientists avoid overgeneralization through replication.
- Scientists avoid illogical reasoning by being as careful and deliberate in their thinking as they are in their observations.
- The scientific study of crime guards against, but does not prevent, ideological and political beliefs from influencing research findings.
- Social science involves three fundamental aspects: theory, data collection, and data analysis.
- Social scientific theory addresses what is, not what should be.
- Social scientists are interested in explaining aggregates, not individuals.
- Although social scientists observe people, they are primarily interested in discovering relationships that connect variables.
- Explanations may be idiographic or nomothetic.
- Data may be quantitative or qualitative.
- Theories may be inductive or deductive.
Key Terms
These terms are defined in the chapter where they are set in boldface and can also be found in the Glossary at the end of the book.

Aggregate Independent
Attribute variable
Deductive reasoning Inductive reasoning
Dependent variable Nomothetic
Empirical Replication
Idiographic Variable

Review Questions and Exercises
1. Review the common errors of personal inquiry discussed in this chapter. Find a newspaper or magazine article about crime that illustrates one or more of those errors. Discuss how a scientist would avoid making that error.
2. Briefly discuss examples of descriptive research and explanatory research about crime rates in a large city near your college or university.
3. Often, things we think are true and supported by considerable experience and evidence turn out not to be true, or at least not true with the certainty we expected. Criminal justice seems especially vulnerable to this phenomenon, perhaps because crime and criminal justice policy are so often the subjects of mass and popular media attention. If news stories, movies, and TV shows all point to growing gang- or drug-related violence, it is easy to assume that these are real problems identified by systematic study. Choose a criminal justice topic or claim that's currently prominent in news stories or entertainment. Consult a recent edition of the Sourcebook of Criminal Justice Statistics (citation below) for evidence to refute the claim.

Online Sessions for Chapter 1
Online Sessions present further examples of selected material covered in each chapter. Your instructor may choose to use these for in-class discussion (as I do with my students) or you can access them at home. To ensure that you always have access to live, correct links for the websites described here, we provide the necessary links on the companion website for Research Methods for Criminal Justice and Criminology at http://www.thomsonedu.com/criminaljustice/maxfield. Once at the companion website, select this specific chapter, click on “Chapter Resources,” then click on “Web Links.”

Theory
See the problem-solving guide by Ron Clarke and John Eck (2005) at the POP Center website. Module 8 (they don’t use chapters or pages) introduces some theoretical concepts that underlie environmental criminology. Then these principles are recommended as a basic framework for problem analysis. http://popcenter.org/learning/60steps/#

Purposes of Research
Description: Most publications on the Bureau of Justice Statistics (BJS) website are examples of descriptive research. Select a publication and summarize its stated research purpose. http://www.ojp.usdoj.gov/bjs/
Exploration: Some BJS publications present information about new or evolving problems. Sometimes you can find an example of exploratory research. The publication “Identity Theft, 2004” is an example. Browse through the publication and describe examples of exploration in the report. http://www.ojp.usdoj.gov/bjs/abstract/it04.htm
Explanation: The POP Center website contains a large selection of guides for specific police problems. Select guide one and find the section titled something like: “Factors contributing to . . .” In most cases, the dependent variable will be named in the title. Identify two independent variables that affect the problem. Briefly describe the ways in which these independent variables appear to operate. http://www.popcenter.org
Application: Later sections of POP Center problem guides describe examples of actions that have been shown to be effective (or not effective) in the past. These are examples of applied research and are usually keyed in some way to independent variables. Briefly describe one effective and one not-so-effective intervention that correspond with the independent variables you selected above.

Variables and Attributes
Select a BJS publication that presents descriptive research. Identify at least two variables and the attributes associated with those variables.

Finding and Assessing Information
The web is now an essential research tool. It’s difficult to think of a topic that cannot be readily examined through web-based research. However, the
vast number of websites can be uneven sources of information. Web blogs, for example, are often geared to particular ideologies and points of view. Most of us have been solicited through phishing attempts masquerading as legitimate websites. The University of California at Berkeley offers guidance on how to evaluate the quality and legitimacy of websites. Here you will also find a rating checklist for assessing individual sites. http://www.lib.berkeley.edu/teachinglib/guides/internet/evaluate.html

Additional Readings

Babbie, Earl, The Sociological Spirit (Belmont, CA: Wadsworth, 1994). The primer in some sociological points of view introduces some of the concepts commonly used in the social sciences.


Maguire, Kathleen, and Pastore, Ann L. (eds.), Sourcebook of Criminal Justice Statistics (Washington, DC: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, annual) (http://www.albany.edu/sourcebook/). For over 30 years, this annual publication has been a source of basic data on criminal justice. If you're not yet familiar with this compendium, Chapter 1 is a good place to start.

Chapter 2

Theory and Criminal Justice Research

We’ll see what distinguishes scientific theory from everyday reasoning and how the social scientific approach to criminal justice research is linked to theory. We’ll also lay a foundation for understanding the research techniques discussed throughout the rest of the book.

Introduction 31

The Creation of Social Science Theory 31

Terms Used in Theory Construction 31

The Traditional Model of Science 33

Two Logical Systems 37

A Deductive Illustration 37

An Inductive Illustration 39

Inductive Theory and Criminal Justice Policy 42

PUTTING IT ALL TOGETHER: THEORIES OF DISCRETION AND DISCRIMINATION 44

Theory, Research, and Public Policy 46

Criminological Evolution: Ecological Theory and Crime Prevention Policy 47

Main Points 48
Introduction

Criminal justice in particular, and human behavior in general, can be studied scientifically.

The evolution of social science has brought a greater emphasis on systematic explanation and a move away from the previous emphasis on description. For example, political scientists now focus on explaining political behavior rather than describing political institutions. The growth of such subfields as econometrics has had this effect in economics, as has historiography in history. And criminal justice and criminology have followed this same trend. Research on the causes of crime and the effects of criminal justice policy has supplemented a previous emphasis on describing strategies for police investigation or corrections management. Understandably, professionals trained and experienced in the more traditional orientations of these fields sometimes objected to the new directions.

This book is grounded in the position that human behavior can be subjected to scientific study as legitimately as can the physicist’s atoms or the biologist’s cells. The study of crime and criminal justice concentrates on particular types of human behavior and so is no less amenable to scientific methods. Our attention now turns to the overall logic of social scientific inquiry as it applies to criminology and criminal justice. That logic is fundamentally rooted in the use of theory to guide inquiry.

The Creation of Social Science Theory

Theory and observation go together in science, but sometimes theory precedes observation, and other times observation comes before theory.

In Chapter 1, we considered some of the foundations of social scientific research as a mode of inquiry involving both logic and observation. Now we'll examine the relationship between theory and research in more depth. First, we'll cover some of the vocabulary of theory construction as a foundation for the chapter. Then we'll consider an example of “the scientific method” as it is traditionally taught. Next we will take a closer look at two models for describing how theory and research work together in the practice of social science. Finally, we consider at some length the important links between theory, research, and policy.

Terms Used in Theory Construction

We have already used some of the commonly terms associated with social science theory. Most people have a general idea of what they mean in everyday language. In this chapter, however, we will examine their meanings more precisely to prepare for our later examination of the links between theory and criminal justice research.

Theory  A theory is a systematic explanation for the observed facts and laws that relate to a particular aspect of life—juvenile delinquency, for example, or perhaps social stratification or political revolution. Joseph Maxwell (2005:42) defines theory as “a set of concepts and the proposed relationships among these, a structure that is intended to represent or model something about the world.”

Jonathan Turner (1974:3) examined several elements of theory. We consider three of them briefly:

- **Concepts.** Turner (1974:5) called concepts the “basic building blocks of theory.” They are abstract elements that represent classes of phenomena within the field of study. The concepts relevant to a theory of juvenile delinquency, for example, would include “juvenile” and “delinquency” for starters. “Peer group”—the people you hang around with and identify with—would be another relevant concept. “Social disorganization” was a central concept in the theory of delinquency tested by Shaw and McKay. “School performance” might also be relevant to a theory of juvenile delinquency.
• **Variables.** A variable is a concept's empirical counterpart. Whereas concepts are in the domain of theory, variables can be observed and can take different values; they vary. Thus, variables require more specificity than concepts. For example, as a variable, contacts with criminal justice agencies might be specified as the number of times someone was arrested, convicted, or incarcerated.

• **Statements.** A theory comprises several types of statements. Principles or laws are one type. Another type of statement is the axiom, a fundamental assertion—taken to be true—on which the theory is grounded. Propositions are a third type of statement: conclusions drawn about the relationships among concepts, based on the logical interrelationships among the axioms. To clarify, an axiom is an assumption about reality, while a proposition expresses relationships among axioms. For example, in a theory of juvenile delinquency, we might begin with axioms such as “Everyone desires material comforts” and “The ability to obtain material comforts legally is greater for the upper class than for the working class.” These axioms might reasonably lead us to the proposition “Working-class youths are more likely to break the law to gain material comforts than upper-class youths are.”

### Objectivity and Subjectivity

We recognize that some things fall into the realm of attitudes, opinions, and subjective points of view. For example, the question of whether Bach or Beethoven was the better composer is a subjective matter, dependent on the experiences and tastes of the person who is making such a judgment. But the existence of this book in your hands is an objective matter, independent of your experience of it. **Objective** is typically defined as “independent of mind,” but our awareness of what might objectively exist comes to us through our minds. As a working principle, social scientists prefer the phrase [intersubjective agreement](#) instead of objectivity. If several of us agree that something exists, then we treat that thing as though it had objective existence. This view of intersubjective agreement as a working version of objectivity is consistent with our earlier consideration, in Chapter 1, of agreement reality.

### Hypotheses

A hypothesis is a specified expectation about empirical reality, derived from propositions. If we continue the present example, a theory might contain the hypothesis “Working-class youths have higher delinquency rates than upper-class youths do.” Such a hypothesis could then be tested through research, as we will see below, in an example drawing on social disorganization theory and delinquency.

### Paradigms

No one ever starts out with a completely clean slate to create a theory. When we mentioned juvenile delinquency as an example of a topic for theory construction, you probably already had some implicit ideas about it. If we had asked you to list some concepts relevant to a theory of juvenile delinquency, you would have been able to make suggestions. We might say that you already have a general point of view—a frame of reference or paradigm. A **paradigm** is a fundamental model or scheme that organizes our view of something.

This may strike you as uncomfortably similar to the definition of theory, and this is a natural point of confusion. The primary distinction is based on organization and structure. In Kenneth Hoover’s (1992:66) view, “A theory is a collection of hypotheses linked by some kind of logical framework.” A paradigm is a structured but broader way of viewing things that affects how we approach research problems. It may be helpful to think of a paradigm as a pair of tinted glasses; everything we look at reflects the tint, no matter how we structure our observations. A theory serves to structure more carefully what we see through the paradigm’s lenses, but what we see is still tinted.

Although a paradigm doesn’t necessarily answer important questions, it tells us where,
and often how, to look for the answers. And, as we’ll see repeatedly, where you look largely determines what you find. Thomas Kuhn (1996:37) describes the importance of paradigms this way:

One of the things a scientific community acquires with a paradigm is a criterion for choosing problems that, while the paradigm is taken for granted, can be assumed to have solutions. To a great extent these are the only problems that the community will admit as scientific or encourage its members to undertake. Other problems, including many that had previously been standard, are rejected as metaphysical, as the concern of another discipline, or sometimes as just too problematic to be worth the time. A paradigm can, for that matter, even insulate the community from those socially important problems that are not reducible to the puzzle form, because they cannot be stated in terms of the conceptual and instrumental tools the paradigm supplies.

Kuhn’s chief interest was in how science advances. Although some progress involves the slow, steady, and incremental improvement of established paradigms (Kuhn calls this “normal science”), he suggests that important scientific progress takes the form of paradigm shifts, as established, agreed-on paradigms are discarded in favor of new ones. Thus, for example, Newtonian physics was replaced by Einstein’s relativity in what Kuhn calls a “scientific revolution.”

We suggest comparing the concept of paradigms with the world of fashion. Just as styles of clothing go in and out of fashion, paradigms that influence criminal justice come and go. We hope some styles and paradigms are gone forever: Contemporary men are free from the powdered wigs of the eighteenth century, women no longer suffer whalebone corsets (good for whales, too), and Western governments have stopped amputating the hands of thieves. Some good ideas have stood the test of time: Summer shorts are blessings for women and men alike.

The Traditional Model of Science

Theory plays a critical role in the traditional model of science. The traditional model of science includes three main elements: theory, operationalization, and observation. We now consider each element in turn, drawing on a classic example of delinquency research to illustrate the interaction between elements.

Theory  According to the traditional model of science, the scientist begins with an interest in some aspect of the real world. Suppose we are interested in discovering some of the broad social factors that contribute to the concentration of crime in urban areas. What kinds of social problems and characteristics of urban areas are associated with higher levels of crime?

As we think about this question, we might come up with things like population density, crowded housing conditions, poverty, unemployment, limited economic opportunities, weakened family ties, and the absence of appropriate role models. These are some of the ideas that influenced the development of social disorganization theories of crime. In the early twentieth century, sociologists began to examine the social disruption produced by rapid population growth in the city of Chicago. What came to be known as the “Chicago School” of criminology began with more general studies of the roots of a broad spectrum of social problems.

Sociologists Ernest Burgess and Robert Park started by describing Chicago’s growth as a pattern of concentric zones (Burgess, 1925; Park and Burgess, 1921). At the core was the city’s central business district, or downtown area. As Chicago expanded, the city developed in a pattern of circles radiating from the core. Each area, from the downtown core to outlying suburbs, displayed particular patterns of land use. Commerce and industry were concentrated in downtown areas, while wealthy families occupied residential areas in outer circles.
Lower-income migrants to Chicago settled in transition zones between the core and more distant areas.

Transition zones were so labeled because of the continuous flow of people in and out of the areas. Higher-income residents tended to move outward to avoid the influx of industrial activity expanding from the city’s core area. Outward-moving families were replaced by lower-income immigrants from Europe (and later white and black migrants from depressed southern states) who came to the United States to pursue economic opportunity or escape political oppression. Persons moving into transition zones found housing that was deteriorating and crowded but close to the factory jobs that had attracted them to Chicago. As their economic situation improved, they moved outward and were replaced by new in-migrants, thus continuing the cycle of transition.

Burgess argued that these patterns of population movements and neighborhood conditions weakened families, social institutions, and other collective ties, thus producing a general phenomenon he called “social disorganization.” Several problems were believed to result from social disorganization, including higher rates of disease, mental illness, and crime. These ideas form the basis of the social disorganization theory of crime.

However, the kinds of social conditions Burgess described as existing in transition zones suggest other possible explanations for the higher rates of crime in these areas. For example, if early twentieth-century immigrants from Europe were concentrated in transition zones, such groups might have brought cultural differences to urban areas that contributed to crime and other social ills. In other words, crime in transition zones might have resulted from a concentration of individuals from particular cultural backgrounds in those areas, rather than being the consequence of social disorganization.

How should we settle the question of whether social disorganization or cultural differences accounted for the concentration of crime in transition zones? The second step in the traditional model of science helps us move toward an answer.

**Operationalization** Operationalization means specifying steps, procedures, or operations for actually identifying and measuring the variables we want to observe. In the present example, operationalization involves deciding how to measure things like crime, social disorganization, and cultural background.

While Park and Burgess were developing their general theories of urban growth and its social consequences, sociologists Clifford Shaw and Henry McKay began a series of studies that tested the social disorganization explanation for crime in transition zones. Their landmark study, originally published in 1942, described how important concepts were operationalized (Shaw and McKay, 1969).

First, they chose three indicators of delinquency to measure crime, each obtained from official city records: (1) alleged delinquents brought before juvenile court, (2) juveniles committed to institutions, and (3) alleged delinquents detained by police but not brought before a juvenile court.

Second, social disorganization was measured even less directly, as is often the case in social research. Indicators were constructed from census data and information provided by local agencies. These indicators included families on relief (an earlier version of welfare and unemployment payments), average monthly rent, home ownership, occupational status, and ethnicity. One measure of ethnicity was whether a head of household was born in another country, referred to as “foreign-born”; another was the percentage of black families.

These indicators made it possible to compare delinquency with social disorganization and ethnic background. But how could Shaw and McKay determine which of these two factors was more important in explaining delinquency? Recall the concept of transition zones:
People were continuously moving into and out of the neighborhoods between core commercial zones and more wealthy residential areas. The notion of transition implies population movement, and the mobility of urban residents in Chicago offered an answer. Shaw and McKay compared changes in delinquency rates over time, focusing on three periods: 1900–1906, 1917–23, and 1927–33. They also examined delinquency among different ethnic groups within each zone. On the one hand, they suggested, if cultural differences accounted for crime, then delinquency rates in outer zones should increase as people from different cultural backgrounds moved out of transition zones and into more stable residential areas. On the other hand, if social disorganization produced delinquency, then the pattern of lower rates in outer zones and higher rates in transition zones should remain stable despite the outward movement of ethnic groups.

Park and Burgess identified five concentric zones in Chicago: Zones I and II were the inner core and transition areas; farther out, zone III included the homes of working-class families, zone IV was a middle-income residential area, and zone V was a suburban or outer area. Within each zone were several distinct neighborhoods. Shaw and McKay divided the city into 140 areas of approximately one square mile each, so that several areas were included in each of the five zones. These 140 areas represented units of analysis.

**Observation** The final step in the traditional model of science involves actual observation—looking at the world and making measurements of what is seen. Having developed theoretical expectations and created a strategy for observation, we next look at the way things are. This step may involve conducting experiments, interviewing people, and visiting what we’re interested in and watching it. The observations may be structured around the testing of specific hypotheses, or the inquiry may be less structured. For Shaw and McKay, it meant poring through published statistics in search of data relevant to their operationalization and then grouping data on delinquency and other indicators according to 140 geographical areas.

Comparing delinquency between ethnic groups, they found that rates were higher for all groups in core areas and transition zones. Although there was wide variation within groups—foreign-born, black, and native white—similar rates for each group were found in similar types of zones. Comparing rates over time, Shaw and McKay (1969:162, 315) found that when foreign-born families moved to outer zones, delinquency rates remained lower in those areas than in inner zones. Summarizing these findings, they concluded:

> In the face of these facts it is difficult to sustain the contention that, by themselves, the factors of race, nativity, and nationality are vitally related to the problem of juvenile delinquency. It seems necessary to conclude, rather, that the significantly higher rates of delinquents found among the children of Negroes, the foreign born, and more recent immigrants are closely related to existing differences in their respective patterns of geographical distribution within the city.

Moreover, the fact that in Chicago the rates of delinquents for many years have remained relatively constant in the areas adjacent to centers of commerce and heavy industry, despite successive changes in the nativity and nationality composition of the population, supports emphatically the conclusion that the delinquency-producing factors are inherent in the community.

Although subsequent research has challenged some of the conclusions Shaw and McKay reached, social disorganization theory continues to influence theory and research in criminal justice and other fields. For example, sociologist William Julius Wilson (1987, 1996) argues that limited opportunities for upward
economic mobility and outward residential mobility contributed to a host of chronic social problems, including crime, among inner-city residents in the late twentieth century. Contemporary environmental theories of crime, which emphasize such concepts as urban form and its influence on human behavior, are rooted in the earlier work of Park and Burgess, and Shaw and McKay (see, for example, Brantingham and Jeffery, 1991).

Notice how Shaw and McKay’s research was based on a more general theory of the links between the conditions of urban neighborhoods and a broad class of social ills. They had certain expectations about the patterns they would find in the numbers if a particular theoretical explanation was accurate. Operationalizing delinquency as an example of a social problem, they found higher rates of delinquency in transition zones, regardless of ethnic or cultural composition. In themselves, of course, these particular data did not prove the case; they did not provide definitive evidence that social disorganization produced delinquency. These data, however, were part of a body of evidence that Shaw and McKay amassed.

We have considered this example at some length for three reasons. First, it illustrates the traditional model of scientific research, moving from theory, to operationalization, to observation. Second, it is more realistic than the hypothetical physics experiments you might recall from a science class. Finally, contemporary research in criminal justice continues to build on this theoretical view of the relationship between crime and the urban environment.

Figure 2.1 provides a schematic diagram of the traditional model of scientific inquiry. In it, we see the researcher beginning with an interest in something or an idea about it. Next comes the development of a theoretical understanding of how a number of concepts, represented by the letters A, B, C, and so on, may be related to each other. The theoretical consider-
lations result in a hypothesis, or an expectation about the way things should be in the world if the theoretical expectations are correct. The notation \( Y = f(X) \) is a conventional way of saying that \( Y \) (for example, delinquency) is a function of or is in some way caused by \( X \) (for example, substandard housing). At that level, however, \( X \) and \( Y \) have general rather than specific meanings.

In the operationalization process, general concepts are translated into specific indicators. Thus, the lowercase \( x \) is a concrete indicator of capital \( X \). As an example, census data on the number of housing units that lack indoor plumbing (\( x \)) are a concrete indicator of substandard housing (\( X \)). This operationalization process results in the formation of a testable hypothesis: For example, did the rate of delinquency among Italian immigrants actually decline as they moved from substandard housing in transition zones to better housing in outer residential areas? Observations aimed at finding out are part of what is typically called hypothesis testing.

This traditional model of science provides a clear and understandable guide to how we can study something carefully and logically. However theory and research in criminal justice are related in other ways.

Two Logical Systems

Theory in criminal justice is used deductively and inductively.

Chapter 1 introduced deductive and inductive ways of linking specific observations and generalizations. Here we consider how deductive and inductive reasoning relate to social science theory. W. I. B. Beveridge (1950:113), a philosopher of science, described these two systems of logic as follows:

Logicians distinguish between inductive reasoning (from particular instances to general principles, from facts to theories) and deductive reasoning (from the general to the particular, applying a theory to a particular case). In induction one starts from observed data and develops a generalization that explains the relationships between the objects observed. On the other hand, in deductive reasoning one starts from some general law and applies it to a particular instance.

The classic illustration of deductive logic is this familiar syllogism: “All people are mortal; Socrates is a person; therefore, Socrates is mortal.” This syllogism, or logical argument, presents a theory and its operationalization. To prove it, we could perform an empirical test of Socrates’ mortality. That is essentially the approach of the traditional model.

Using inductive logic, we might begin by noting that Socrates is mortal and observing a number of other people as well. We might then note that all the observed people are mortals, thereby arriving at the tentative conclusion that all people are mortal.

A Deductive Illustration

Why is there so much plea bargaining in criminal courts? What accounts for differences in conviction rates and sentence lengths among people convicted of similar offenses? These were among the questions addressed by James Eisenstein and Herbert Jacob (1977) in their classic study of court processing in three U.S. cities. Their research provides a good example of the deductive approach.

They began by identifying the shortcomings of popular views of criminal court outcomes. Differences in defendant and judicial characteristics and in the formal requirements of court procedure were not satisfactory explanations. Furthermore, they argued, previous studies of criminal courts had failed to disentangle the combined influence of differences in judges, lawyers, defendants, and court procedures.

Their novel approach was to view trial courts as organizations and to appeal to various
theories of organizational behavior to explain trial court outcomes. Courts in the United States are formal adversarial organizations, meaning that attorneys representing each side are supposed to do battle before a neutral judge. However, said Eisenstein and Jacob, even though prosecutors and defense lawyers are formal adversaries, they work together to achieve the common goal of processing defendants and so have incentives to cooperate with each other. Similarly, judges are supposed to play a neutral role, but they also have to manage their courtrooms and move along large numbers of cases in a timely fashion.

Viewing the judge, prosecutor, and defense attorney as a work group helps explain their mutual interests. In organization theory, work groups are small collections of people who share responsibility for certain tasks. If the work group is stable, with its members working together over time, they come to know one another and develop shared incentives for maintaining group cohesion and harmony. Among the shared goals of courtroom work groups are disposing of caseloads and reducing uncertainty by sharing information about cases and avoiding surprises. Of course, group members also want to do justice, but justice is an elusive concept; other goals are easier to identify and have more immediate relevance to the work group. Finally, like most people who must work together, judges, prosecutors, and defense attorneys find their jobs more pleasant and easier if they get along with one another.

Eisenstein and Jacob drew several hypotheses from this theory, most notably that plea bargaining would be more common in courtrooms with stable work groups. Let’s break this hypothesis down into parts to see how it deduces a specific prediction about trial courts from general organization theory. First, the longer people work together in a small group (stability), the better they will come to know one another and pursue common goals in a cooperative way. Second, members of the courtroom work group have a common goal of processing cases. Third, plea bargains are the products of negotiation. Therefore, stable courtroom work groups will more often process cases through negotiated guilty pleas than through formal trials.

Having framed this hypothesis, Eisenstein and Jacob gathered data from felony courts (those that handle more serious crimes) in Baltimore, Chicago, and Detroit. Their findings generally supported the work group hypothesis by comparing the characteristics of court organization with indicators of case outcomes. However, they found that certain other features of court organization also helped explain case processing.

Plea bargaining was common in Chicago and Detroit for slightly different reasons, each consistent with organizational features. Judges, prosecutors, and public defenders in Chicago were regularly assigned to specific courtrooms for extended periods. This, and the traditional politics of accommodation in Chicago, promoted plea bargaining, which was how 70 percent of convicted offenders were processed.

Detroit courts had less stable work groups but used a system of court administration that played a much stronger role in managing caseloads than in Chicago. Judges and prosecutors took active steps to promote the rapid settlement of cases. As a result, 82 percent of convicted defendants pleaded guilty in plea bargains. The more aggressive management style also resulted in faster processing of cases than in Chicago.

The situation was different in Baltimore. Court management was haphazard, judges were regularly rotated to different courtrooms, and there was a great deal of turnover among prosecutors. As a result, courtroom work groups were less stable, and the informal goals of negotiation gave way to more formal processing of felony cases—only 45 percent were settled by plea bargains. Eisenstein and Jacob also pointed out that in Baltimore, where neither informal work group norms nor formal court management were very strong, there tended to be greater dif-
ferences in sentences between judges. Where formal and informal organizational controls were weaker, individual differences in sentencing became more common.

This example clearly illustrates the deductive model. Beginning with general, theoretical expectations about the impact of organizational characteristics on court operations, Eisenstein and Jacob derived concrete hypotheses that linked specific measurable variables. The actual empirical data could then be analyzed to determine whether the deductive expectations were supported by empirical reality.

An Inductive Illustration

Often, social scientists begin constructing a theory by observing aspects of social life, seeking to discover patterns that may point to more or less universal principles. Barney Glaser and Anselm Strauss (1967) coined the term grounded theory to describe this inductive method of theory construction. Field research—the direct observation of events in progress—is frequently used to develop theories. Or research drawing on data from surveys might reveal patterns of results that suggest particular theoretical explanations. Here’s an example.

Criminal justice researchers have increasingly recognized that much crime is concentrated—among offenders, places, and even victims. A landmark study of youths in Philadelphia underscored the concentration of delinquency among a relatively small number of youths (Wolfgang, Figlio, and Sellin, 1972). Researchers have also formulated the concept of hot spots, relatively small places where a number of crimes are concentrated (Sherman, Gartin, and Buerger, 1989). Research done mostly in England discovered that some individual people or places are repeatedly victimized by crime. For example, Ken Pease reports that about 2 percent of victims of property crime in England account for over 40 percent of property crime victimizations (Pease 1998:3). In fact, Pease and Laycock (1996) link the ideas of hot spots and repeat victims by arguing that hot spots often emerge because individual people or buildings are repeatedly targeted. In this way, they described repeat victims as “hot dots,” and the clustering of repeat victims is what produces hot spots. Figure 2.2 shows an example.

Townsley, Homel, and Chaseling (2003) carried this one step further in wondering whether some burglaries in hot spot areas might be conceived as “near repeat victimizations.” In doing so they drew on their understanding of two different mechanisms for repeat burglary victimization. First, people and place may be repeatedly victimized because of inherently unequal risks—referred to as risk heterogeneity. Houses located in areas where burglars live are at greater
risk than targets located in other areas. Houses screened by dense vegetation are at greater risk because burglars can work unobserved.

The second explanation for repeat victimization assumes that offenders learn something about individual dwellings from breaking into them. After burglars break into a particular house or apartment they know how to break into it later. When offenders learn how to get into a dwelling, it reduces some element of uncertainty that exists if they were to select a target about which less was known. Burglars also know what’s available to be taken at a later time. This mechanism of repeat burglary is known as event dependency, where the event of a later burglary is partly dependent on what was learned during the first.

The idea of a near-repeat burglary recognizes how each of these two mechanisms might contribute to a burglary hot spot. Risk heterogeneity means that homes in areas where burglars operate are at greater risk. So if a burglary occurs a couple of doors down from an earlier burglary, that’s a near-repeat that occurs because burglars are working in a particular neighborhood. With respect to event dependency, Townsley and associates proposed that burglars learn something about an area in the course of breaking into one home. That means dwellings near those already targeted are at greater risk because offenders have learned something about features of the area that offer clues for selecting additional targets. Burglars may learn something about the street layout,
After talking with area residents and business owners and thinking about the problem, the community prosecutor reasoned that traditional approaches like stepped-up patrol, police visibility, and arrests might work for a time. The increased police presence could not be maintained over a long period, however, and problems would no doubt reappear after police moved on to other areas.

Instead, the prosecutor arranged an initial police sweep to disperse campers from the Gulch. This was accompanied by a city-sponsored clean-up that removed tents, derelict structures, and countless truckloads of debris. The next step was a cooperative effort among police and area residents. Signs reading “No Camping” on one side and listing nearby homeless shelters on the reverse were posted in the area. Citizens were enlisted to patrol the Gulch on a regular basis and to notify police about the appearance of any campsites. Encouraged by the support they had received from police and the prosecutor’s office, area residents were diligent in watching for campers and dissuading new arrivals from setting up camp.


Listening to the concerns of residents and business owners near Sullivan’s Gulch and observing the area firsthand, the neighborhood prosecutor “built” a theory. His theory was based on street-level experience, experience that differed from the prevailing suite-level perspective of traditional prosecution. The theory was inductive, grounded in observations that were then used to formulate more general statements about the links between disorder and serious crime in an area of Portland. And these links suggested a theory of action that guided justice policy.

*Source:* Adapted from Boland (1996:36–37).

The researchers examined police burglary reports from five urban and suburban areas in Queensland, Australia. As it happens, police records in Australia and the United Kingdom are more uniform and much more detailed than police records for most areas in the United States. Repeat burglaries were defined as those addresses that had two or more burglaries over 34 months. Near-repeats, identified through a fairly complex statistical estimation process, were burglaries that occurred within about 200 yards of a previous burglary over the 34-month study period.

Next, Townsley and associates enlisted the help of a real estate agent to learn something about the nature of residences in their study area. This was an important part of a research strategy to learn how physical features of areas were related to burglary rates. They distinguished newer suburban developments with similar housing from more established areas that tended to have older, more varied housing. The agent also helped researchers estimate the market value and age of houses in different neighborhoods, and whether or not areas were large-scale developments.

Analyzing the official police data, Townsley and associates found repeat victimization was more common in areas with older homes, where the types and features of houses varied quite a lot. This was consistent with previous research that suggested event dependency was
the mechanism. Once burglars had targeted an older home, subsequent burglaries were more like to strike the same dwelling, presumably because offenders knew how to break in and what they would find.

Findings were different in newer, larger developments. Here researchers found higher levels of near repeat burglaries. Burglars were more likely to strike houses close to a home previously victimized than they were to return to the original target.

Considering this pattern of results, Townsley and associates proposed a mechanism of contagion for understanding repeat and near-repeat burglary:

One way to think of the relationship between housing diversity and target vulnerability is through the infection analogy. . . . Housing diversity is an indicator of how contagious burglary victimization is, in that similar areas allow transmission of burglaries, whereas diverse areas restrict transmission. (2003:630)

Just as infectious diseases spread when people are packed together, as on an airplane or bus, burglary moves through a neighborhood made up of many newer homes of similar design. Once burglars have learned how to break into one house, they can more easily strike another nearby house of similar design. But in older neighborhoods where houses are more varied, burglars tend to return to the same target.

The researchers used inductive reasoning to develop their contagion theory of burglary. Examining patterns of offenses across different types of neighborhoods led them to compare the spread of crime to the spread of disease. Finally, the researchers link their findings to theoretical foundations of environmental criminology, emphasizing the central role of place. Just as Burgess, Shaw and McCay found delinquency rates linked to urban zones, Townsley, Homel, and Chaseling now propose that features of housing settlement patterns affect how burglaries spread through a community.

**Inductive Theory and Criminal Justice Policy**

Consider a different way of looking at inductive theory building, suggested by Barbara Hart, legal director of the Pennsylvania Coalition Against Domestic Violence. Speaking on the topic of how to increase the utilization of criminal justice research by justice professionals, Hart argued that the development of grounded theory was commonplace among many practitioners. By this she meant that probation officers, judges, case managers, victim services counselors, and others tend to formulate general explanations to fit the patterns they observe in individual cases. As Hart put it: “Practitioners engage in theory building as they interpret experience.” Others refer to a “theory of action” that embraces experience-based theories that guide public officials (Campbell, 1979; Clarke, 1997).

A theory of action may then guide the future actions of justice professionals, until they discover cases or patterns of cases that don’t fit their theory. See the box titled “Grounded Theory and Community Prosecution” for an example.

**A Graphic Contrast** Figure 2.3 shows a graphic comparison of the deductive and inductive methods. In both cases, let’s assume we are interested in the relationship between the number of hours spent studying for an exam and the grade earned on that exam. Using the deductive method, we begin by examining the matter logically. Doing well on an exam reflects a student’s ability to recall and to manipulate information. Both of these abilities should be increased by exposure to the infor-

---

mation before the exam. In this fashion, we arrive at a hypothesis suggesting a positive relationship between the number of hours spent studying and the grade earned on the exam. We say *positive* because we expect grades to increase as the hours of studying increase. If increased hours produced lower grades, that would be called a negative relationship. The hypothesis is represented by the graph in part 1(a) of Figure 2.3.

Our next step, using the deductive method, is to make observations relevant to testing our hypothesis. The shaded area in part 1(b) of the figure represents perhaps hundreds of observations of different students, noting how many hours they studied and what grades they got. Finally, in part 1(c) of the figure, we compare the hypothesis and the observations. Because observations in the real world seldom (if ever) match our expectations perfectly, we must
decide whether the match is close enough to consider the hypothesis confirmed. Put differently, can we conclude that the hypothesis describes the general pattern that exists, granting some variations in real life?

Now let’s address the same research question using the inductive method. In this case, we begin—as in part 2(a) of the figure—with a set of observations. Curious about the relationship between hours spent studying and grades earned, we simply arrange to collect some relevant data. Then we look for a pattern that best represents or summarizes our observations. In part 2(b) of the figure, the pattern is shown as a curved line running through the center of the curving mass of points.

The pattern found among the points in this case—and shown in part 2(c) of the figure—suggests that with 1–15 hours of studying each additional hour generally produces a higher grade on the exam. Studying 15–25 hours actually seems to slightly lower the grade. Studying
research might be structured to test those theories. Here are two brief examples:

- **Behavior is the product of underlying attitudes.** Rooted in psychological theory, this perspective assumes that police stop more minority drivers because of underlying prejudices. Testing this would require measuring these underlying attitudes by administering a standard questionnaire.

- **Organizational behavior and myths.** Organization theory would claim that the actions of individual police in traffic stops are the result of organization policies. As an explanation of racial profiling, this perspective would mean that police managers encourage individual police to stop minority drivers. Further, Crank and Langworthy (1992) describe how organizations have traditions, myths, and beliefs that can affect actions by individuals. If organization beliefs assume that minorities are more often involved in drug or weapons violations, those beliefs can influence actions by police in traffic enforcement.

### Traffic Enforcement Is Different

In his work on the New Jersey State Police, Carsten Andresen (2005) drew on two bodies of theory to guide his two basic research questions: (1) what influences police decisions to stop a particular vehicle, and (2) what influences the actions police take after making a stop.

Andresen adapted a theory of police discretion to address the first question: In the absence of clear directives, police can choose which cars to stop based on whatever criteria they wish to use. He tested this explanation by interviewing troopers and observing their actions while accompanying troopers on a large number of routine patrols.

A synthesis of theories rooted in the psychology of interactions guided Andresen’s research on the second question. He reasoned that a traffic stop is a transaction between police and drivers. Like other transactions, the parties involved in a traffic stop negotiate, though not as obviously as we might negotiate when buying some item at a garage sale. Troopers sought some concession of fault or contrition from the driver. In exchange, they gave cooperative drivers a break by not issuing a ticket or by writing a ticket for a lesser traffic violation. Negotiations broke down if drivers were aggressive, denied traffic violation, or complained that everyone else was speeding too; these drivers usually received a traffic citation.

Andresen used theory in two ways. Observing patterns of “bargaining” between police and drivers, he began to form some more general explanation through inductive reasoning. Then he consulted research literature in other areas, drawing on the organizational psychology theory of reciprocity to “deduce” an explanation for the patterns he observed but had been able to only partly explain.

more than 25 hours, however, results in a return to the initial pattern: More hours produce higher grades. Using the inductive method, then, we end up with a tentative conclusion about the relationship between the two variables. The conclusion is tentative because the observations we have made cannot be taken as a test of the pattern; those observations are the source of the pattern we’ve created.

In summary, the scientific norm of logical reasoning provides a bridge between theory and research—a two-way bridge. Scientific inquiry in practice typically involves an alternation between deduction and induction. During the deductive phase, we reason toward observations; during the inductive phase, we reason from observations. Both logic and observation are essential. In practice, both deduction and induction are routes to the construction of social theories. This chapter’s installment of our running example illustrates how theory has helped researchers formulate and test
explanations for patterns of traffic stops. Conversely, some researchers have argued that the lack of theory to guide research has been a problem in studies of racial profiling.

**Theory, Research, and Public Policy**

Theory in criminal justice guides both research and public policy.

Because crime is an important social problem, and not simply a social artifact of interest to researchers, much research in criminal justice is closely linked to public policy. It is interesting and important to determine why some people commit crimes while others do not, but such research takes on added meaning when we consider how new knowledge can be used to formulate policy. Research on criminal justice policy is an example of applied research; research results are applied to specific questions about how government officials and the general public should respond to crime.

Criminal justice theory is just as important in structuring applied research questions as it is in directing basic research. Theory, research, and criminal justice policy are linked in two ways. First, theory is used to guide basic research. The results from this research may point to specific policies. For example, a presiding judge in Phoenix who wishes to process cases faster would be wise to learn about empirical research on trial courts. Eisenstein and Jacob (1977) found that court administrators in Detroit scheduled meetings between prosecutors and defense attorneys in which plea bargains were discussed, and they required that reports of such meetings be forwarded to judges. The Phoenix presiding judge might then direct all judges to require pretrial meetings between prosecutors and defense attorneys to negotiate the terms of plea bargains. Similarly, Nancy Merritt examined how this theoretical perspective on trial courts as organizations helped explain changes in sentencing practices after “get-tough” legislation in Oregon (Merritt, 2006).

To understand the second way criminal justice theory and policy are related, consider the similarities between a hypothesis and a specific program. At the beginning of this chapter, we said described hypotheses as specified expectations about empirical reality. This is also true of public policy programs: They are specified expectations about what empirical reality will result from a particular policy action.

Theories of social disorganization include “if-then” statements such as “If an urban neighborhood exhibits signs of social disorder, then crime and delinquency in that neighborhood will be higher compared with neighborhoods having fewer signs of social disorder.” This implies a public policy if-then statement: “If we implement a new program to reduce social disorder in an urban neighborhood, then crime and delinquency will decline.” Complementary if-then statements such as these underlie the rationale for community policing and the actions taken by Boland’s community prosecutor (see the box earlier in this chapter). With community-based crime control strategies, police and prosecutors take action against neighborhood problems of social disorder that may contribute to crime and delinquency.

So theory, research, and policy are related in two similar ways: (1) Theory structures research, which, in turn, is consulted to develop policy, and (2) policies take the form of if-then statements, which implies that they are subject to empirical tests. This second type of link between theory and policy forms the basis for evaluation research—studies that assess whether some program is achieving its intended goals. A program is simply a hypothesis that some specific action will produce some specific result. A program may therefore be empirically evaluated in the same way a research hypothesis can be empirically evaluated. Our
discussion of evaluation research in Chapter 12
draws extensively on the parallel logic of theory
and public policy.

Criminological Evolution:
Ecological Theory and
Crime Prevention Policy

We conclude this chapter by offering an exam-
ple of how theory, research, and policy are in-
tertwined. Further, we believe after reading the
final section of this chapter you will be able to
trace principles of contemporary criminologi-
cal theory to foundations from the Chicago
School, discussed earlier.

In his introduction to a collection of case
studies on situational crime prevention, Ron-
ald Clarke (1997) describes how this approach
to criminal justice policy evolved from more
general ecological theories of crime. The word
ecological is significant: Just as humans and
other organisms are affected by environmen-
tal forces in nature (weather, global climate
change, water quality), human behavior is par-
tially a function of the physical and social en-
vironment in which that behavior takes place.
Environmental criminologists link criminal
behavior to physical and social space. As de-
scribed by Paul and Patricia Brantingham
(1991:12–13), applying principles of social ecol-
yogy to criminology has two components. First,
the physical features of a city—urban form—
are partly the results of conflict over a scarce
resource, space. Cities are densely populated
because land costs are high and many people
want to locate near workplaces, stores, and en-
tertainment venues. Second, human behavior
is affected by physical form. People behave dif-
ferently in the downtown areas of large cities,
for example, than they do in suburban shop-
ing malls (Wikström, 1995).

Different urban forms present greater or
lesser opportunities for crime by bringing po-
tential offenders and victims close to each
other. For example, the Brantinghams (1991:
49) describe how dispersed shopping areas and
strip malls are especially vulnerable to high
rates of property crime. This is because such
shopping areas are designed to provide con-
venient access to large numbers of people in
automobiles. And just like shoppers, property
offenders value ease of access and anonymity
and are therefore attracted to dispersed shop-
ing centers and strip developments in large
metropolitan areas.

Notice how these fundamental principles of
ecological theory are consistent with social dis-
organization theory: Crime is more common
in urban transition zones where the physical
environment (crowded housing) is unpleasant
and the social environment (poverty, unem-
ployment, disease) is undesirable.

Ecological theories rooted in research by
Shaw and McKay eventually affected the think-
ing of architects and city planners, whose work
focuses on the design of buildings, streets,
parks, and other physical features of the urban
landscape. One of the most influential works
that blended concepts of urban design and
criminal justice was Oscar Newman’s (1972)
book Defensible Space. Newman argued that ur-
ban housing should be constructed to enhance
the ability of residents to monitor and con-
trol their environment. Too often, the design
of large public housing projects in U.S. cities
made it difficult for residents to distinguish
threatening strangers from neighbors going
about their business.

Around the same time Newman’s book ap-
peared, similar ideas about the link between
urban design, human behavior, and crime were
influencing the development of crime preven-
tion policy in England. Drawing on the eco-
nomics paradigm of rational choice, research-
ers in the British Home Office Research Unit
extended Newman’s ideas about how the phy-
sical environment affects potential victims
(be they people or places), proposing that de-
cisions made by criminals are also influenced
by urban design (Clarke and Mayhew, 1980; Cornish and Clarke, 1986). Offenders may not formally weigh the costs and benefits of committing a crime, but Cornish and Clarke assert that offenders do evaluate potential targets and make choices based at least in part on opportunities.

Such reasoning led Clarke and others to propose crime prevention policies that try to reduce the opportunities for crime. This approach is called “situational crime prevention,” defined by Clarke (1997:4) as “opportunity reducing measures that are (1) directed at highly specific forms of crime (2) that involve the management, design, or manipulation of the immediate environment in as systematic and permanent a way as possible (3) so as to increase the effort and risks of crime and reduce the rewards as perceived by a wide range of offenders.” Clarke describes several examples of situational crime prevention techniques. The effort required to steal cars can be increased by the use of steering wheel locks; rewards of theft from automobiles can be reduced if car owners install removable stereo systems; the risks (to thieves) of auto theft can be increased by more thoughtful design of parking lots.

This example illustrates the deductive approach to moving from theory to policy development. Situational crime prevention policies were deduced, over a period of years, from ecological theories that described the relationships between crime and the physical and social environment of urban areas.

Now consider how Clarke’s description of the steps involved in designing specific situational crime prevention programs illustrates the inductive approach to policy development (1997:5):

1. Collect data about the nature and dimensions of the specific crime problem;
2. analysis of the situational conditions that . . . facilitate the commission of the crimes in question;
3. systematic study of possible means of blocking opportunities for these particular crimes, including analysis of costs;
4. implementation of the most promising, feasible and economic measures; and
5. monitoring of results and dissemination of experience.

Steps 1 and 2 involve collecting data and searching for patterns; steps 3 and 4 are equivalent to formulating a tentative conclusion and operating hypothesis based on the observed patterns. Step 5 is the applied research counterpart to hypothesis testing: Monitor and evaluate results, and then report whether the program achieved its intended objectives.

Throughout this chapter, we have seen various aspects of the links between theory and research in criminal justice inquiry. In the deductive model, research is used to test theories. In the inductive model, theories are developed from the analysis of research data. Although we focused on two logical models for linking theory and research, criminal justice researchers have developed a great many variations on these themes. Sometimes theoretical issues are introduced merely as a background for empirical analyses. Other studies cite selected empirical data to bolster theoretical arguments.

The ground we covered in this chapter reinforces our claim that there is no simple recipe for conducting criminal justice research. It is far more open-ended than the traditional view of science would suggest. Ultimately, science rests on two pillars: logic and observation. As we’ll see throughout this book, they can be fit together in many patterns.

**Main Points**

- Although people speak of science as being “objective,” it’s more a matter of intersubjective agreement—different scientists agree on their observations and conclusions.
- The traditional image of science includes theory, operationalization, and observation.
• Social scientific theory and research are linked through two logical methods: (1) Deduction involves the derivation of expectations or hypotheses from theories, and (2) induction involves the development of generalizations from specific observations.
• Science is a process involving an alternation of deduction and induction.
• In grounded theory, observations contribute to theory development.
• Theory in criminology and criminal justice is frequently adapted from other disciplines in the social and, less often, natural sciences.
• Criminal justice theory, research, and policy are linked in two ways: (1) Theory influences basic research, which may suggest new policy developments, and (2) policies are formulated like hypotheses and may therefore be subject to empirical test.

✪ Key Terms
Grounded theory  Objectivity
Hypothesis  Paradigm
Hypothesis testing  Theory
Intersubjective agreement

✪ Review Questions and Exercises
1. In his book Crime and Everyday Life, Marcus Fel-son expands on “routine activity theory” as an explanation for why crime occurs. According to this theory, crimes happen when three elements come together: (a) a motivated offender, (b) an attractive victim, and (c) absence of capable guardians. Select one property crime and one violent crime. Apply routine activity theory to explain examples of each. Are you using theory in an inductive or deductive way?
2. Review the relationships between theory and research discussed in this chapter. Select a research article from an academic journal, and classify the relationship between theory and research you find there.
3. A study by Clarke and Mayhew (1988) examined change in suicide rates in England following the reformulation of household gas used for cooking and heating. Beginning in the 1960s, household gas was “detoxified”—its carbon monoxide content was sharply reduced to make it less toxic when inhaled. Based on your understanding of theory as discussed in this chapter: (1) Do you think suicide rates would have changed? How? (2) Explain what mechanisms you think might be at work.

✪ Online Sessions for Chapter 2
Online Sessions present further examples of selected material covered in each chapter. Your instructor may choose to use these for in-class discussion (as I do with my students) or you can access them at home. To ensure that you always have access to live, correct links for the websites described here, we provide the necessary links on the companion website for Research Methods for Criminal Justice and Criminology at http://www.thomsonedu.com/criminaljustice/maxfield. Once at the companion website, select this specific chapter, click on “Chapter Resources,” then click on “Web Links.”

Critical Theory
The American Society of Criminology Division of Critical Criminology is a resource for those interested in a class of theories associated with leftist ideology. Critical criminology treats crime as largely a political construct, and criminal justice agencies are viewed as instruments for maintaining political control. Critical criminology theories guide research and its interpretation. http://www.critcrim.org/

Routine Activity Theory
To learn more about routine activity theory, visit this website, maintained by a crime prevention unit in the British Home Office. You’ll find links to a few publications that draw heavily on routine activity theory. http://www.crimereduction.gov.uk/skills08.htm

Environmental Criminology
Several sites offer resources based in environmental criminology. The Crime Prevention Service at Rutgers University is one example. http://crimeprevention.rutgers.edu/index.htm

Principles associated with crime prevention through environmental design (CPTED) underlie services available from the Vancouver Design Centre for CPTED. http://www.designcentreforcpted.org/Pages/EnvCrim.html

One of the most fascinating spin-offs from CPTED is the Design Against Crime (DAC) Research Centre, based in England. Here you will find information and links about a variety of research projects and consumer goods featuring
design-based crime prevention. For example, check out the chair designed to protect against the theft of handbags. Or visit the karrysafe.com site, where you can buy the high-security Karryfront screamer bag that provides: “Secure portable storage for all makes of laptop and accessories. Inbuilt anti-attack alarm will automatically start screaming if the bag is removed by force from your possession.” [http://www.designagainstcrime.com/index.html](http://www.designagainstcrime.com/index.html)

**Additional Readings**

Clarke, Ronald V., “Situational Crime Prevention,” in Michael Tonry and David Farrington (eds.), *Crime and Justice: An Annual Review of Research* (Chicago: University of Chicago Press, 1995), pp. 91–150. In this essay, Clarke lays out the genesis of situational crime prevention, linking it to past and evolving theories of crime and delinquency. This is an excellent example of synthesizing theory, research, crime prevention policy, and research methods.

Felson, Marcus, *Crime and Everyday Life*, 3rd ed. (Thousand Oaks, CA: Sage, 2002). Felson begins with a simple theory that has nonetheless been embraced by many researchers for its apparent power. Then he deduces explanations for different types of crime. Some explanations make more sense than do others; in fact, some explanations border on the strange. Figuring out which is which is fun, and not that difficult.

Gladwell, Malcolm, *The Tipping Point* (Boston: Little, Brown, 2000). Increases and decreases in crime are examples of social epidemics, the focus of Gladwell’s fascinating book. What sorts of social and other forces come together to produce large changes? Gladwell includes an extensive and provocative discussion of crime in New York City. Compare what we have to say about paradigms in this chapter with Gladwell’s discussion of tipping points. You can find some basic information on tipping points at Gladwell’s website: [http://www.gladwell.com](http://www.gladwell.com).

Gottfredson, Michael R., and Hirschi, Travis, *A General Theory of Crime* (Stanford, CA: Stanford University Press, 1990). This book presents a good example of a comprehensive theory, in which the authors argue that crime should be viewed as one type of human behavior that has much in common with other types of behavior. Along the way, the authors comment on how theories from different social and natural science disciplines contribute to theory in criminology and criminal justice.

Kuhn, Thomas, *The Structure of Scientific Revolutions*, 3rd ed. (Chicago: University of Chicago Press, 1996). This book is an exciting and innovative recasting of the nature of scientific development. Kuhn disputes the notion of gradual change and modification in science, arguing instead that established “paradigms” tend to persist until the weight of contradictory evidence brings about their rejection and replacement by new paradigms. This short book is at once stimulating and informative.

Chapter 3

Ethics and Criminal Justice Research

We’ll examine some of the ethical considerations that must be taken into account along with the scientific ones in the design and execution of research. We’ll consider different types of ethical issues and ways of handling them.

Introduction 52

Ethical Issues in Criminal Justice Research 52

No Harm to Participants 53

Voluntary Participation 57

Anonymity and Confidentiality 57

Deceiving Subjects 59

Analysis and Reporting 59

Legal Liability 60

Special Problems 61

Promoting Compliance with Ethical Principles 63

Putting It All Together: Ethics and Racial Profiling Research 64

Codes of Professional Ethics 64

Institutional Review Boards 66

Institutional Review Board Requirements and Researcher Rights 69

Ethics and Juvenile Gang Members 70

Two Ethical Controversies 72

Trouble in the Tearoom 72

The Stanford Prison Experiment 73

Discussion Examples 75

Main Points 76
Introduction

Despite our best intentions, we don’t always recognize ethical issues in research.

Most of this book focuses on scientific and procedures and constraints. We’ll see that the logic of science suggests certain research procedures, but we’ll also see that some scientifically “perfect” study designs are not feasible, because they would be too expensive or take too long to execute. Throughout the book, we’ll deal with workable compromises.

Before we get to scientific and practical constraints on research, it’s important to explore another essential consideration in doing criminal justice research in the real world—ethics. Just as certain designs or measurement procedures are impractical, others are constrained by ethical problems. We introduce the issue of ethics by quoting from Earl Babbie’s book *The Practice of Social Research* (2004:62):

Several years ago, I was invited to sit in on a planning session to design a study of legal education in California. The joint project was to be conducted by a university research center and the state bar association. The purpose of the project was to improve legal education by learning which aspects of the law school experience were related to success on the bar exam. Essentially, the plan was to prepare a questionnaire that would get detailed information about the law school experiences of individuals. People would be required to answer the questionnaire when they took the bar exam. By analyzing how people with different kinds of law school experiences did on the bar exam, it would be possible to find out what sorts of things worked and what didn’t. The findings of the research could be made available to law schools, and ultimately legal education could be improved.

The exciting thing about collaborating with the bar association was that all the normally aggravating logistical hassles would be handled. There would be no problem getting permission to administer questionnaires in conjunction with the exam, for example, and the problem of nonresponse could be eliminated altogether.

I left the meeting excited about the prospects for the study. When I told a colleague about it, I glowed about the absolute handling of the nonresponse problem. Her immediate comment turned everything around completely. “That’s unethical. There’s no law requiring the questionnaire, and participation in research has to be voluntary.” The study wasn’t done.

It now seems obvious that requiring participation would have been inappropriate. You may have seen that before you read the comment by Babbie’s colleague.

All of us consider ourselves ethical—not perfect perhaps, but more ethical than most of humanity. The problem in criminal justice research—and probably in life—is that ethical considerations are not always apparent to us. As a result, we often plunge into things without seeing ethical issues that may be obvious to others and even to ourselves when they are pointed out. Our excitement at the prospect of a new research project may blind us to obstacles that ethical considerations present.

Any of us can immediately see that a study that requires juvenile gang members to demonstrate how they steal cars is unethical. You’d speak out immediately if we suggested interviewing people about drug use and then publishing what they said in the local newspaper. But, as ethical as we think we are, we are likely to miss the ethical issues in other situations—not because we’re bad, but because we’re human.

Ethical Issues in Criminal Justice Research

A few basic principles encompass the variety of ethical issues in criminal justice research.

In most dictionaries and in common usage, ethics is typically associated with morality, and both deal with matters of right and
wrong. But what is right and what is wrong? What is the source of the distinction? For various individuals, the sources vary from religion, to political ideology, to pragmatic observations of what seems to work and what doesn’t.

Webster’s New World Dictionary (4th ed.) is typical among dictionaries in defining ethical as “conforming to the standards of conduct of a given profession or group.” Although the relativity embedded in this definition may frustrate those in search of moral absolutes, what we regard as moral and ethical in day-to-day life is no more than a matter of agreement among members of a group. And, not surprisingly, different groups have agreed on different ethical codes of conduct. If someone is going to live in a particular society, then, it is extremely useful to know what that society considers ethical and unethical. The same holds true for the criminal justice research “community.”

Anyone preparing to do criminal justice research should be aware of the general agreements shared by researchers about what’s proper and improper in the conduct of scientific inquiry. Ethical issues in criminal justice can be especially challenging because our research questions frequently address illegal behavior that people are anxious to conceal. This is true of offenders and, sometimes, people who work in criminal justice agencies.

The sections that follow explore some of the more important ethical issues and agreements in criminal justice research. Our discussion is restricted to ethical issues in criminal justice research, not ethics in criminal justice policy and practice. Thus, we will not consider such issues as the morality of the death penalty, acceptable police practices, the ethics of punishment, or codes of conduct for attorneys and judges. If you are interested in substantive ethical issues in criminal justice policy, consult Jocelyn Pollock (2003) or Richard Hall and associates (1999) for an introduction. The journal Criminal Justice Ethics includes a wide range of articles.

No Harm to Participants

Weighing the potential benefits from doing research against the possibility of harm to the people being studied—or harm to other people—is a fundamental ethical dilemma in all research. For example, biomedical research can involve potential physical harm to people or animals. Social research may cause psychological harm or embarrassment in people who are asked to reveal information about themselves. Criminal justice research has the potential to produce both physical and psychological harm, as well as embarrassment. Although the likelihood of physical harm may seem remote, it is worthwhile to consider possible ways it might occur.

Harm to subjects, researchers, or third parties is possible in field studies that collect information from or about persons engaged in criminal activity; this is especially true for field research. Studies of drug crimes may involve locating and interviewing active users and dealers. For example, Bruce Johnson and associates (1985) studied heroin users in New York, recruiting subjects by spreading the word through various means. Other researchers have studied dealers in Detroit (Mieczkowski, 1990), and St. Louis (Jacobs, 1999). Collecting information from active criminals presents at least the possibility of violence against research subjects by other drug dealers.

Potential danger to field researchers should also be considered. For instance, Peter Reuter and associates (Reuter, MacCoun, and Murphy, 1990) selected their drug dealer subjects by consulting probation department records. The researchers recognized that sampling persons from different Washington, DC, neighborhoods would have produced a more generalizable group of subjects, but they rejected that approach because mass media reports of widespread drug-related violence generated concern about the safety of research staff (1990:119). Whether such fears were warranted is unclear, but this example does illustrate how safety issues can affect criminal justice research.
A more sobering example of possible safety concerns is a series of experimental studies of bystander intervention by Bibb Latané and John Darley (1970). These researchers staged crimes that, in some instances, could have produced harm to either subjects or research staff. They were interested in circumstances when witnesses to a “crime” would or would not intervene. In one experiment, staging a liquor store holdup, a bystander called police, who responded with drawn guns.

Other researchers acknowledge the potential for harm in the context of respect for ethical principles. The box titled “Ethics and Extreme Field Research” by Dina Perrone, Rutgers University, discusses the challenges and dilemmas faced by researchers in high-risk environments. As a female ethnographer studying active drug use in a New York dance club, Perrone encountered awkward and difficult situations. Her main purpose was to study the use of ecstasy and other drugs in rave club settings. She became a participant observer in an all-night dance club (The Plant) where the use of club drugs was common. She covertly observed activities in the club, partly masking her role as a researcher by assuming the role of club-goer.

Though she was required to comply with university institutional review board guidelines, published codes and regulations offered limited guidance for many of the situations she experienced. As a result, she had to use her best judgment, learning from past experiences to make immediate decisions regarding ethical issues. She was forced to make decisions about how to handle drug episodes, so as not to place her research or her informants in any danger. Because her research was conducted in a dance club that is also a place for men to pick up women, she faced problems in getting information from subjects while watching out for her physical safety.

Drug Episodes and Subject Safety

I witnessed many drug episodes—adverse reactions to various club drugs—in my visits to The Plant. I watched groups trying to get their friends out of “K-holes” resulting from Ketamine or “Special K.” I even aided a subject throwing up. Being a covert observer made it difficult to handle these episodes. There were times in the club when I felt as though I was the only person not under the influence of a mind-altering substance. This led me to believe that I had better judgment than the other patrons. Getting involved in these episodes, however, risked jeopardizing my research.

During my first observation, I tried to intervene in what appeared to be a serious drug episode but was warned off by an informant. I was new to the club and unsure what would happen if I got involved. If I sought help from club staff or outsiders in dealing with acute drug reactions, patrons as well as the bouncers would begin to question why I kept coming there. I needed to gain the trust of the patrons in order to enlist participants in my research. Furthermore, the bouncers could throw me out of the club, fearing I was a troublemaker who would summon authorities.

As a researcher, I have an ethical responsibility to my participants, and as a human being, I have an ethical responsibility to my conscience. I decided to be extra cautious during my research and to pay close attention to how drug episodes are handled. I would first consult my informants and follow their suggestions. But if I ever felt a person suffering a drug episode was at risk while other patrons were neither able nor inclined to help, I would intervene to the best of my ability.

Sexual Advances in the Dance Club

The Plant is also partly a “meat market.” Unlike most bars and dance clubs, the patrons’ attire and the dance club entertainment are highly...
Extreme Field Research” gives examples of subtle and not-so-subtle ethical dilemmas encountered by a Rutgers University researcher in her study of drug use in rave clubs.

More generally, John Monahan and associates (1993) distinguish three different groups at potential risk of physical harm in their research on violence. First are research subjects themselves. Women at risk of domestic violence, for example, may be exposed to greater danger if assailants learn they have disclosed past victimizations to researchers. Second, researchers might trigger attacks on themselves when they interview subjects who have a

Ethical Decision Rules Evolving from Experience

Academic associations have formulated codes of ethics and professional conduct, but limited guidance is available for handling issues that arise in some types of ethnographic research. Instead, like criminal justice practitioners, those researchers have to make immediate decisions based on experience and training, without knowing how a situation will unfold. Throughout my research, I found myself in situations that I would normally avoid and would probably never confront. Should I help her get through a drug episode? If I don’t, will she be okay? If I walk away from this aggressive guy, will he follow me? Does he understand that I wanted to talk to him just for research?

The approach I developed to tackle these issues was mostly gained by consulting with colleagues and reading other studies. An overarching theme regarding all codes of ethics is that ethnographers must put the safety and interests of their participants first, and they must recognize that their informants are more knowledgeable about many situations than they are. Throughout the research, I used my judgment to make the best decisions possible when handling these situations. To decide when to intervene during drug episodes, I followed the lead of my informants. Telling men that my informant was my boyfriend and walking away were successful tactics in turning away sexual advances.
The potential for psychological harm to subjects exists when interviews are used to collect information. For example, crime surveys that ask respondents about their experiences as victims of crime may remind them of a traumatic, or at least an unpleasant, experience. Surveys may also ask respondents about illegal behaviors such as drug use or crimes they have committed. Talking about such actions with interviewers can be embarrassing.

Some researchers have taken special steps to reduce the potential for emotional trauma in interviews of domestic violence victims (Tjaden and Thoennes, 2000). One of the most interesting examples involves the use of self-completed computer questionnaires in the British Crime Survey (Mirrlees-Black, 1999). Rather than verbally respond to questions from interviewers, respondents read and answer questions on a laptop computer. This procedure affords a greater degree of privacy for research subjects. Although the fact often goes unrecognized, subjects can also be harmed by the analysis and reporting of data. Every now and then, research subjects read the books published about the studies they participated in. Reasonably sophisticated subjects can locate themselves in the various indexes and tables of published studies. Having done so, they may find themselves characterized—though not identified by name—as criminals, deviants, probation violators, and so forth.

Largely for this reason, information on the city of residence of victims identified in the National Crime Victimization Survey is not available to researchers or the public. The relative rarity of some types of crime means that if crime victimization is reported by city of residence individual victims might recognize the portrayal of their experience or might be identified by third parties.

Recent developments in the use of crime mapping software have raised similar concerns. Many police departments now use some type of computer-driven crime map, and some have made maps of small areas available to the public on the web. As Tom Casady (1999) points out, this raises new questions of privacy as individuals might be able to identify crimes directed against their neighbors. Researchers and police alike must recognize the potential for such problems before publishing or otherwise displaying detailed crime maps. See crime maps for cities in the San Diego metropolitan area for examples (http://www.arjis.org; accessed 28 March 2006).

By now, it should be apparent that virtually all research runs some risk of harming other people somehow. A researcher can never completely guard against all possible injuries, yet some study designs make harm more likely than others do. If a particular research procedure seems likely to produce unpleasant effects for subjects—asking survey respondents to report deviant behavior, for example—the researcher should have firm scientific grounds for doing so. If researchers pursue a design that is essential and also likely to be unpleasant for subjects, they will find themselves in an ethical netherworld, forced to do some personal agonizing.

As a general principle, possible harm to subjects may be justified if the potential benefits of the study outweigh the harm. Of course, this raises a further question of how to determine whether possible benefits offset possible harms. There is no simple answer, but as we will see, the research community has adopted certain safeguards that help subjects to make such determinations themselves.

Not harming people is an easy norm to accept in theory, but it is often difficult to ensure in practice. Sensitivity to the issue and experi-
ence in research methodology, however, should improve researchers’ efforts in delicate areas of inquiry. Review Dina Perrone’s observations in the box “Ethics and Extreme Field Research” for examples.

Voluntary Participation
Criminal justice research often intrudes into people's lives. The interviewer’s telephone call or the arrival of an email questionnaire signals the beginning of an activity that respondents have not requested and that may require a significant portion of their time and energy. Being selected to participate in any sort of research study disrupts subjects’ regular activities.

A major tenet of medical research ethics is that experimental participation must be voluntary. The same norm applies to research in criminal justice. No one should be forced to participate. But this norm is far easier to accept in theory than to apply in practice.

For example, prisoners are sometimes used as subjects in experimental studies. In the most rigorously ethical cases, prisoners are told the nature—and the possible dangers—of the experiment; they are told that participation is completely voluntary; and they are further instructed that they can expect no special rewards (such as early parole) for participation. Even under these conditions, volunteers often are motivated by the belief that they will personally benefit from their cooperation. In other cases, prisoners—or other subjects—may be offered small cash payments in exchange for participation. To people with very low incomes, small payments may be an incentive to participate in a study they would not otherwise endure.

When an instructor in an introductory criminal justice class asks students to fill out a questionnaire that she or he plans to analyze and publish, students should always be told that their participation in the survey is completely voluntary. Even so, students might fear that nonparticipation will somehow affect their grade. The instructor should therefore be especially sensitive to the implied sanctions and make provisions to obviate them. For example, students could be asked to drop the questionnaires in a box near the door prior to the next class.

Notice how this norm of voluntary participation works against a number of scientific concerns or goals. In the most general terms, the goal of generalizability is threatened if experimental subjects or survey respondents are only the people who willingly participate. The same is true when subjects’ participation can be bought with small payments. Research results may not be generalizable to all kinds of people. Most clearly, in the case of a descriptive study, a researcher cannot generalize the study findings to an entire population unless a substantial majority of a scientifically selected sample actually participates—both the willing respondents and the somewhat unwilling.

Field research (the subject of Chapter 10) has its own ethical dilemmas in this regard. Often, a researcher who conducts observations in the field cannot even reveal that a study is being done, for fear that this revelation might significantly affect what is being studied. Imagine, for example, that you are interested in whether the way stereo headphones are displayed in a discount store affects rates of shoplifting. Therefore, you plan a field study in which you will make observations of store displays and shoplifting. You cannot very well ask all shoppers whether they agree to participate in your study.

The norm of voluntary participation is an important one, but it is sometimes impossible to follow. In cases in which researchers ultimately feel justified in violating it, it is all the more important to observe the other ethical norms of scientific research.

Anonymity and Confidentiality
The clearest concern in the protection of the subjects’ interests and well-being is the protection of their identity. If revealing their behavior or responses would injure them in any way, adherence to this norm becomes crucial. Two
techniques—anonymity and confidentiality—assist researchers in this regard, although the two are often confused.

**Anonymity** A research subject is considered anonymous when the researcher cannot associate a given piece of information with the person. **Anonymity** addresses many potential ethical difficulties. Studies that use field observation techniques are often able to ensure that research subjects cannot be identified. Researchers may also gain access to nonpublic records from courts, corrections departments, or other criminal justice agencies in which the names of persons have been removed.

One example of anonymity is a web-based survey where no login or other identifying information is required. Respondents anonymously complete online questionnaires that are then tabulated. Likewise, a telephone survey is anonymous if residential phone numbers are selected at random and respondents are not asked for identifying information. Interviews with subjects in the field are anonymous if the researchers neither ask for nor record the names of subjects.

Ensuring anonymity makes it difficult to keep track of which sampled respondents have been interviewed, because researchers did not record their names. Nevertheless, in some situations, the price of anonymity is worth paying. In a survey of drug use, for example, we may decide that the likelihood and accuracy of responses will be enhanced by guaranteeing anonymity.

Respondents in many surveys cannot be considered anonymous because an interviewer collects the information from individuals whose names and addresses are known. Other means of data collection may similarly make it impossible to guarantee anonymity for subjects. If we wished to examine juvenile arrest records for a sample of ninth-grade students, for example, we would need to know their names even though we might not be interviewing them or having them fill out a questionnaire.

**Confidentiality** **Confidentiality** means that a researcher is able to link information with a given person’s identity but essentially promises not to do so publicly. In a survey of self-reported drug use, for example, the researcher is in a position to make public the use of illegal drugs by a given respondent, but the respondent is assured that this will not be done. Similarly, if field interviews are conducted with juvenile gang members, researchers can certify that information will not be disclosed to police or other officials. Studies using court or police records that include individuals’ names may protect confidentiality by not including any identifying information.

Some techniques ensure better performance on this guarantee. To begin, field or survey interviewers who have access to respondent identifications should be trained in their ethical responsibilities. As soon as possible, all names and addresses should be removed from data collection forms and replaced by identification numbers. A master identification file should be created linking numbers to names to permit the later correction of missing or contradictory information. This file should be kept under lock and key and be made available only for legitimate purposes.

Whenever a survey is confidential rather than anonymous, it is the researcher’s responsibility to make that fact clear to respondents. He or she must never use the term anonymous to mean confidential. Note, however, that research subjects and others may not understand the difference. For example, a former assistant attorney general in New Jersey once demanded that Maxfield disclose the identities of police officers who participated in an anonymous study. It required repeated explanations of the difference between “anonymous” and “confidential” before the lawyer finally understood that it was not possible to identify participants. In any event, subjects should be assured that the information they provide will be used for research purposes only and not be disclosed to third parties.
Deceiving Subjects

We've seen that the handling of subjects’ identities is an important ethical consideration. Handling our own identity as a researcher can be tricky, too. Sometimes, it’s useful and even necessary to identify ourselves as researchers to those we want to study. It would take a master con artist to get people to participate in a laboratory experiment or complete a lengthy questionnaire without letting on that research was being conducted. We should also keep in mind that deceiving people is unethical; in criminal justice research, deception needs to be justified by compelling scientific or administrative concerns.

Sometimes, researchers admit that they are doing research but fudge about why they are doing it or for whom. For example, Cathy Spatz Widom and associates interviewed victims of child abuse some 15 years after their cases had been heard in criminal or juvenile courts (Widom, Weiler, and Cottler, 1999). Widom was interested in whether child abuse victims were more likely than a comparison group of non-victims to have used illegal drugs. Interviewers could not explain the purpose of the study without potentially biasing responses. Still, it was necessary to provide a plausible explanation for asking detailed questions about personal and family experiences. Widom’s solution was to inform subjects that they had been selected to participate in a study of human development. She also prepared a brochure describing her research on human development that was distributed to respondents.

Although we might initially think that concealing our research purpose by deception would be particularly useful in studying active offenders, James Inciardi (1993), in describing methods for studying “crack houses,” makes a convincing case that this is inadvisable. First, concealing our research role when investigating drug dealers and users implies that we are associating with them for the purpose of obtaining illegal drugs. Faced with this situation, a researcher would have the choice of engaging in illegal behavior or offering a convincing explanation for declining to do so. Second, masquerading as a crack house patron would have exposed the researcher to the considerable danger of violence that was found to be common in such places. Because the choice of committing illegal acts or becoming a victim of violence is really no choice at all, Inciardi (1993:152) advises researchers who study active offenders in field settings: “Don’t go undercover.”

Analysis and Reporting

As criminal justice researchers, then, we have ethical obligations to our subjects of study. At the same time, we have ethical obligations to our colleagues in the scientific community; a few comments on those obligations are in order. In any rigorous study, the researcher should be more familiar than anyone else with the technical shortcomings and failures of the study. Researchers have an obligation to make those shortcomings known to readers. Even though it’s natural to feel foolish admitting mistakes, researchers are ethically obligated to do so.

Any negative findings should be reported. There is an unfortunate myth in social scientific reporting that only positive discoveries are worth reporting (and journal editors are sometimes guilty of believing that as well). And this is not restricted to social science. For example, Helle Krogh Johansen and Peter Gotzsche (1999) describe how published research on new drugs tends to focus on successful experiments. Unsuccessful research on new formulations is less often published, which leads pharmaceutical researchers to repeat studies of drugs already shown to be ineffective. Largely because of this bias, researchers at the Johns Hopkins University Medical School have established the Journal of Negative Observations in Genetic Oncology (NOGO), dedicated to publishing negative findings from cancer research [http://www.path.jhu.edu/NOGO/]. In social science, as in medical research, it is often as important to know that two things are not related as to know that they are.
Similarly, researchers should avoid the temptation to save face by describing findings as the product of a carefully planned analytic strategy when that is not the case. Many findings are unexpected even though they may seem obvious in retrospect. Suppose you uncover an interesting relationship by accident—so what? Embroidering such situations with descriptions of fictitious hypotheses is dishonest and tends to mislead inexperienced researchers into thinking that all scientific inquiry is rigorously preplanned and organized. Fortunately, evaluation researchers recognize the value of reporting unexpected findings and have developed evaluation approaches that increase the likelihood of discovering unanticipated benefits (Tilley, 2000).

In general, science progresses through honesty and openness, and is retarded by ego defenses and deception. We can serve our fellow researchers—and the scientific community as a whole—by telling the truth about all the pitfalls and problems experienced in a particular line of inquiry. With luck, this will save others from the same problems.

Legal Liability
Two types of ethical problems expose researchers to potential legal liability. First, assume you are making field observations of criminal activity such as street prostitution that is not reported to police. Under criminal law in many states, you might be arrested for obstructing justice or being an accessory to a crime. Potentially more troublesome is the situation in which participant observation of crime or deviance draws researchers into criminal or deviant roles themselves—smuggling cigarettes into a lockup in order to obtain the cooperation of detainees, for example.

The second and more common potential source of legal problems involves knowledge that research subjects have committed illegal acts. Self-report surveys or field interviews may ask subjects about crimes they have committed. If respondents report committing offenses they have never been arrested for or charged with, the researcher’s knowledge of them might be construed as obstruction of justice. Or research data may be subject to subpoena by a criminal court. Because disclosure of research data that could be traced to individual subjects violates the ethical principle of confidentiality, a new dilemma emerges.

Fortunately, federal law protects researchers from legal action in most circumstances, provided that appropriate safeguards are used to protect research data. Research plans for 2002 published by organizations in the Office of Justice Programs summarized this protection: “[Research] information and copies thereof shall be immune from legal process, and shall not, without the consent of the person furnishing such information, be admitted as evidence or used for any purpose in any action, suit, or other judicial, legislative, or administrative proceedings” (42 U.S. Code §22.28a). This not only protects researchers from legal action but also can be valuable in assuring subjects that they cannot be prosecuted for crimes they describe to an interviewer or field worker. For example, Bruce Johnson and associates (1985:219) prominently displayed a Federal Certificate of Confidentiality at their research office to assure heroin dealers that they could not be prosecuted for crimes disclosed to interviewers. Savvier than many people about such matters, heroin users were duly impressed.

Note that such immunity requires confidential information to be protected. We have already discussed the principle of confidentiality, so this bargain should be an easy one to keep.

Somewhere between legal liability and physical danger lies the potential risk to field researchers from law enforcement. Despite being up front with crack users about his role as a researcher, Inciardi (1993) points out that police could not be expected to distinguish him from his subjects. Visibly associating with offenders in natural settings brings some risk of being arrested or inadvertently being an
accessory to crime. Thus, Inciardi on one occasion fled the scene of a robbery and on another was caught up in a crack-house raid. Another example is the account Bruce Jacobs (1996) gives of his contacts with police while he was studying street drug dealers. Exercises presented at the end of the chapter ask you to think more carefully about the ethical issues involved in Jacobs’s contact with police.

Special Problems
Certain types of criminal justice studies present special ethical problems in addition to those we have mentioned. Applied research, for example, may evaluate some existing or new program. Evaluations frequently have the potential to disrupt the routine operations of agencies being studied. Obviously, it is best to minimize such interferences whenever possible.

Staff Misbehavior
While conducting applied research, researchers may become aware of irregular or illegal practices by staff in public agencies. They are then faced with the ethical question of whether to report such information. For example, investigators conducting an evaluation of an innovative probation program learned that police visits to the residences of probationers were not taking place as planned.1 Instead, police assigned to the program had been submitting falsified log sheets and had not actually checked on probationers.

What is the ethical dilemma in this case? On the one hand, researchers were evaluating the probation program and so were obliged to report reasons it did or did not operate as planned. Failure to deliver program treatments (home visits) is an example of a program not operating as planned. Investigators had guaranteed confidentiality to program clients—the offenders assigned to probation—but no such agreement had been struck with program staff. On the other hand, researchers had assured agency personnel that their purpose was to evaluate the probation program, not individuals’ job performance. If researchers disclosed their knowledge that police were falsifying reports, they would violate this implied trust.

What would you have done in this situation? We will tell you what the researchers decided at the end of this chapter. You should recognize, however, how applied research in criminal justice agencies can involve a variety of ethical issues.

Research Causes Crime
Because criminal acts and their circumstances are complex and imperfectly understood, some research projects have the potential to produce crime or influence its location or target. Certainly, this is a potentially serious ethical issue for researchers.

Most people agree that it is unethical to encourage someone to commit an offense solely for the purpose of a research project. What’s more problematic is recognizing situations in which research might indirectly promote offending. Scott Decker and Barrik Van Winkle (1996) discuss such a possibility in their research on gang members. Some gang members offered to illustrate their willingness to use violence by inviting researchers to witness a drive-by shooting. Researchers declined all such invitations (1996:46). Another ethical issue was the question of how subjects used the $20 cash payments they received in exchange for being interviewed (1996:51):

We set the fee low enough that we were confident that it would not have a criminogenic effect. While twenty dollars is not a small amount of money, it is not sufficient to purchase a gun or bankroll a large drug buy. We are sure that some of our subjects used the money for illegal purposes. But, after all, these were individuals who were regularly engaged in delinquent and criminal acts.

You may or may not agree with the authors’ reasoning in the last sentence. But their con-

---

1 Information about this example from personal communication between researchers and one of the authors.
consideration of how cash payments would be used by active offenders represents an unusually careful recognition of the ethical dilemmas that emerge in studying active offenders.

A different type of ethical problem is the possibility of crime displacement in studies of crime prevention programs. For example, consider an experimental program to reduce street prostitution in one area of a city. Researchers studying such a program might designate experimental target areas for enhanced enforcement, as well as nearby comparison areas that will not receive an intervention. If prostitution is displaced from target areas to adjacent neighborhoods, the evaluation study contributes to an increase in prostitution in the comparison areas.

In a review of more than 50 evaluations of crime prevention projects, René Hesseling (1994) concludes that displacement tended to be associated with programs targeting street prostitution, bank robbery, and certain combinations of offenses. The type of crime prevention action also made a difference, with displacement more common for target-hardening programs. For example, installing security screens on ground-floor windows in some buildings seemed to displace burglary to less protected structures. Similarly, adding steering column locks to new cars tended to increase thefts of older cars (Felson and Clarke, 1998).

At the same time, Hesseling demonstrates that displacement is by no means inevitable. Ronald Clarke and John Eck (2005) further argue that researchers and public officials unrealistically assume a deterministic model of offending behavior. Instead, offenders are easily dissuaded by a variety of crime prevention measures.

In any event, when it does occur, displacement tends to follow major policy changes that are not connected with criminal justice research. Researchers cannot be expected to control actions by criminal justice officials that may benefit some people at the expense of others. However, it is reasonable to expect researchers involved in planning an evaluation study to anticipate the possibility of such things as displacement and bring them to the attention of program staff.

Withholding of Desirable Treatments  Certain kinds of research designs in criminal justice can lead to different kinds of ethical questions. Suppose, for example, researchers believe that diverting domestic violence offenders from prosecution to counseling reduces the possibility of repeat violence. Is it ethical to conduct an experiment in which some offenders are prosecuted but others are not?

You may recognize the similarity between this question and those faced by medical researchers who test the effectiveness of experimental drugs. Physicians typically respond to such questions by pointing out that the effectiveness of a drug cannot be demonstrated without such experiments. Failure to conduct research, even at the potential expense of subjects not receiving the trial drugs, would therefore make it impossible to develop new drugs or to distinguish beneficial treatments from those that are ineffective and even harmful.

One solution to this dilemma is to interrupt an experiment if preliminary results indicate that a new policy or drug does in fact produce improvements in a treatment group. For example, Michael Dennis (1990) describes how such plans were incorporated into a long-term evaluation of enhanced drug treatment counseling. If preliminary results had indicated that the new counseling program reduced drug use, researchers and program staff were prepared to provide enhanced counseling to subjects in the control group. Dennis recognized this potential ethical issue and planned his elaborate research design to accommodate such midstream changes. Similarly, Martin Killias and associates (2000) planned to interrupt their experimental study of heroin prescription in Switzerland if compelling evidence pointed to benefits from that approach to treated drug dependency.
Mandatory Reporting The situation is somewhat murkier for researchers studying certain kinds of family violence. Following the Federal Child Abuse Prevention and Treatment Act of 1974, all states developed child protection agencies and adopted mandatory reporting laws. Specific provisions vary, but in general, people who learn about possible cases of child abuse must report them to designated state agencies. This certainly seems to be a worthwhile goal, but what about researchers who learn about possible child maltreatment in the course of a survey? In most states, such requirements apply only to health professionals and teachers. But in eight states, anyone who suspects a case of child maltreatment must report it to designated authorities.

Notice how this is consistent with one ethical principle—protection of human subjects by reporting possible victims—but at odds with another principle—confidentiality. A Bureau of Justice Statistics report on human subjects protection suggests that researchers warn subjects at the beginning of an interview that any information disclosed about child abuse must be reported to authorities (Sieber, 2001). But that threatens researchers’ ability to learn about child abuse. Another approach, adopted by Lianne Woodward and David Fergusson (2000), is to interview subjects age 18 and older, asking about experiences of abuse victimization when they were children. This is an imperfect solution, but it illustrates the trade-offs between our interest in protecting research subjects and our interest in studying the phenomenon of child abuse. For more examples and guidance in this vexing research/ethical area, see Mandated Reporting of Suspected Child Abuse: Ethics, Law, and Policy (Kalichman, 2000).

Research in criminal justice, especially applied research, can pose a variety of ethical dilemmas, only some of which we have mentioned here. See the “Additional Readings” at the end of this chapter for more information. Our running example also illustrates various ethical questions that emerge in conducting research on racial profiling and traffic enforcement.

Promoting Compliance with Ethical Principles

Codes of ethics and institutional review boards are two main ways of promoting compliance with ethical principles.

No matter how sensitive they might be to the rights of individuals and possible ways subjects might be harmed, researchers are not always the best judges of whether or not adequate safeguards are used. In 1974 the National Research Act was signed into law after a few highly publicized examples of unethical practices in medical and social science research. A few years later, what has become known as The Belmont Report prescribed a brief, but comprehensive set of ethical principles for protecting human subjects (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). In only six pages, three principles were presented:

1. Respect for persons: Individuals must be allowed to make their own decisions about participation in research, and those with limited capacity to make such decisions should have special protection.
2. Beneficence: Research should do no harm to participants, and seek to produce benefits.
3. Justice: The benefits and burdens of participating in research should be distributed fairly.

Copious federal regulations have stemmed from these three principles. But in most cases, the research community has adopted two general mechanisms for promoting ethical research practices: codes of professional ethics and institutional review boards.

We’ll consider one of these examples near the end of this chapter.
Codes of Professional Ethics

If the professionals who design and conduct research projects can fail to recognize ethical problems, how can such problems be avoided? One approach is for researchers to consult one of the codes of ethics produced by professional associations. Formal codes of conduct describe what is considered acceptable and unacceptable professional behavior. The American Psychological Association (2002) code of

PUTTING IT ALL TOGETHER

ETHICS AND RACIAL PROFILING RESEARCH

Conducting research on racial profiling potentially involves researchers in a variety of ethical questions. First, researchers in New Jersey (Andresen, 2005; Maxfield and Kelling, 2005), North Carolina (Smith, Tomaskovic-Devey, Zingraff, et al., 2003), and Pennsylvania (Engel, Calnon, Liu, et al., 2004; Engel, Calnon, Tillyer, et al., 2005) interacted with human subjects, primarily police, supervisors, command officers, and administrative staff. Second, all projects required access to confidential files maintained by state police agencies. Third, when Rutgers researchers accompanied New Jersey state troopers on patrol, they came into contact with private citizens, often people stopped for some traffic offense. Fourth, accompanying troopers on routine patrol exposed researchers to high-speed driving on congested highways and other potential dangers. Finally, it became clear that voluntary participation by state troopers was not always entirely voluntary.

Police as Human Subjects

Studying state troopers as they patrolled New Jersey highways raised ethical issues that are somewhat different from those involving private individuals and their everyday lives. On the one hand, researchers simply observed what public servants are paid to do. Troopers patrol in public view and interact with drivers as part of their daily routine. In a sense, researchers became not much different from curious bystanders who are naturally drawn to real-life examples of the stuff of television dramas.

But field research observing police is different in certain respects. Researchers accompanied troopers for an extended time, listening in on their conversations with each other and on messages broadcast over the radio. Troopers drive on public roads, but much of what happens inside patrol units is private, just as the general public enjoys some measure of privacy inside personal vehicles. Riding with state police revealed candid actions and comments. Maxfield and Andresen learned firsthand what sorts of driving infractions were routinely overlooked and which ones were certain to warrant a traffic stop. Researchers witnessed actions that were inconsistent with New Jersey State Police regulations, and heard a fair share of racist and sexist comments.

Researchers in Pennsylvania had access to all records of police stops. Among other things they were interested in whether some troopers stopped higher proportions of minority drivers. However, researchers wanted to protect the identity of individual troopers; they wished to study only aggregate patterns, not individual police. Accordingly, after they received traffic stop records, the identifying badge number of the trooper who made the stop was deleted from each data record.

Confidential Information

At the time of the research in New Jersey, state police were under scrutiny by the U.S. Department of Justice. Federal officials and their representatives had access to information about police stops; some of this was shared with Rutgers researchers who were bound to protect the confi-
Chapter 3  Ethics and Criminal Justice Research  65

Ethics is quite detailed, reflecting the different professional roles of psychologists in research, clinical treatment, and educational contexts. Many of the ethical questions that criminal justice researchers are likely to encounter are addressed in the ethics code of the American Sociological Association (1997). Paul Reynolds (1979:442–49) has created a composite code for the use of human subjects in research, drawing on 24 codes of ethics published by national...
associations of social scientists. The National Academy of Sciences publishes a very useful booklet on a variety of ethical issues, including the problem of fraud and other forms of scientific misconduct (Committee on Science, Engineering, and Public Policy, 1995).

The two national associations representing criminology and criminal justice researchers have one code of ethics between them. The Academy of Criminal Justice Sciences (ACJS) based its code of ethics on that developed by the American Sociological Association. ACJS members are bound by a very general code that reflects the diversity of its membership: “Most of the ethical standards are written broadly, to provide applications in varied roles and varied contexts. The Ethical Standards are not exhaustive—conduct that is not included in the Ethical Standards is not necessarily ethical or unethical” (Academy of Criminal Justice Sciences, 2000:1).

After years of inaction, a committee of the American Society of Criminology (ASC) proposed a draft code of ethics in 1998, one that drew extensively on the code for sociology. But no ethics code had been adopted as of July 2006, and the ASC withdrew its draft code from circulation in 1999. In personal correspondence with Maxfield in 2003, a prominent ASC officer expressed doubt that any sort of code would be approved soon; eventually, this person felt, a very brief statement of general principles might be approved. In the meantime, the ASC website includes the following statement on a page titled “Code of Ethics”:

The American Society of Criminology has not formally adopted a code of ethics. We would suggest that persons interested in this general topic examine the various codes of ethics adopted by other professional associations. (http://asc41.com/ethicspg.html; accessed 31 March 2006)

Links to a number of ethics codes for social science associations are listed, including those listed above and the British Society of Criminology.

What can we make of the inability of the largest professional association of criminologists to agree on a code of ethics? The wide variety of approaches to doing research in this area probably has something to do with it. Criminologists also encounter a range of ethical issues and have diverging views on how those issues should be addressed. Finally, we have seen examples of the special problems that criminologists face in balancing ethics and research. Not all of these problems have easy solutions that can be embodied in a code.

Even when they exist, professional codes of ethics for social scientists cannot be expected to prevent unethical practices in criminal justice research any more than the American Bar Association’s Code of Professional Responsibility eliminates breaches of ethics by lawyers. For this reason, and in reaction to some controversial medical and social science research, the U.S. Department of Health and Human Services (HHS) has established regulations protecting human research subjects. These regulations do not apply to all social science or criminal justice research. It is, however, worthwhile to understand some of their general provisions. Material in the following section is based on the Code of Federal Regulations, Title 45, Chapter 4.6. And those regulations are themselves rooted in the Belmont Report.

Institutional Review Boards

Government agencies and nongovernment organizations (including universities) that conduct research involving human subjects must establish review committees, known as institutional review boards (IRBs). These IRBs have two general purposes. First, board members make judgments about the overall risks to human subjects and whether these risks are acceptable, given the expected benefits from actually doing the research. Second, they determine whether the procedures to be used include adequate safeguards regarding the safety,
confidentiality, and general welfare of human subjects.

Under HHS regulations, virtually all research that uses human subjects in any way, including simply asking people questions, is subject to IRB review. The few exceptions potentially include research conducted for educational purposes and studies that collect anonymous information only. However, even those studies may be subject to review if they use certain special populations (discussed later) or procedures that might conceivably harm participants. In other words, it’s safe to assume that most research is subject to IRB review if original data will be collected from individuals whose identities will be known. If you think about the various ways subjects might be harmed and the difficulty of conducting anonymous studies, you can understand why this is the case.

Federal regulations and IRB guidelines address other potential ethical issues in social research. Foremost among these is the typical IRB requirement for dealing with the ethical principle of voluntary participation.

**Informed Consent** The norm of voluntary participation is usually satisfied through informed consent—informing subjects about research procedures and then obtaining their consent to participate. Although this may seem like a simple requirement, obtaining informed consent can present several practical difficulties. It requires that subjects understand the purpose of the research, possible risks and side effects, possible benefits to subjects, and the procedures that will be used.

If you accept that deception may sometimes be necessary, you will realize how the requirement to inform subjects about research procedures can present something of a dilemma. Researchers usually address this problem by telling subjects at least part of the truth or offering a slightly revised version of why the research is being conducted. In Widom’s study of child abuse, subjects were partially informed about the purpose of the research—human development—one component of which is being a victim of child abuse, which the subjects were not told.

Another potential problem with obtaining informed consent is ensuring that subjects have the capacity to understand the descriptions of risks, benefits, procedures, and so forth. Researchers may have to provide oral descriptions to participants who are unable to read. For subjects who do not speak English, researchers should be prepared to describe procedures in their native language. And if researchers use specialized terms or language common in criminal justice research, participants may not understand the meaning and thus be unable to grant informed consent. For example, consider this statement: “The purpose of this study is to determine whether less restrictive sanctions such as restitution produce heightened sensitivity to social responsibility among persistent juvenile offenders and a decline in long-term recidivism.” Can you think of a better way to describe this study to delinquent 14-year-olds? Figure 3.1 presents a good example of an informed consent statement that was used in a study of juvenile burglars. Notice how the statement describes research procedures clearly and unambiguously tells subjects that participation is voluntary.

Other guidelines for obtaining informed consent include explicitly telling people that their participation is voluntary and assuring them of confidentiality. However, it is more important to understand how informed consent addresses key ethical issues in conducting criminal justice research. First, it ensures that participation is voluntary. Second, by informing subjects of procedures, risks, and benefits, researchers are empowering them to resolve the fundamental ethical dilemma of whether the possible benefits of the research offset the possible risks of participation.

**Special Populations** Federal regulations on human subjects include special provisions for certain types of subjects, and two of these are
particularly important in criminal justice research—juveniles and prisoners. Juveniles, of course, are treated differently from adults in most aspects of the law. Their status as a special population of human subjects reflects the legal status of juveniles, as well as their capacity to grant informed consent. In most studies that involve juveniles, consent must be obtained both from parents or guardians and from the juvenile subjects themselves.

In some studies, however, such as those that focus on abused children, it is obviously not desirable to obtain parental consent. Decker and Van Winkle faced this problem in their study of St. Louis gang members. See the box “Ethics and Juvenile Gang Members” for a discussion of how they reconciled the conflict between two ethical principles and satisfied the concerns of their university’s IRB.

Prisoners are treated as a special population for somewhat different reasons. Because of their ready accessibility for experiments and interviews, prisoners have frequently been used in biomedical experiments that produced serious harm (Mitford, 1973). Recognizing this, HHS regulations specify that prisoner subjects may not be exposed to risks that would be considered excessive for nonprison subjects. Furthermore, undue influence or coercion cannot be used in recruiting prisoner subjects. Informed consent statements presented to prospective subjects must indicate that a decision

---

You and your parents or guardian are invited to participate in a research study of the monitoring program that you were assigned to by the Juvenile Court. The purpose of this research is to study the program and your reactions to it. In order to do this a member of the research team will need to interview you and your parents/guardians when you complete the monitoring program. These interviews will take about 15 minutes and will focus on your experiences with the court and monitoring program, the things you do, things that have happened to you, and what you think. In addition, we will record from the court records information about the case for which you were placed in the monitoring program, prior cases, and other information that is put in the records after you are released from the monitoring program.

Anything you or your parents or guardian tell us will be strictly confidential. This means that only the researchers will have your answers. They will not under any conditions (except at your request) be given to the court, the police, probation officers, your parents, or your child!

Your participation in this research is voluntary. If you don’t want to take part, you don’t have to! If you decide to participate, you can change your mind at any time. Whether you participate or not will have no effect on the program, probation, or your relationship with the court.

The research is being directed by Dr. Terry Baumer and Dr. Robert Mendelsohn from the Indiana University School of Public and Environmental Affairs here in Indianapolis. If you ever have any questions about the research or comments about the monitoring program that you think we should know about, please call one of us at 274-0531.

---

Consent Statement

We agree to participate in this study of the Marion County Juvenile Monitoring Program. We have read the above statement and understand what will be required and that all information will be confidential. We also understand that we can withdraw from the study at any time without penalty.

Juvenile __________________________________________________________ Date: _____________________

Parent/Guardian __________________________________________________________________________________

Parent/Guardian __________________________________________________________________________________

Researcher ______________________________________________________________________________________

Figure 3.1 Informed Consent Statement for Evaluation of Marion County Juvenile Monitoring Program
not to participate in a study will have no influence on work assignments, privileges, or parole decisions. To help ensure that these ethical issues are recognized, if an IRB reviews a project in which prisoners will be subjects, at least one member of that IRB must be either a prisoner or someone specifically designated to represent the interests of prisoners. Figure 3.2 presents a checklist required by the Rutgers University IRB for proposed research involving prisoners.

Regarding the last item in Figure 3.2, randomization is generally recognized as an ethical procedure for selecting subjects or deciding which subjects will receive an experimental treatment. HHS regulations emphasize this in describing special provisions for using prison subjects: “Unless the principal investigator provides to the [IRB] justification in writing for following some other procedures, control subjects must be selected randomly from the group of available prisoners who meet the characteristics needed for that particular research project” (45 CFR 46.305[4]).

Institutional Review Board Requirements and Researcher Rights

Federal regulations contain many more provisions for IRBs and other protections for human subjects. Some researchers believe that such regulations actually create problems by setting constraints on their freedom and professional judgments in conducting research. Recall that potential conflict between the rights of researchers to discover new knowledge and the rights of subjects to be free from unnecessary harm is a fundamental ethical dilemma. It is at least inconvenient to have outsiders review a research proposal. Or a researcher may feel insulted by the implication that the potential benefits of research will not outweigh the potential harm or inconvenience to human subjects.

Much of the original impetus for establishing IRBs had to do with medical experimentation on humans. Many social research study designs are regarded as exempt from IRB review under federal guidelines (45 CFR 46.101[b]). Here are the most relevant for criminal justice research:

1. Research conducted in established or commonly accepted educational settings, involving normal educational practices.

2. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior, unless: (i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects’ responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects’ financial standing, employability, or reputation.

3. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior if: (i) the human subjects are elected or appointed public officials or candidates for public office; or (ii) any personally identifiable information will be maintained as confidential throughout the research and thereafter.

4. Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.
Scott Decker and Barrik Van Winkle faced a range of ethical issues in their study of gang members. Many of these should be obvious given what has been said so far in this chapter. Violence was common among subjects and presented a real risk to researchers. Decker and Van Winkle (1996:252) reported that 11 of the 99 members of the original sample had been killed since the project began in 1990. There was also the obvious need to assure confidentiality to subjects.

Their project was supported by a federal agency and administered through a university, so Decker and Van Winkle had to comply with federal human subjects guidelines as administered by the university institutional review board (IRB). And because many of the subjects were juveniles, they had to address federal regulations concerning that special population. Foremost among these was the normal requirement that informed consent for juveniles include parental notification and approval.

You may immediately recognize the conflicting ethical principles at work here, together with the potential for conflict. The promise of confidentiality to gang members is one such principle that was essential for the researchers to obtain candid reports of violence and other law-breaking behavior. But the need for confidentiality conflicted with initial IRB requirements to obtain parental consent for their children to participate in the research:

This would have violated our commitment to maintain the confidentiality of each subject, not to mention the ethical and practical difficulties of finding and...
informing each parent. We told the Human Subjects Committee that we would not, in effect, tell parents that their child was being interviewed because they were an active gang member, knowledge that the parents may not have had. (Decker and Van Winkle, 1996:52)

You might think deception would be a possibility—informing parents that their child was selected for a youth development study, for example. This would not, however, solve the logistical difficulty of locating parents or guardians, some of whom had lost contact with their children. Furthermore, it was likely that even if parents or guardians could be located, suspicions about the research project and the reasons their children were selected would prevent many parents from granting consent. Loss of juvenile subjects in this way would compromise the norm of generality as we have described it in this chapter and elsewhere.

Finally, waiving the requirement for parental consent would have undermined the legal principle that the interests of juveniles must be protected by a supervising adult. Remember that researchers are not always the best judges of whether sufficient precautions have been taken to protect subjects. Here is how Decker and Van Winkle (1996:52) resolved the issue with their IRB:

We reached a compromise in which we found an advocate for each juvenile member of our sample; this person—a university employee—was responsible for making sure that the subject understood (1) their rights to refuse or quit the interview at any time without penalty and (2) the confidential nature of the project. All subjects signed the consent form.

Source: Adapted from Decker and Van Winkle (1996).

(5) Research and demonstration projects which are conducted by or subject to the approval of Department or Agency heads, and which are designed to study, evaluate, or otherwise examine: (i) public benefit or service programs; (ii) procedures for obtaining benefits or services under those programs; (iii) possible changes in or alternatives to those programs or procedures; or (iv) possible changes in methods or levels of payment for benefits or services under those programs.

These exempt categories include quite a number of common models for doing criminal justice research. However, exempt means that research proposals do not have to be subject to full IRB review. In most cases, researchers who believe that their projects qualify for exemption under these categories must still have IRBs review the research to verify that exemption.

If this strikes you as confusing, you should feel reassured to learn that researchers, IRBs, and federal regulators are often confused as well. Given the unclear language of HHS regulations, many university IRBs have become extremely cautious in reviewing research proposals. See Christopher Shea's (2000) discussion for examples of problems resulting from this. Professional associations and research-oriented federal agencies have tried to offer guidance on what is and is not subject to various levels of IRB approval. Joan Sieber (2001) prepared an analysis of human subjects issues associated with large surveys for the Bureau of Justice Statistics. Always alert
for possible restrictions on academic freedom, the American Association of University Professors (2001) published a useful summary of how IRBs have come to regulate social science research.

There is some merit in such concerns; however, we should not lose sight of the reasons that IRB requirements and other regulations were created. Researchers are not always the best judges of the potential for their work to harm individuals. In designing and conducting criminal justice research, they may become excited about learning how to better prevent crime or more effectively treat cocaine addiction. That excitement and commitment to scientific advancement may lead researchers to overlook possible harms to individual rights or well-being. You may recognize this as another way of asking whether the ends justify the means. Because researchers are not always disinterested parties in answering such questions, IRBs are established to provide outside judgments. Also recognize that IRBs can be sources of expert advice on how to resolve ethical dilemmas. Decker and Van Winkle shared their university’s concern about balancing confidentiality against the need to obtain informed consent from juvenile subjects; together, they were able to fashion a workable compromise.

Virtually all colleges and universities have IRBs. Consult the Rutgers University website [http://orsp.rutgers.edu/humans/default.php#general] for an example, or visit the IRB website at your institution.

Another reason for creating regulations to protect human subjects, and IRBs to monitor compliance with those regulations, is the perceived failure of other means, together with ethical controversies raised by some actual studies (Homan, 1991). In the next section, we will briefly describe two research projects that provoked widespread ethical controversy and discussion. These are not the only two controversial projects that have been done; they simply illustrate ethical issues in the real world.

### Two Ethical Controversies

**Two examples of controversial studies illustrate key issues in ethics.**

The first project studied sexual behavior in public rest rooms, and the second examined how “prisoners” and “guards” reacted in a simulated prison setting.

#### Trouble in the Tearoom

As a graduate student, Laud Humphreys became interested in studying homosexual behavior. He developed a special interest in the casual and fleeting homosexual acts engaged in by some nonhomosexuals. In particular, his research interest focused on male homosexual acts between strangers who met in the public rest rooms in parks, called “tearooms” among homosexuals. The result was the publication of *The Tearoom Trade* (Humphreys, 1975).

What particularly interested Humphreys about the tearoom activity was that the participants seemed to lead otherwise conventional lives as “family men.” Thus, it was important to them that they remain anonymous in their tearoom visits. How would you study something like that?

Humphreys’ approach took advantage of the social structure of the situation. Typically, the tearoom encounter involved three people: the two men actually engaged in the homosexual act and a lookout, called the “watchqueen.” Thus, Humphreys began to show up at public rest rooms, offering to serve as watchqueen whenever it seemed appropriate. Because the watchqueen’s payoff was the chance to watch the action, Humphreys was able to observe behavior in natural settings, just as he would if he were studying drug dealers or jaywalking.

To round out his understanding of the tearoom trade, Humphreys needed to know something more about the people who participated. Given that the men probably would not have been thrilled about being interviewed by their watchqueen, Humphreys came up with a dif-
ferent solution. Whenever possible, he noted the license plate numbers of participants’ cars and tracked down their names and addresses through the police. Humphreys then visited the men at their homes, disguising himself enough to avoid recognition and claiming that he was conducting a survey. In that fashion, he collected the personal information he was unable to get in the rest rooms.

Humphreys’ research provoked considerable controversy, both within and outside the social scientific community. Some critics charged Humphreys with a gross invasion of privacy in the name of science: What men did in public rest rooms was their own business and not Humphreys’. Others were concerned about the deceit involved; Humphreys had lied to the participants by leading them to believe he was only a voyeur. People who believed that the tearoom participants, because they were in a public facility, were fair game for observation nonetheless protested the follow-up survey. They argued that it was unethical for Humphreys to trace the participants to their homes and to interview them under false pretenses. Still others justified Humphreys’ research. The topic, they said, was worth studying and couldn’t be studied any other way. They regarded the deceit as essentially harmless, noting that Humphreys was careful not to harm his subjects by disclosing their tearoom activities.

The tearoom trade controversy has never been resolved. It remains a topic of debate regarding research ethics and probably always will be because it stirs emotions and reveals ethical issues people disagree about. What do you think? Was Humphreys ethical in doing what he did? Are there parts of the research you feel were acceptable and other parts that were not? Whatever your opinion, you are sure to find others who disagree with you.

The Stanford Prison Experiment
The second research controversy differs from the first in many ways. Whereas Humphreys’ study involved participant observation, this study setting was in the laboratory. And whereas the first study examined a form of human deviance, this one focused on how people behave in formal institutions.

Few people would disagree that prisons are dehumanizing. Inmates forfeit freedom, of course, but their incarceration also results in a loss of privacy and individual identity. Violence is among the realities of prison life that people point to as evidence of the failure of prisons to rehabilitate inmates.

Although the problems of prisons have many sources, psychologists Curtis Haney, Craig Banks, and Philip Zimbardo (1973) were interested in two general explanations. The first was the dispositional hypothesis—prisons are brutal and dehumanizing because of the types of people who run them and are incarcerated in them. Inmates have demonstrated their disrespect for legal order and their willingness to use deceit and violence; persons who work as prison guards may be disproportionately authoritarian and sadistic. The second was the situational hypothesis—the prison environment itself creates brutal, dehumanizing conditions independent of the kinds of people who live and work in the institutions.

Haney and associates set out to test the situational hypothesis by creating a functional prison simulation in which healthy, psychologically normal male college students were assigned to roles as prisoners and guards. The “prison” was constructed in the basement of a psychology department building: three 6 x 9 foot “cells” furnished with only a cot, a prison “yard” in a corridor, and a 2 x 7 foot “solitary confinement cell.” Twenty-one subjects were selected from 75 volunteers after screening to eliminate those with physical or psychological problems. Offered $15 per day for their participation, the 21 subjects were randomly assigned to be either guards or prisoners.

All subjects signed contracts that included instructions about prisoner and guard roles for the planned 2-week experiment. “Prisoners” were told that they would be confined and
under surveillance throughout the experiment, and their civil rights would be suspended; they were, however, guaranteed that they would not be physically abused.

“Guards” were given minimal instructions, most notably that physical aggression or physical punishment of “prisoners” was prohibited. Together with a “warden,” they were generally free to develop prison rules and procedures. The researchers planned to study how both guards and prisoners reacted to their roles, but guards were led to believe that the purpose of the experiment was to study prisoners.

If you had been a prisoner in this experiment, you would have experienced something like the following after signing your contract: First, you would have been arrested without notice at your home by a real police officer, perhaps with neighbors looking on. After being searched and taken to the police station in handcuffs, you would have been booked, fingerprinted, and placed in a police detention facility. Next, you would have been blindfolded and driven to “prison,” where you would have been stripped, sprayed with a delousing solution, and left to stand naked for a period of time in the “prison yard.” Eventually, you would have been issued a prison uniform (a loose overshirt stamped with your ID number), fitted with an ankle chain, led to your cell, and ordered to remain silent. Your prison term would then have been totally controlled by the guards.

Wearing mirrored sunglasses, khaki uniforms, and badges and carrying nightsticks, guards supervised prisoner work assignments and held lineups three times per day. Although lineups initially lasted only a few minutes, guards later extended them to several hours. Prisoners were fed bland meals and accompanied by guards on three authorized toilet visits per day.

The behavior of all subjects in the prison yard and other open areas was videotaped; audiotapes were made continuously while prisoners were in their cells. Researchers administered brief questionnaires throughout the experiment to assess emotional changes in prisoners and guards. About 4 weeks after the experiment concluded, researchers conducted interviews with all subjects to assess their reactions.

Haney and associates (1973:88) had planned to run the prison experiment for 2 weeks, but they halted the study after 6 days because subjects displayed “unexpectedly intense reactions.” Five prisoners had to be released even before that time because they showed signs of acute depression or anxiety.

Subjects in each group accepted their roles all too readily. Prisoners and guards could interact with each other in friendly ways because guards had the power to make prison rules. But interactions turned out to be overwhelmingly hostile and negative. Guards became aggressive, and prisoners became passive. When the experiment ended prematurely, prisoners were happy about their early “parole,” but guards were disappointed that the study would not continue.

Haney and colleagues justify the prison simulation study in part by claiming that the dispositional/situational hypotheses could not be evaluated using other research designs. Clearly, the researchers were sensitive to ethical issues. They obtained subjects’ consent to the experiment through signed contracts. Prisoners who showed signs of acute distress were released early. The entire study was terminated after less than half of the planned 2 weeks had elapsed when its unexpectedly harsh impact on subjects became evident. Finally, researchers conducted group therapy debriefing sessions with prisoners and guards and maintained follow-up contacts for a year to ensure that subjects’ negative experiences were temporary.

Two related features of this experiment raise ethical questions, however. First, subjects were not fully informed of the procedures. Although we have seen that deception, including something less than full disclosure, can often be justified, in this case deception was partially due
to the researchers' uncertainty about how the prison simulation would unfold. This relates to the second and more important ethical problem: Guards were granted the power to make up and modify rules as the study progressed, and their behavior became increasingly authoritarian. Comments by guards illustrate their reactions as the experiment unfolded (Haney, Banks, and Zimbardo, 1973:88):

“They [the prisoners] didn't see it as an experiment. It was real and they were fighting to keep their identity. But we were always there to show them just who was boss.”

“During the inspection, I went to cell 2 to mess up a bed which the prisoner had made and he grabbed me, screaming that he had just made it. . . . He grabbed my throat, and although he was laughing, I was pretty scared. I lashed out with my stick and hit him in the chin (although not very hard), and when I freed myself I became angry.”

“Acting authoritatively can be fun. Power can be a great pleasure.”

How do you feel about this experiment? On the one hand, it provided valuable insights into how otherwise normal people react in a simulated prison environment. Subjects appeared to suffer no long-term harm, in part because of precautions taken by researchers. Paul Reynolds (1979:139) found a certain irony in the short-term discomforts endured by the college student subjects: “There is evidence that the major burdens were borne by individuals from advantaged social categories and that the major benefactors would be individuals from less advantaged social categories [actual prisoners], an uneven distribution of costs and benefits that many nevertheless consider equitable.” On the other hand, researchers did not anticipate how much and how quickly subjects would accept their roles. The experiment had to be halted prematurely. In discussing their findings, Haney and associates (1973:90) note: “Our results are . . . congruent with those of Milgram who most convincingly demonstrated the proposition that evil acts are not necessarily the deeds of evil men, but may be attributable to the operation of powerful social forces.” This quote illustrates the fundamental dilemma—balancing the right to conduct research against the rights of subjects. Is it ethical for researchers to create powerful social forces that lead to evil acts?

Discussion Examples

Research ethics, then, is an important and ambiguous topic. The difficulty of resolving ethical issues cannot be an excuse for ignoring them, however. You need to keep ethics in mind as you read other chapters in this book and whenever you plan a research project.

To further sensitize you to the ethical component of criminal justice and other social research, we’ve prepared brief descriptions of 10 real and hypothetical research situations. Can you see the ethical issue in each? How do you feel about it? Are the procedures described ultimately acceptable or unacceptable? It would be very useful to discuss these examples with other students in your class.

1. A researcher studies speeding on urban expressways by using a small radar gun to detect speeders, while an assistant records the license plate numbers of cars traveling more than 10 miles per hour over the posted limit. The speeders’ addresses are traced, and they receive a mailed questionnaire with a cover letter beginning: “You were observed traveling more than 10 miles per hour over the speed limit on [date and location].”

2. Researchers make field observations of social activities as part of a study of police strategies to reduce certain crimes in specified neighborhoods. While making ob-

---

3 Here the authors are referring to controversial research on obedience to authority by Lester Milgram (1965).
servations during a weekday afternoon, a researcher witnesses a residential burglary. The incident is recorded in field notes but not reported to police.

3. In a federally funded study of a probation program, a researcher discovers that one participant was involved in a murder while on probation. Public disclosure of this incident might threaten the program that the researcher believes, from all evidence, is beneficial. Judging the murder to be an anomaly, the researcher does not disclose it to federal sponsors or describe it in published reports.

4. As part of a course on domestic violence, a professor requires students to telephone a domestic violence hotline, pretend to be a victim, and request help. Students then write up a description of the assistance offered by hotline staff and turn it in to the professor.

5. Studying aggression in bars and nightclubs, a researcher records observations of a savage fight in which three people are seriously injured. Ignoring pleas for help from one of the victims, the researcher retreats to a rest room to write up notes from these observations.

6. This quote is from a report on crack dealers by Bruce Jacobs (1996:364–65, n.5):

I informed police of my research, realizing that I could not withhold information from authorities if they should subpoena it. This never happened, perhaps because I told police that the research was about street life and urbanism rather than about gangs and crack distribution per se. Although technically this was a violation of the law, many other sociologists studying deviant populations have done the same thing to acquire valid observational data. . . . In addition, my university’s Human Subject Review committee approved this research because I obtained written informed consent from all respondents and included safeguards to protect their anonymity and confidentiality.

7. In a study of state police, researchers learn that officers have been instructed by superiors to “not sign anything.” Fearing that asking officers to sign informed consent statements will sharply reduce participation, researchers seek some other way to satisfy their university IRB. What should they do?

8. While visiting a police department as part of an evaluation of community policing sponsored by the National Institute of Justice, two researchers accompany a police captain to a local restaurant for lunch. The captain insists on paying. The tab is less than half the listed price of menu items ordered by the captain and the two researchers.

9. A researcher studying juvenile gangs is asked by a federal funding agency not to publish findings indicating that gang members are less often involved in illegal drug sales than are nongang members.

10. In the example mentioned on page 61, the researchers disclosed to public officials that police were not making visits to probationers as called for in the program intervention. Published reports describe the problem as “irregularities in program delivery.”

Main Points

- In addition to technical and scientific considerations, criminal justice research projects are shaped by ethical considerations.
- What’s ethically “right” and “wrong” in research is ultimately a matter of what people agree is right and wrong.
- Researchers tend not to be the best judges of whether their own work adequately addresses ethical issues.
- Most ethical questions involve weighing the possible benefits of research against the potential for harm to research subjects.
- Criminal justice research may generate special ethical questions, including the potential for legal liability and physical harm.
- Scientists agree that participation in research should, in general, be voluntary. This norm,
however, can conflict with the scientific need for generalizability.

• Most scientists agree that research should not harm subjects unless they willingly and knowingly accept the risks of harm.

• Anonymity and confidentiality are two ways to protect the privacy of research subjects.

• Compliance with ethical principles is promoted by professional associations and by regulations issued by the Department of Health and Human Services (HHS).

• HHS regulations include special provisions for two types of subjects of particular interest to many criminal justice researchers: prisoners and juveniles.

• Institutional review boards (IRBs) play an important role in ensuring that the rights and interests of human subjects are protected. But some social science researchers believe that IRBs are becoming too restrictive.

**Key Terms**

- Anonymity
- Confidentiality

**Review Questions and Exercises**

1. Obtain a copy of the Academy of Criminal Justice Sciences’ (2000) code of ethics at this website: [http://www.acjs.org](http://www.acjs.org). Read the document carefully. How would the code apply to Laud Humphreys’ tearoom research? What about the prison simulation?

2. Discuss the general trade-offs between the requirements of sound scientific research methods and the need to protect human subjects. Where do tensions exist? Cite illustrations of tensions from two or more examples of ethical issues presented in this chapter.

3. Review the box “Ethics and Juvenile Gang Members,” noting that Decker and Van Winkle developed an informed consent form for their subjects. Try your hand at preparing such a form, keeping in mind the various ethical principles discussed in this chapter.

**Online Sessions for Chapter 3**

Online Sessions present further examples of selected material covered in each chapter. Your instructor may choose to use these for in-class discussion (as I do with my students) or you can access them at home. To ensure that you always have access to live, correct links for the websites described here, we provide the necessary links on the companion website for *Research Methods for Criminal Justice and Criminology* at [http://www.thomsonedu.com/criminaljustice/maxfield](http://www.thomsonedu.com/criminaljustice/maxfield). Once at the companion website, select this specific chapter, click on “Chapter Resources,” then click on “Web Links.”

**University IRB** The chapter presented excerpts from the Rutgers University materials concerning human subject research. Much more information is available on the human subjects site maintained by the Rutgers Office of Research and Sponsored Programs. Here’s the address: [http://orsp.rutgers.edu/Humans/default.php#general](http://orsp.rutgers.edu/Humans/default.php#general)

We suggest you also visit the IRB site for your own college or university. Use a search engine and type “IRB” and the name of your university.

**Belmont Report** View the Belmont Report at the Department of Health and Human Services site. You will find quite a lot more information here as well, but notice how most issues are at least touched on by the Belmont Report. [http://www.hhs.gov/ohrp](http://www.hhs.gov/ohrp)

**Stanford Prison Simulation** We described some elements of the Stanford Prison Simulation in this chapter. One of the researchers, Phillip Zimbardo, maintains a website for the experiment, linking it to contemporary issues like the abuse of prisoners at Abu Ghraib during the U.S. military action in Iraq. [http://www.prisonexp.org/](http://www.prisonexp.org/)

**Fraud, Conflict of Interest** Biomedical research consumes billions of dollars annually. Funding agencies, federal regulators, and the general public have become more aware of two related problems. First is fraudulent research—fudging data or otherwise presenting results that are known to be false. Second is laundering research sponsored by drug companies. In this case, in-house research is conducted by pharmaceutical companies (known as “pharmas”) to develop and test new drugs. If research finds drugs to be ineffective, no results are published. If research finds some positive impact from the new drugs, physicians or medical researchers are enlisted as paid “authors” of articles ghostwritten by pharma staff, which are then published in medical journals. A website maintained by the Alliance for Human Protection in Research covers these and related questions that affect human subject protection: [http://www.ahrp.org/cms/](http://www.ahrp.org/cms/)
Additional Readings

Als-Nielsen, Bodil, Chen, Wendong, Gluud, Christian, and Kjaergard, Lise L. 2003. “Association of Funding and Conclusions in Randomized Drug Trials: A Reflection of Treatment Effect or Adverse Events?” Journal of the American Medical Association 290 (7, 20 August):921–27. A brief, interesting and nontechnical analysis of research on the effects of new drugs. The authors find that research sponsored by drug companies is much more likely to find that drugs are effective, compared to research sponsored by nonprofit organizations. What do you make of that?

American Association of University Professors, Protecting Human Beings: Institutional Review Boards and Social Science Research (Washington, DC: American Association of University Professors/Redbook, 2001) (http://www.aaup.org/statements/Redbook/repirb.htm; accessed 24 January 2003). Largely in response to concern about overly restrictive institutional review boards, the AAUP convened a series of meetings with representatives from major social science professional groups; this report summarizes the discussion. The most interesting sections address the expansion of human subjects’ protections in social science and even historical research.

Committee on Science, Engineering, and Public Policy, On Being a Scientist: Responsible Conduct in Research, 2nd ed. (Washington, DC: National Academy Press, 1995). This monograph covers a range of issues, including research fraud and other misconduct by researchers. Although many of the issues are specific to the natural sciences, criminologists will find much valuable material.


Sieber, Joan E., Summary of Human Subjects Protection Issues Related to Large Sample Surveys (Washington, DC: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, 2001). The Bureau of Justice Statistics administers the National Crime Victimization Survey, in which almost 100,000 subjects are interviewed every 6 months. Sieber’s report discusses how such large surveys should address a variety of ethical issues, including privacy, information that interviewers are required to report to appropriate authorities, special populations, and special topics in connection with informed consent.
Posing questions properly is often more difficult than answering them. Indeed, a properly phrased question often seems to answer itself. We sometimes discover the answer to a question in the very process of clarifying the question for someone else.

Basically, scientific research is a process for achieving generalized understanding through observation. Part Three of this book will describe some of the specific methods of observation for criminal justice research. But first, Part Two deals with the posing of proper questions, the structuring of inquiry.

Chapter 4 addresses some of the fundamental issues that must be considered in planning a research project. It examines questions of causation, the units of analysis in a research project, the important role of time, and the kinds of things we must consider in proposing to do research projects.

Chapter 5 deals with the specification of what it is we want to study—a process known as conceptualization—and the measurement of the concepts we specify. We’ll look at some of the terms that we use casually in everyday life, and we’ll see how essential it is to be clear about what we really mean by such terms when we do research. Once we are clear on what we mean when we use certain terms, we are in a position to create measurements of what those terms refer to. The process of devising steps or operations for measuring what we want to study is known as operationalization.

Chapter 6 focuses on a specific, difficult, but important measurement problem: measuring crime. We’ll see many different approaches to conceptualization and operationalization and discuss the strengths and weaknesses of each in measuring crime. In doing so, we’ll refer to standards of measurement quality described in Chapter 5.

Chapter 7 concentrates on the general design of a criminal justice research project. A criminal justice research design specifies a strategy for finding out something, for structuring a research project. Chapter 7 describes commonly used strategies for experimental and quasi-experimental research. Each is adapted in some way from the classical scientific experiment.
Chapter 4

General Issues in Research Design

Here, we’ll examine some fundamental principles about conducting empirical research: causation, and variations on whom or what is to be studied, when, and how. We’ll also take a broad overview of the research process.

<table>
<thead>
<tr>
<th>Introduction</th>
<th>81</th>
</tr>
</thead>
<tbody>
<tr>
<td>Causation in the Social Sciences</td>
<td>82</td>
</tr>
<tr>
<td>Criteria for Causality</td>
<td>82</td>
</tr>
<tr>
<td>Necessary and Sufficient Causes</td>
<td>83</td>
</tr>
<tr>
<td>Validity and Causal Inference</td>
<td>84</td>
</tr>
<tr>
<td>Statistical Conclusion Validity</td>
<td>85</td>
</tr>
<tr>
<td>Internal Validity</td>
<td>86</td>
</tr>
<tr>
<td>External Validity</td>
<td>86</td>
</tr>
<tr>
<td>Construct Validity</td>
<td>86</td>
</tr>
<tr>
<td>Validity and Causal Inference Summarized</td>
<td>88</td>
</tr>
<tr>
<td>Does Drug Use Cause Crime?</td>
<td>88</td>
</tr>
<tr>
<td>Introducing Scientific Realism</td>
<td>89</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Units of Analysis</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individuals</td>
<td>91</td>
</tr>
<tr>
<td>Groups</td>
<td>91</td>
</tr>
<tr>
<td>Organizations</td>
<td>92</td>
</tr>
<tr>
<td>Social Artifacts</td>
<td>92</td>
</tr>
<tr>
<td>Units of Analysis in Review</td>
<td>93</td>
</tr>
<tr>
<td>The Ecological Fallacy</td>
<td>94</td>
</tr>
<tr>
<td>Reductionism</td>
<td>95</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The Time Dimension</th>
<th>95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Units of Analysis in the National Youth Gang Survey</td>
<td>96</td>
</tr>
<tr>
<td>Cross-Sectional Studies</td>
<td>96</td>
</tr>
</tbody>
</table>

(continued)
Introduction

Causation, units, and time are key elements in planning a research study.

Science is an enterprise dedicated to “finding out.” No matter what we want to find out, though, there are likely to be a great many ways of doing so. Topics examined in this chapter address how to plan scientific inquiry—how to design a strategy for finding out something. Often, criminal justice researchers want to find out something that involves questions of cause and effect. They may want to find out more about things that make crime more likely to occur or about policies that they hope will reduce crime in some way.

Let’s say you are interested in studying corruption in government. That’s certainly a worthy and appropriate topic for criminal justice research. But what specifically are you interested in? What do you mean by “corruption”? What kinds of behavior do you have in mind? And what do you mean by “government”? Whom do you want to study: All public employees? Only sworn police officers? Civilian employees? Elected officials? What is your purpose? Do you want to find out how much corruption there is? Do you want to learn why corruption exists? These are the kinds of questions that need to be answered in research design.

In practice, all aspects of research design are interrelated. They are separated here and in subsequent chapters so that we can explore particular topics in detail. We start with a discussion of causation in social science, the foundation of explanatory research. We then examine units of analysis—the what or whom to study. Deciding on units of analysis is an important part of all research, partly because people sometimes inappropriately use data measuring one type of unit to say something about a different type of unit.

Next, we consider alternative ways of handling time in criminal justice research. It is sometimes appropriate to examine a static cross section of social life, but other studies
follow social processes over time. In this regard, researchers must consider the time order of events and processes in making statements about cause.

We then provide a brief overview of the whole research process. This serves two purposes: (1) It provides a map to the remainder of this book, and (2) it conveys a sense of how researchers go about designing a study. The chapter concludes with guidelines for preparing a research proposal. Often, the actual conduct of research needs to be preceded by a detailed plan of our intentions—perhaps to obtain funding for a major project or to get an instructor’s approval for a class assignment. We’ll see that preparing a research proposal offers an excellent way to ensure that we have considered all aspects of our research in advance.

Causation in the Social Sciences

_Causation is the focus of explanatory research._

Cause and effect are implicit in much of what we have examined so far. One of the chief goals of social science researchers is to explain why things are the way they are. Typically, we do that by specifying the causes for the way things are: Some things are caused by other things.

The general notion of causation is both simple and complex. On the one hand, we could have ignored the issue altogether. We could simply plug in the terms _cause_ and _effect_ and have little difficulty understanding them. On the other hand, an adequate discourse on causation would require a whole book or even a series of books. We decided to adopt a middle ground, providing more than a commonsense perspective on causation but not attempting to be definitive.

Much of our discussion in this section describes issues of causation and validity for social science in general. Recall from Chapters 1 and 2 that criminal justice research and theory are most strongly rooted in the social sciences. Furthermore, social science research methods are adapted from those used in the physical sciences. Many important and difficult questions about causality and validity occupy researchers in criminal justice, but our basic approach requires stepping back a bit to consider the larger picture of how we can or cannot assert that some cause actually produces some effect.

At the outset, it’s important to keep in mind that cause in social science is inherently _probabilistic_, a point we introduced in Chapter 1. We say, for example, that certain factors make delinquency more or less _likely_ within groups of people. Thus, victims of childhood abuse or neglect are _more likely_ to report alcohol abuse as adults (Schuck and Widom, 2001). Recidivism is _less likely_ among offenders who receive more careful assessment and classification at institutional intake (Cullen and Gendreau, 2000).

Criteria for Causality

We begin our consideration of cause by examining what criteria must be satisfied before we can infer that something causes something else. Recall the discussion of idiographic and nomothetic modes of explanation in Chapter 1. We now consider, first, criteria for inferring idiographic causation, then criteria for inferring cause in the nomothetic mode of explanation.

Joseph Maxwell (2005:106–7), writes that criteria for assessing an idiographic explanation are (1) how credible and believable it is and (2) whether alternative explanations (“rival hypotheses”) were seriously considered and found wanting. The first criterion relates to logic as one of the foundations of science. We demand that our explanations make sense, even if the logic is sometimes complex. The second criterion reminds us of Sherlock Holmes’s dictum that when all other possibilities have been eliminated the remaining explanation, however improbable, must be the truth.

Regarding nomothetic explanation, we examine three specific criteria for causality, as
described by William Shadish, Thomas Cook, and Donald Campbell (2002):

- Two variables must vary together—they must be empirically correlated.
- The cause must be occur before the effect.
- The empirical correlation between cause and effect is not due to some other factor.

The first requirement in a causal relationship between two variables is that the two variables be empirically correlated with each other—they must occur together. It makes no sense to say that exploding gunpowder causes a bullet to leave the muzzle of a gun if, in observed reality, a bullet does not come out after the gunpowder explodes.

As straightforward as it may seem, criminal justice research has difficulties with this requirement. In the probabilistic world of nomothetic models of explanation, at least, we encounter few perfect measures of correlation. Most judges sentence repeat drug dealers to prison, but some don’t. Many gang members use illegal drugs, but many do not. We are forced to ask, therefore, how strong the empirical relationship must be for that relationship to be considered causal.

The second requirement for inferring a causal relationship between two variables is that the cause precede the effect in time. It makes no sense to imagine something being caused by something else that happened later on. A bullet leaving the muzzle of a gun does not cause the gunpowder to explode; it works the other way around.

As simple and obvious as this criterion may seem, criminal justice research suffers many problems in this regard. Often, the time order connecting two variables is simply unclear. Which comes first: drug use or crime? And even when the time order seems clear, exceptions may be found. For example, we normally assume that obtaining a master’s degree in management is a cause of more rapid advancement in a state department of corrections. Yet corrections executives might pursue graduate education after they have been promoted and recognize that advanced training in management skills will help them do their job better.

The third requirement for a causal relationship is that the observed empirical correlation between two variables cannot be explained away as being due to the influence of some third variable that causes both of them. For example, we may observe that drug markets are often found near bus stops, but this does not mean that bus stops encourage drug markets. A third variable is at work here: Groups of people naturally congregate in reasonably close proximity at bus stops, and street drug markets are often found where people naturally congregate.

To sum up, most social researchers consider two variables to be causally related—one causes the other—if (1) there is an empirical correlation between them, (2) the cause precedes the effect in time, and (3) the relationship is not found to result from the effects of some third variable on each of the two initially observed. Any relationship that satisfies all these criteria is causal.

**Necessary and Sufficient Causes**

Recognizing that virtually all causal relationships in criminal justice are probabilistic is central to understanding other points about cause. Within the probabilistic model, it is useful to distinguish two types of causes: necessary and sufficient causes. A **necessary** cause is a condition that must be present for the effect to follow. For example, it is necessary for someone to be charged with a criminal offense before a person can be convicted. But being charged is not sufficient. Some people who are charged are ruled innocent. You must plead guilty or be found guilty by the court. Figure 4.1A illustrates a necessary cause—it is necessary to be charged before one can be convicted.

A **sufficient** cause, in contrast, is a condition that more or less guarantees the effect in question. Thus, for example, pleading guilty to some criminal charge is a sufficient cause for being convicted. It is also possible to be
convicted through a trial as well, but pleading guilty is sufficient. Figure 4.1B illustrates this state of affairs.

The discovery of a cause that is both necessary and sufficient to produce some result is the most satisfying outcome in research. If we are studying juvenile delinquency, we want to discover a single condition that (1) has to be present for delinquency to develop and (2) always results in delinquency. Then we will surely feel that we know precisely what causes juvenile delinquency. Unfortunately, we seldom discover causes that are both necessary and sufficient. Most causal relationships that criminal justice researchers work with are probabilistic and partial—we are able to partly explain cause and effect in some percentage of cases we observe.

**Validity and Causal Inference**

Scientists assess the truth of statements about cause by considering threats to validity.

Paying careful attention to cause-and-effect relationships is crucial in criminal justice research. Cause and effect are also key elements of applied studies, in which a researcher may be interested in whether, for example, a new
mandatory sentencing law actually causes an increase in the prison population.

When we are concerned with whether we are correct in inferring that a cause produced an effect, we are concerned with the validity of causal inference. In the words of Shadish, Cook, and Campbell (2002:34), validity is “the approximate truth of an inference. . . . When we say something is valid, we make a judgment about the extent to which relevant evidence supports that inference as being true or correct.” They emphasize that approximate is an important word because one can never be absolutely certain about cause.

It is almost never possible to conclusively establish the validity of causal inference. We cannot make any sort of absolute judgment about whether our statements about cause are true. Instead, social scientists try to address different validity threats in causal inference—reasons we might be incorrect in stating that some cause produced some effect. As Maxwell (2005:106) puts it, “A key concept for validity is thus the validity threat: a way you might be wrong” (emphasis in original). Here we will summarize the threats to four general categories of validity: statistical conclusion validity, internal validity, construct validity, and external validity. Chapter 7 discusses each type in more detail, linking the issue of validity to different ways of designing research.

Statistical Conclusion Validity

Statistical conclusion validity refers to our ability to determine whether a change in the suspected cause is statistically associated with a change in the suspected effect. This corresponds with the first criterion for inferring cause: Are two variables related to each other? If we suspect that using illegal drugs causes people to commit crimes, one of the first things we will be interested in is the common variation between drug use and crime. If drug users and nonusers commit equal rates of crime, and if about the same proportions of criminals and noncriminals use drugs, there will be no statistical relationship between measures of drug use and criminal offending. That seems to be the end of our investigation of the causal relationship between drugs and crime.

Basing conclusions on a small number of cases is a common threat to statistical conclusion validity. For example, suppose a researcher studies 10 drug users and 10 nonusers and compares the numbers of times these subjects are arrested for other crimes over a 6-month period. The researcher might find that the 10 users were arrested an average of three times while nonusers averaged two arrests in 6 months. There is a difference in arrest rates, but is it a significant difference? Statistically, the answer is no because so few drug users were included in the study. Researchers cannot have much confidence in statements about cause if their findings are based on a small number of cases.

Threats to statistical conclusion validity might also have the opposite effect of suggesting that covariation is present when, in fact, there is no cause-and-effect relationship. The reasons for this are again somewhat technical and require a basic understanding of statistical inference. However, some of the superstitious behavior exhibited by gamblers provides a rough-and-ready example.

Let’s say you like to play the state lottery; nothing serious, you simply buy a couple of tickets each week. Many people play a system whereby they pick numbers that correspond to their date of birth or their license plate number. Expecting these lucky numbers to influence your chances of winning the lottery is an example of failing to recognize that winning numbers are picked by a random process. If you always pick your favorite uncle’s birthday—say, July 7, 1977 (7777)—and buy two tickets per week for 5 years, your lucky number might come up simply by chance in the random system used to pick winning numbers. But there is no statistical conclusion validity to the inference that consistently picking your
birthday will cause you to win the lottery. Your winning will be due to the random effects of how numbers are selected.

**Internal Validity**

Even if we have found that two variables appear to be related to each other, **internal validity** threats challenge causal statements about the observed covariation. An observed association between two variables has internal validity if the relationship is, in fact, causal and not due to the effects of one or more other variables. Whereas statistical conclusion threats are most often due to random error, internal validity problems result from nonrandom or systematic error. Simply put, threats to the internal validity of a proposed causal relationship between two indicators usually arise from the effects of one or more other variables. Notice how this validity threat relates to the third requirement for establishing a causal relationship: eliminating other possible explanations for the observed relationship.

If, for example, we observe that convicted drug users sentenced to probation are rearrested less often than drug users sentenced to prison, we might be tempted to infer that prison sentences cause recidivism. Although being in prison might have some impact on whether someone commits more crimes in the future, in this case it is important to look for other causes of recidivism. One likely candidate would be prior criminal record. Convicted drug users without prior criminal records are more likely to be sentenced to probation, whereas persons with previous convictions more often receive prison terms. Research on criminal careers has found that the probability of reoffending increases with the number of prior arrests and convictions (Chaiken and Chaiken, 1982). In this case, a third variable—prior convictions—may explain some or all of the observed tendency of prison sentences to be associated with recidivism. Prior convictions are associated with both sentence—prison or probation—and subsequent convictions.

**External Validity**

Are findings about the impact of mandatory arrest for family violence in Minneapolis similar to findings in Milwaukee? Can community crime prevention organizations successfully combat drug use throughout a city, or do they work best in areas with only minor drug problems? Electronic monitoring may be suitable as an alternative sentence for convicted offenders, but can it work as an alternative to jail for defendants awaiting trial? Such questions are examples of issues in **external validity**: Do research findings about cause and effect apply equally to different cities, neighborhoods, and populations?

In a general sense, external validity is concerned with whether research findings from one study can be reproduced in another study, often under different conditions. Because crime problems and criminal justice responses can vary so much from city to city or from state to state, researchers and public officials are often especially interested in external validity. For example, a Kansas City evaluation found sharp reductions in gun-related crimes in hot spots that had been targeted for focused police patrols (Sherman and Rogan, 1995). Because these results were promising, similar projects were launched in two other cities—Indianapolis (McGarrell et al., 2001) and Pittsburgh (Cohen and Ludwig, 2003). In both cases, researchers found that police actions targeting hot spots for gun violence reduced gun-related crime and increased seizures of illegal firearms. Having similar findings in Indianapolis and Pittsburgh enhanced the external validity of original results from Kansas City.

**Construct Validity**

This type of validity is concerned with how well an observed relationship between variables that a researcher has measured represents the underlying causal process of interest. In this sense, **construct validity** refers to generalizing from what we observe and measure to the
real-world things in which we are interested. The concept of construct validity is thus closely related to issues in measurement, as we will see in Chapters 5 and 6.

To illustrate construct validity, let’s consider the supervision of police officers—specifically, whether close supervision causes police officers to write more traffic tickets. We might define “close supervision” in this way: A police sergeant drives his own marked police car in such a way as to always keep a patrol car in view.

This certainly qualifies as close supervision, but you may recognize a couple of problems. First, two marked patrol cars present a highly visible presence to motorists, who might drive more prudently and thus reduce the opportunities for patrol officers to write traffic tickets. Second, and more central to the issue of construct validity, this represents a narrow definition of the construct “close supervision.” Patrol officers may be closely supervised in other ways that cause them to write more traffic tickets. For example, sergeants might closely supervise their officers by reviewing their ticket production at the end of each shift. Supervising subordinates by keeping them in view is only one way of exercising control over their behavior. It may be appropriate for factory workers, but it is not practical for police, representing a very limited version of the construct “supervision.”

The well-known Kansas City Preventive Patrol Experiment, discussed in Chapter 1, provides another example of construct validity problems. Recall that the experiment sought to determine whether routine preventive patrol caused reductions in crime and fear of crime and increases in arrests. This causal proposition was tested by comparing measures of crime, fear, and arrests in proactive beats (with twice the normal level of preventive patrol), reactive beats (with no preventive patrol), and control beats (with normal levels of preventive patrol). Researchers found no significant differences in levels of crime, fear, and arrests.

Richard Larson (1975) discussed several difficulties with the experiment’s design. One important problem relates to the visibility of police presence, a central concept in preventive patrol. It is safe to assume that the ability of routine patrol to prevent crime and enhance feelings of safety depends crucially on the visibility of police. It makes sense to assume further that withdrawing preventive patrol, as was done in the reactive beats, reduces the visibility of police. But by how much? Larson explored this question in detail and suggested that two other features of police operations during the Kansas City experiment partially compensated for the absence of preventive patrol and produced a visible police presence.

First, the different types of experimental beats were adjacent to one another; one reactive beat shared borders with three control and three proactive beats. This enhanced the visibility of police in reactive beats in two ways: (1) Police in adjoining proactive and control beats sometimes drove around the perimeter of reactive beats, and (2) police often drove through reactive beats on their way to some other part of the city.

Second, many Kansas City police officers were skeptical about the experiment and feared that withdrawing preventive patrol in reactive beats would create problems. Partly as a result, police who responded to calls for service in the reactive areas more frequently used lights and sirens when driving to the location of complaints. A related action was that police units not assigned to the calls for service nevertheless drove into the reactive beats to provide backup service.

Each of these actions produced a visible police presence in the reactive beats. People who lived in these areas were unaware of the experiment and, as you might expect, did not know whether a police car happened to be present because it was on routine patrol, was on its way to some other part of the city, or was responding to a call for assistance. And, of course, the
use of lights and sirens makes police cars much more visible.

Larson’s point was that the construct of police visibility is only partly represented by routine preventive patrol. A visible police presence was produced in Kansas City through other means. Therefore, the researchers’ conclusion that routine preventive patrol does not cause a reduction in crime or an increase in arrests suffers from threats to construct validity. Construct validity is a frequent problem in applied studies, in which researchers may oversimplify complex policies and policy goals.

**Validity and Causal Inference Summarized**

The four types of validity threats can be grouped into two categories: bias and generalizability. Internal and statistical conclusion validity threats are related to systematic and nonsystematic bias, respectively. Problems with statistical procedures produce nonsystematic bias, while an alternative explanation for an observed relationship is an example of systematic bias. In either case, bias calls into question the inference that some cause produced some effect. Failing to consider the more general cause-and-effect constructs that operate in an observed cause-and-effect relationship results in research findings that cannot be generalized to real-world behaviors and conditions. And a cause-and-effect relationship observed in one setting or at one time may not operate in the same way in a different setting or at a different time.

Shadish, Cook, and Campbell (2002:39) summarized their discussion of these four validity threats by linking them to the types of questions that researchers ask in trying to establish cause and effect. Test your understanding by writing the name of each validity threat after the appropriate question from Cook and Campbell.

1. How large and reliable is the covariation between the presumed cause and effect? ____________
2. Is the covariation causal, or would the same covariation have been obtained without the treatment? ____________
3. How generalizable is the locally embedded causal relationship over varied persons, treatments, observations, and settings? ____________
4. What general constructs are involved in the persons, settings, treatments, and observations used in the experiment? ____________

**Does Drug Use Cause Crime?**

As a way of illustrating issues of validity and causal inference, we will consider the relationship between drug use and crime. Drug addiction is thought to cause people who are desperate for a fix and unable to secure legitimate income to commit crimes in order to support their habit. In recent decades, researchers with both popular culture and criminal justice policy orientations have explored this issue.

Discussing the validity of causal statements about drug use and crime requires carefully specifying two key concepts—drug use and crime—and considering the different ways these concepts might be related. Jan and Marcia Chaiken (1990) provide unusually careful and well-reasoned insights that will guide our consideration of links between drugs and crime. Let’s begin with issues in basic research on the drugs-crime connection and then discuss the implications of this research for criminal justice policy.

First is the question of temporal order: Which comes first, drug use or crime? Research summarized by Chaiken and Chaiken provides no conclusive answer. In an earlier study of prison inmates, Chaiken and Chaiken (1982) found that 12 percent of their adult subjects committed crimes after using drugs for at least 2 years, while 15 percent committed predatory crimes 2 or more years before using drugs.

Many studies have found that some drug users commit crimes and that some criminals use drugs, but Chaiken and Chaiken (1990:234) conclude that “drug use and crime participa-
tion are weakly related as contemporaneous products of factors generally antithetical to traditional United States lifestyles.” Stated somewhat differently, drug use and crime (as well as delinquency) are each deviant activities produced by other underlying causes. A statistical association between drug use and crime clearly exists. But the presence of other factors indicates that the relationship is not directly causal, thus bringing into question the internal validity of causal statements about drug use and crime.

To assess the construct validity of research on drugs and crime, let’s think for a moment about different patterns of each behavior, rather than assume that drug use and crime are uniform behaviors. Many adolescents in the United States experiment with drugs, just as many—especially males—commit delinquent acts or petty crimes. A large number of adults may be occasional users of illegal drugs as well. Many different patterns of drug use, delinquency, and adult criminality have been found in research in other countries. Pudney (2002), and Bennett and Holloway (2005) report on varying patterns of use in England and Wales. Because there is no simple way to describe either construct, searching for a single cause-and-effect relationship misrepresents a complex causal process.

Both construct and external validity are concerned with generalizations. Problems with the external validity of research on drugs and crime are similar to those revolving around construct validity. The relationship between occasional marijuana use and delinquency among teenagers is different from that between occasional cocaine use and adult crime; each relationship, in turn, varies from that between heroin addiction and persistent criminal behavior among adults.

The issue of external validity comes into sharper focus when we shift from basic research that seeks to uncover fundamental causal relationships to criminal justice policy. Chaiken and Chaiken argue that any uniform policy to reduce the use of all drugs among all population groups will have little effect on serious crime.

Basic and applied research on the relationships among drug use and crime readily illustrates threats to the validity of causal inference. It is often difficult to find a relationship because there is so much variation in drug use and crime participation (statistical conclusion validity threat). A large number of studies have demonstrated that, when statistical relationships are found, both drug use and crime can be attributed to other, often multiple, causes (internal validity threat). Different patterns among different population groups mean that there are no readily identifiable cause-and-effect constructs (construct validity). Because of these differences, policies developed to counter drug use among the population as a whole cannot be expected to have much of an impact on serious crime (external validity).

None of the above is to say that there is no cause-and-effect relationship between drug use and crime. However, Chaiken and Chaiken have clearly shown that there is no simple causal connection.

Introducing Scientific Realism

In our final consideration of cause and effect in this chapter, we revisit the distinction between idiographic and nomothetic ways of explanation. Doing research to find what causes what most often reflects nomothetic concerns. We wish to find causal explanations that apply generally to situations beyond those we actually study in our research. At the same time, researchers and public officials are often interested in understanding specific causal mechanisms in more narrowly defined situations—what we have described as the idiographic mode of explanation.

Scientific realism bridges idiographic and nomothetic approaches to explanation by seeking to understand how causal mechanisms operate in specific contexts. Traditional approaches to finding cause and effect usually try to isolate causal mechanisms from other possible influences, something you should now recognize
as trying to control threats to internal validity. The scientific realist approach views these other possible influences as contexts in which causal mechanisms operate. Rather than try to exclude or otherwise control possible outside influences, scientific realism studies how such influences are involved in cause-and-effect relationships.

For example, earlier in this chapter, we noted that electronic monitoring as a condition of probation might apply to some populations but not others. We framed this as a question of external validity in the traditional way of considering nomothetic causation. A scientific realist approach would consider the causal mechanism underlying electronic monitoring to be effective in some contexts but not in others. As another example, we reviewed at some length the cause-and-effect conundrum surrounding drug use and crime. That review was framed by traditional nomothetic research to establish cause and effect. A scientific realism approach to the question would recognize that drug use and crime co-occur in some contexts but not in others.

We say that scientific realism bridges idiographic and nomothetic modes of explanation because it exhibits elements of both. Because it focuses our attention on very specific questions, scientific realism seems idiographic: “Will redesigning the Interstate 78 exit in Newark, New Jersey, cause a reduction in suburban residents seeking to buy heroin in this neighborhood?” But this approach is compatible with more general questions of causation: “Can the design of streets and intersections be modified to make it more difficult for street drug markets to operate?” Changing an expressway exit ramp to reduce drug sales in Newark is a specific example of cause and effect that is rooted in the more general causal relationship between traffic patterns and drug markets.

These illustrations of the scientific realist approach to cause and effect are examples of research for the purpose of application, a topic treated at length by British researchers Ray Pawson and Nick Tilley (1997). Application is a type of explanatory research as we indicated in Chapter 1. In later chapters, we call on scientific realism as a strategy for designing explanatory research (Chapter 7) and conducting evaluations (Chapter 12). Sorting out causes and effects is one of the most difficult challenges of explanatory research. Our attention now turns to two other important considerations that emerge in research for explanation and other purposes: units of analysis and the time dimension.

### Units of Analysis

*To avoid mistaken inferences, researchers must carefully specify the people or phenomena that will be studied.*

In criminal justice research, there is a great deal of variation in what or who is studied—what are technically called *units of analysis.* Individual people are often units of analysis. Researchers may make observations describing certain characteristics of offenders or crime victims, such as age, gender, or race. The descriptions of many individuals are then combined to provide a picture of the population that comprises those individuals.

For example, we may note the age and gender of persons convicted of drunk driving in Fort Lauderdale over a certain period. Aggregating these observations, we might characterize drunk-driving offenders as 72 percent men and 28 percent women, with an average age of 26.4 years. This is a descriptive analysis of convicted drunk drivers in Fort Lauderdale. Although the description applies to the group of drunk drivers as a whole, it is based on the characteristics of individual people convicted of drunk driving.

The same situation could exist in an evaluation study. Suppose we wish to determine whether an alcohol education program reduces repeat arrests of first-time drunk-driving offenders. First, we administer the education program to half the persons convicted of drunk
driving in Fort Lauderdale over a period of 2 months; the other half does not receive the program. Next, we observe drunk-driving arrest records for 12 months, keeping track of people who were previously convicted and of whether they received the alcohol education program. Combining our observations, we find that 5 percent of those who received the program were arrested again and that 20 percent of those who did not receive alcohol education were rearrested within a year. The purpose of the study is to evaluate a program, but individual drunk drivers are still the units of analysis.

Units of analysis in a study are typically also the units of observation. Thus, to study what steps people take to protect their homes from burglary, we might observe, perhaps through interviews, individual household residents. Sometimes, however, we “observe” units of analysis indirectly. For example, we might ask individuals about crime prevention measures for the purpose of describing households. We might want to find out whether homes with double-cylinder deadbolt locks are burglarized less often than homes with less substantial protection. In this case, our units of analysis are households, but the units of observation are individual household members who are asked to describe burglaries and home protection to interviewers.

**Individuals**

Any variety of individuals may be the units of analysis in criminal justice research. This point is more important than it may initially seem. The norm of generalized understanding in social science should suggest that scientific findings are most valuable when they apply to all kinds of people. In practice, however, researchers seldom study all kinds of people. At the very least, studies are typically limited to people who live in a single country, although some comparative studies stretch across national boundaries.

As the units of analysis, individuals may be considered in the context of their membership in different groups. Examples of circumscribed groups whose members may be units of analysis at the individual level are police, victims, defendants in criminal court, correctional inmates, gang members, and active burglars. Note that each of these terms implies some population of individual persons. Descriptive studies having individuals as their units of analysis typically aim to describe the population that comprises those individuals.

**Groups**

A variety of groups may also be the units of analysis for criminal justice research. This is not the same as studying the individuals within a group. If we study the members of a juvenile gang to learn about teenagers who join gangs, the individual (teen gang member) is the unit of analysis. But if we study all the juvenile gangs in a city to learn the differences between big gangs and small ones, between gangs selling drugs and gangs stealing cars, and so forth, the unit of analysis is the social group (gang).

Police beats or patrol districts might be the units of analysis in a study. A police beat can be described in terms of the total number of people who live within its boundaries, total street mileage, annual crime reports, and whether the beat includes a special facility such as a park or high school. We can then determine, for example, whether beats that include a park report more assaults than beats without such facilities or whether auto thefts are more common in beats with more street mileage. Here, the individual police beat is the unit of analysis.

Other examples of units at the group level are households, city blocks, census tracts, cities, counties, and other geographic regions. Each of these terms also implies some population of groups. Street gang implies some population that includes all street gangs. The population of street gangs could be described, say, in terms of its geographic distribution throughout a city. An explanatory study of street gangs might discover, for example, whether large
gangs are more likely than small ones to engage in inter-gang warfare.

**Organizations**

Formal political or social organizations may also be the units of analysis in criminal justice research. An example is correctional facilities, which implies, of course, a population of all correctional facilities. Individual facilities might be characterized in terms of their number of employees, status as state or federal prisons, security classification, percentage of inmates who are from racial or ethnic minority groups, types of offenses for which inmates are sentenced to each facility, average length of sentence served, and so forth. We might determine whether federal prisons house a larger or smaller percentage of offenders sentenced for white-collar crimes than do state prisons. Other examples of formal organizations suitable as units of analysis are police departments, courtrooms, probation offices, drug treatment facilities, and victim services agencies.

When social groups or formal organizations are the units of analysis, their characteristics are often derived from the characteristics of their individual members. Thus, a correctional facility might be described in terms of the inmates it houses—gender distribution, average sentence length, ethnicity, and so on. In a descriptive study, we might be interested in the percentage of institutions housing only females. Or, in an explanatory study, we might determine whether institutions housing both males and females report, on the average, fewer or more assaults by inmates on staff compared with male-only institutions. In each example, the correctional facility is the unit of analysis. In contrast, if we ask whether male or female inmates are more often involved in assaults on staff, then the individual inmate is the unit of analysis.

Some studies involve descriptions or explanations of more than one unit of analysis. Consider, for example, an evaluation of community policing programs in selected neighborhoods of a large city. In such an evaluation, we might be interested in how citizens feel about the program (individuals), whether arrests increased in neighborhoods with the new program compared with those without it (groups), and whether the police department’s budget increased more than the budget in a similar city (organizations). In such cases, it is imperative that researchers anticipate what conclusions they wish to draw with regard to what units of analysis.

**Social Artifacts**

Yet another potential unit of analysis may be referred to as “social artifacts,” or the products of social beings and their behavior. One class of social artifacts is stories about crime in newspapers and magazines or on television. A newspaper story might be characterized by its length, placement on front or later pages, size of headlines, and presence of photographs. A researcher could analyze whether television news features or newspaper reports provide the most details about a new police program to increase drug arrests.

Social interactions are also examples of social artifacts suitable for criminal justice research. Police crime reports are an example. We might analyze assault reports to find how many involved three or more people, whether assaults involve strangers or people with some prior acquaintance, or whether they more often occurred in public or private locations. At first, crime reports may not seem to be social artifacts, but consider for a moment what they represent. When a crime is reported to the police, officers usually record what happened from descriptions by victims or witnesses. For instance, an assault victim may describe how he suffered an unprovoked attack while innocently enjoying a cold beer after work. However, witnesses to the incident might claim that the “victim” started the fight by insulting the “offender.” The responding police officer must interpret who is telling the truth in trying to sort out the circumstances of a violent social
interaction. The officer’s report becomes a social artifact that represents one among the population of all assaults.

Records of different types of social interactions are common units of analysis in criminal justice research. Criminal history records, meetings of community anti-crime groups, presentence investigations, and interactions between police and citizens are examples. Notice that each example requires information about individuals but that social interactions between people are the units of analysis.

In his analysis of juvenile arrests, Howard Snyder (2005) offers examples of how social artifacts sometimes stand in for individuals as units of analysis. Police record what Snyder terms an arrest event. Each arrest event includes at least one person, but because juveniles often commit crimes in groups, a single arrest event might include a large number of juveniles. Furthermore, individual juveniles might be arrested more than once over some time period, so the number of arrests is not the same as the number of juveniles arrested.

Units of Analysis in Review

The purpose of this section has been to specify what is sometimes a confusing topic, in part because criminal justice researchers use a variety of different units of analysis. Although individual people are often the units of analysis, that is not always the case. Many research questions can more appropriately be answered through the examination of other units of analysis.

The concept of unit of analysis may seem more complicated than it needs to be. Understanding the logic of units of analysis is more important than memorizing some list of the units. It is irrelevant what we call a given unit of analysis—a group, a formal organization, or a social artifact. It is essential, however, that we be able to identify what our unit of analysis is. We must decide whether we are studying assaults or assault victims, police departments or police officers, courtrooms or judges, and prisons or prison inmates. Without keeping this point in mind, we run the risk of making assertions about one unit of analysis based on the examination of another.

To test your grasp of the concept of units of analysis, here are some examples of real research topics. See if you can determine the unit of analysis in each. (The answers are given later in the chapter, on page 112.)

1. “Taking into account preexisting traffic fatality trends and several other relevant factors, the implementation of the emergency cellular telephone program resulted in a substantial and permanent reduction in the monthly percentage of alcohol-related fatal crashes” (D’Alessio, Stolzenberg, and Terry, 1999:463–64).

2. “Our analysis provides, at best, extremely weak support for the hypothesis that curfews reduce juvenile crime rates. Of the offense and victimization measures, only burglary, larceny, and simple assault arrests significantly decreased after cities adopted curfew statutes. These decreases occurred only for revised laws, and only the reductions in larceny appeared in both the county and city–county samples” (McDowall, Loftin, and Wiersema, 2000:88).

3. “Despite the record number of homicides in Chicago in 1974, Chicago’s murder rate was actually lower than those in Detroit, Cleveland, Washington, DC, and Baltimore” (Wilson, 1987:22).

4. “On average, probationers were 31 years old, African American, male, and convicted of drug or property offenses. Most lived with family, and although they were not married, many were in exclusive relationships (44 percent) and had children (47 percent)” (MacKenzie, Browning, Skroban, and Smith, 1999:433).

5. “Seventy-five percent (n = 158) of the cases were disposed at district courts, and 3 percent (n = 6) remained pending. One percent of the control and 4 percent of the
experimental cases were referred to drug treatment court” (Taxman and Elis, 1999:42).

6. “The department’s eight Field Operations Command substations (encompassing 20 police districts and 100 patrol beats) comprised the study’s cross-sectional units. From January 1986 through July 1989, a period of 43 months, data were collected monthly from each substation for a total of 344 cases” (Kessler, 1999:346).

7. For most violent crimes, American rates are among the highest—along with Australia, Canada, Spain, and France—but not the highest. Chances of being robbed, being assaulted, or being the victim of a stranger rape are higher in several other Western countries” (Tonry, 1999:421).

8. “Of all the files opened for Arabs, 32 percent were crimes against the person, whereas of all the files opened for Jews, 30 percent were in that category. Proportionally, more files of property and drug offenses were opened for Jews than for Arabs. In addition, Arabs are over-represented in public order offenses” (Mesch and Fishman, 1999:184).

9. “An enormous variation of deviant activities was represented in our sample of 1,485 news items. We categorized these deviant activities into five general types for analysis: violence, economic, political, ideological/cultural, and diversionary” (Ericson, Baranek, and Chan, 1991:243–44).

10. “Approximately half of the [burglars] (54 percent) admitted to 50 or more lifetime burglaries. . . . Included in this group are 44 offenders who had committed at least 100 such crimes. At the other extreme are 11 individuals who had participated in nine or fewer residential break-ins” (Wright and Decker, 1994:13).

The Ecological Fallacy
At this point, it is appropriate to introduce two important concepts related to units of analysis and causation: the ecological fallacy and reductionism. The first concept—the ecological fallacy—refers to the danger of making assertions about individuals as the unit of analysis based on the examination of groups or other aggregations. Such assertions are often made in connection with causation, where researchers observe associations between aggregate units and make statements about causality between individual units.

Suppose we are interested in learning about robbery in different police precincts of a large city. Let’s assume that we have information on how many robberies were committed in each police precinct of Chicago for the year 2005. Assume also that we have census data describing some of the characteristics of those precincts. Our analysis of such data might show that a large number of 2005 robberies occurred in the downtown precinct and that the average family income of persons who live in downtown Chicago (the “Loop”) was substantially higher than that in other precincts in the city. We might be tempted to conclude that high income downtown residents are more likely to be robbed than are people who live in other parts of the city—that robbers select richer victims.

In reaching such a conclusion, we run the risk of committing the ecological fallacy by using observations aggregated into police precincts to make some statement about the socioeconomic status of individual robbery victims. We refer to this as a fallacy because lower-income people who did not live in the downtown area might also have been robbed there in 2005. Victims might be commuters to jobs in the Loop, people visiting downtown theaters or restaurants, passengers on subway or elevated train platforms, or homeless persons who are not counted by the census. Our problem is that we examined police precincts as our units of analysis, but we wish to draw conclusions about individual people.

The same problem will arise if we discover that incarceration rates are higher in states that have a large proportion of elderly residents. We will not know whether older people are actually imprisoned more often. Or, if we find higher suicide rates in cities with large
nonwhite populations, we cannot be sure whether more nonwhites than whites committed suicide.

Don’t let these warnings against the ecological fallacy lead you to commit what is called an “individualistic fallacy.” Some students approaching criminal justice research for the first time have trouble reconciling general patterns of attitudes and actions with individual exceptions they know of. If, for example, you read a newspaper story about a Utah resident visiting New York who is murdered on a subway platform, the fact remains that most visitors to New York and most subway riders are not at risk of murder. Similarly, mass media stories and popular films about drug problems in U.S. cities frequently focus on drug use and dealing among African Americans. But that does not mean that most African Americans are drug users or that drugs are not a problem among whites.

The individualistic fallacy can be especially troublesome for beginning students of criminal justice. Newspapers, local television news, and television police dramas often present unusual or highly dramatized versions of crime problems and criminal justice policy. These messages may distort the way many people initially approach research problems in criminal justice.

Reductionism

A second concept relating to units of analysis is reductionism. Basically, reductionism is an overly strict limitation on the kinds of concepts and variables to be considered as causes in explaining the broad range of human behavior represented by crime and criminal justice policy. Economists may tend to consider only economic variables (marginal value, expected utility); sociologists may consider only sociological variables (values, norms, roles); psychologists may consider only psychological variables (personality types, compulsive personality disorder). For example, why did violent crime by juveniles decline from 1995 through 2002, after an increase from 1985 through about 1994? Was it the result of changes in family structure? Fluctuation in economic opportunities for teenagers? Diminished fascination with images of violence, power, and flashy lifestyles among drug dealers in popular media? Social scientists from different disciplines tend to look at some explanations for crime problems and ignore the others. Explaining crime solely in terms of economic factors is economic reductionism; explaining crime solely in terms of psychological factors is psychological reductionism.

Reductionism of any type tends to suggest that particular units of analysis or variables are more relevant than others. If we consider the changing family structure as the cause of increased juvenile crime, our unit of analysis will be families. An economist, though, might use the 50 states as the units of analysis and compare juvenile crime rates and economic conditions. A psychologist might choose individual juveniles as the units of analysis to determine how watching violent films affects personality development.

Like the ecological fallacy, reductionism involves the use of inappropriate units of analysis. The appropriate unit of analysis for a given research question is not always clear and is often debated by social scientists, especially across disciplinary boundaries.

The box titled “Units of Analysis in the National Youth Gang Survey” offers several additional examples of using inappropriate units of analysis. It also illustrates that lack of clarity about units of analysis in criminal justice results in part from difficulties in directly measuring the concepts we want to study.

The Time Dimension

Because time order is a requirement for causal inferences, the time dimension of research requires careful planning.

We saw earlier in this chapter how the time sequence of events and situations is a critical element in determining causation. Time is also involved in the generalizability of research
findings. Do the descriptions and explanations that result from a particular study accurately represent the situation of 10 years ago or 10 years from now, or do they represent only the present state of affairs? In general, observations may be made more or less at one time point, or they may be deliberately stretched over a longer period. Observations made at more than one time point can look forward or backward.

Cross-Sectional Studies
Many criminal justice research projects are designed to study some phenomenon by taking a cross section of it at one time and analyzing that cross section carefully. Exploratory and descriptive studies are often cross-sectional. A single U.S. census, for instance, is a study aimed at describing the U.S. population at a given time. A single wave of the National Crime Victimization Survey (NCVS) is a descriptive cross-sectional study that estimates how many people have been victims of crime in a given time.

A cross-sectional exploratory study might be conducted by a police department in the form of a survey that examines what residents believe to be the sources of crime problems in their neighborhood. In all likelihood, the study will ask about crime problems in a single

**UNITS OF ANALYSIS IN THE NATIONAL YOUTH GANG SURVEY**

In 1997, the third annual National Youth Gang Survey was completed for the federal Office of Juvenile Justice and Delinquency Prevention (OJJDP). This survey reflects keen interest in developing better information about the scope of youth gangs and their activities in different types of communities around the country. As important and useful as this effort is, the National Youth Gang Survey—especially reports of its results—illustrates how some ambiguities can emerge with respect to units of analysis.

A variety of attempts, often creative, are used to gather information from or about active offenders. Partly this is because it is difficult to systematically identify offenders for research. Studying youth gangs presents more than the usual share of problems with units of analysis. Are we interested in gangs (groups), gang members (individuals), or offenses (social artifact) committed by gangs?

Following methods developed in earlier years, the 1997 National Youth Gang Survey was based on a sample of law enforcement agencies. The sample was designed to represent different types of communities: rural areas, suburban counties, small cities, and large cities. Questionnaires were mailed to the police chief for municipalities and to the sheriff for counties (National Youth Gang Center, 1999:3). Questions asked respondents to report on gangs and gang activity in their jurisdiction—municipality for police departments, and unincorporated service area for sheriffs’ departments. Here are examples of the types of questions included in the survey:

1. How many youth gangs were active in your jurisdiction?
2. How many active youth gang members were in your jurisdiction?
3. In your jurisdiction, what percent of street sales of drugs were made by youth gang members? [followed by list: powder cocaine, crack cocaine, marijuana, heroin, methamphetamine, other]
4. Does your agency have the following? [list of special youth gang units]

Notice the different units of analysis embedded in these questions. Seven are stated or implied.
time frame, with the findings used to help the department explore various methods of introducing community policing.

Cross-sectional studies for explanatory or evaluation purposes have an inherent problem. We have seen that inferring cause requires that a cause precede an effect in time, but cross-sectional studies produce observations made at only one time. For example, a survey might ask respondents whether their home has been burglarized and whether they have any special locks on their doors, hoping to explain whether special locks prevent burglary. Because the questions about burglary victimization and door locks are asked at only one time, it is not possible to determine whether burglary victims installed locks after a burglary or whether special locks were already in place but did not prevent the crime. Some of the ways we can deal with the difficult problem of determining time order will be discussed in the section on approximating longitudinal studies.

**Longitudinal Studies**

Research projects known as **longitudinal studies** are designed to permit observations over an extended period. An example is a researcher who observes the activities of a neighborhood anticrime organization from the time of its inception until its demise. Analysis of
newspaper stories about crime or numbers of prison inmates over time are other examples. In the latter instances, it is irrelevant whether the researcher’s observations are made over the course of the actual events under study or at one time—for example, examining a year’s worth of newspapers in the library or 10 years of annual reports on correctional populations.

Three special types of longitudinal studies should be noted here: trend, cohort, and panel studies. **Trend studies** look at changes within some general population over time. An example is a comparison of UCR figures over time, showing an increase in reported crime from 1960 through 1993 and then a decline through 2005. Or a researcher might want to know whether changes in sentences for certain offenses were followed by increases in the number of people imprisoned in state institutions. In this case, a trend study might examine annual figures for prison population over time, comparing totals for the years before and after new sentencing laws took effect.

**Cohort studies** examine more specific populations (cohorts) as they change over time. Typically, a cohort is an age group, such as those people born during the 1980s, but it can also be based on some other time grouping. Cohorts are often defined as a group of people who enter or leave an institution at the same time, such as persons entering a drug treatment center during July, offenders released from custody in 2002, or high school seniors in March 2003.

In a classic cohort study, Marvin Wolfgang and associates (Wolfgang, Figlio, and Sellin, 1972) studied all males born in 1945 who lived in the city of Philadelphia from their 10th birthday through age 18 or older. The researchers examined records from police agencies and public schools to determine how many boys in the cohort had been charged with delinquency or arrested, how old they were when first arrested, and what differences there were in school performance between delinquents and nondelinquents.

**Panel studies** are similar to trend and cohort studies except that observations are made on the same set of people on two or more occasions. The NCVS is a good example of a descriptive panel study. A member of each household selected for inclusion in the survey is interviewed seven times at 6-month intervals. The NCVS serves many purposes, but it was developed initially to estimate how many people were victims of various types of crimes each year. It is designed as a panel study so that persons can be asked about crimes that occurred in the previous 6 months, and two waves of panel data are combined to estimate the nationwide frequency of victimization over a 1-year period. Panel studies are often used in evaluation research, in which the same persons are interviewed both before and after a new program is introduced.

Though longitudinal studies tend to be expensive and difficult to conduct, panel studies face a special problem: panel attrition. Some of the respondents studied in the first wave of a study may not participate in later waves. The danger is that those who drop out of the study may not be typical and may thereby distort the results of the study. Suppose we are interested in evaluating the success of a new drug treatment program by conducting weekly drug tests on a panel of participants for a period of 10 months. Regardless of how successful the program appears to be after 10 months, if a substantial number of people drop out of our study, we can expect that treatment was less effective in keeping them off drugs.

**Approximating Longitudinal Studies**

It may be possible to draw conclusions about processes that take place over time even when only cross-sectional data are available. It is worth noting some of the ways to do that.

Sometimes, cross-sectional data imply processes that occur over time on the basis of simple logic. Consider, for example, a study of gun ownership and violence by Swiss researcher
Martin Killias (1993). Killias compared rates of gun ownership as reported in an international crime survey to rates of homicide and suicide committed with guns. He was interested in the possible effects of gun availability on violence: Do nations with higher rates of gun ownership also have higher rates of gun violence?

Killias reasoned that inferring causation from a cross-sectional comparison of gun ownership and homicides committed with guns would be ambiguous. Gun homicide rates could be high in countries with high gun ownership rates because the availability of guns was higher. Or people in countries with high gun ownership rates could have bought guns to protect themselves, in response to rates of homicide. Cross-sectional analysis would not make it possible to sort out the time order of gun ownership and gun homicides.

But does that reasoning hold for gun suicides? Killias argued that the time order in a relationship between gun ownership and gun suicides is less ambiguous. It makes much more sense that suicides involving guns are at least partly a result of gun availability. But it is not reasonable to assume that people might buy guns in response to high gun suicide rates.

Logical inferences like that made by Killias may also be sometimes made whenever the time order of variables is clear. For example, if we discover in a cross-sectional study of high school students that men are more likely than women to smoke marijuana, we can conclude that gender might affect the propensity to use marijuana, not the other way around. Thus, even though our observations are made at only one time, we are justified in drawing conclusions about processes that take place across time.

Retrospective Studies

Retrospective research, which asks people to recall their pasts, is another common way of approximating observations over time. In a study of recidivism, for example, we might select a group of prison inmates and analyze their history of delinquency or crime. Or suppose we are interested in whether college students convicted of drunk driving are more likely to have parents with drinking problems than college students with no drunk-driving record. Such a study is retrospective because it focuses on the histories of college students who have or have not been convicted of drunk driving.

The danger in this technique is evident. Sometimes, people have faulty memories. Retrospective recall is one way of approximating observations across time, but it must be used with caution. Retrospective studies that analyze records of past arrests or convictions suffer from different problems: Records may be unavailable, incomplete, or inaccurate.

A more fundamental problem in retrospective research hinges on how subjects are selected and how subject selection affects the kinds of questions such studies can address. Imagine that you are a juvenile court judge, and you’re troubled by what appears to be a large number of child abuse cases in your court. Talking with a juvenile caseworker, you wonder whether the parents of these children were abused or neglected during their own childhood. Together, you formulate a hypothesis about the intergenerational transmission of violence: Victims of childhood abuse later abuse their own children. How might you go about investigating that hypothesis?

Given your position as a judge who regularly sees abuse victims, you will probably consider a retrospective approach that examines the backgrounds of families appearing in your court. Let’s say you and the caseworker plan to investigate the family backgrounds of 20 abuse victims who appear in your court during the next 3 months. The caseworker consults with a clinical psychologist from the local university and obtains copies of a questionnaire, or “protocol,” that has been used by researchers to study the families of child abuse victims. After interviewing the families of 20 victims, the caseworker reports to you that 18 of the 20 child victims have a mother or father who was abused as a child. It seems safe to
conclude that your hypothesis about the intergenerational transmission of violence is strongly supported, because 90 percent (18 out of 20) of abuse/neglect victims brought before your court come from families with a history of child abuse.

Think for a moment about how you approached the question of whether child abuse breeds child abuse. You began with abuse victims and retrospectively established that many of their parents had been abused. However, this is different from the question of how many victims of childhood abuse later abuse their own children. That question requires a prospective approach, in which you begin with childhood victims and then determine how many of them later abuse their own children.

To clarify this point, let’s shift from the hypothetical study to actual research that illustrates the difference between prospective and retrospective approaches to the same question. Rosemary Hunter and Nancy Kilstrom (1979) conducted a study of 255 infants and their parents. The researchers began by selecting families of premature infants in a newborn intensive care unit. Interviews with the parents of 255 infants revealed that either the mother or the father in 49 of the families had been the victim of abuse or neglect; 206 families revealed no history of abuse. In a prospective follow-up study, Hunter and Kilstrom found that within 1 year, 10 of the 255 infants had been abused. Nine of those 10 infant victims were from the 49 families with a history of abuse, while 1 abused infant was from the 206 families with no background of abuse.

Figure 4.2A illustrates these prospective results graphically. Infants in 18 percent (9 out of 49) of families with a history of abuse showed signs of abuse within 1 year of birth, while less than 1 percent of infants born to parents with no history of abuse were themselves abused within 1 year. Although that is a sizable difference, notice that the 18 percent figure for continuity of abuse is very similar to the 19 percent rate of abuse discovered in the histories of all 255 families.

Now consider what Hunter and Kilstrom would have found if they had begun with the 10 abused infants at time 2 and then checked their family backgrounds. Figure 4.2B illustrates this retrospective approach. A large majority of the 10 infant victims (90 percent) had parents with a history of abuse.

You probably realize by now that the prospective and retrospective approaches address fundamentally different questions, even though the questions may appear similar on the surface:

Prospective: What percentage of abuse victims later abuse their children?
(18 percent; Figure 4.2A)

Retrospective: What percentage of abuse victims have parents who were abused?
(90 percent; Figure 4.2B)

In a study of how child abuse and neglect affect drug use, Cathy Spatz Widom and associates (Widom, Weiler, and Cotler, 1999) present a similar contrast of prospective and retrospective analysis. Looking backward, 75 percent of subjects with a drug abuse diagnosis in semi-clinical interviews were victims of childhood abuse or neglect. Looking forward, 35 percent of childhood victims and 34 percent of nonvictims had a drug abuse diagnosis.

More generally, Robert Sampson and John Laub (1993:14) comment on how retrospective and prospective views yield different interpretations about patterns of criminal offending over time:

Looking back over the careers of adult criminals exaggerates the prevalence of stability. Looking forward from youth reveals the success and failures, including adolescent delinquents who go on to be normal functioning adults. This is the paradox noted [by Lee Robins] earlier: adult criminality seems to be always preceded by childhood misconduct, but most conduct-disordered children do not become antisocial or criminal adults (Robins, 1978). (emphasis in original)
Our intention here is not to suggest that retrospective studies have no value. Rather, we want to point out how the time dimension is linked to the framing of research questions. A retrospective approach is limited in its ability to reveal how causal processes unfold over time. A retrospective approach is therefore not well suited to answer questions such as how many childhood victims of abuse or neglect later abuse their own children. A retrospective study can be used, however, to compare whether childhood victims are more likely than nonvictims to have a history of abuse in their family background.

**The Time Dimension Summarized**

Joel Devine and James Wright (1993:19) offer a clever metaphor that distinguishes longitudinal studies from cross-sectional ones. Think of a cross-sectional study as a snapshot, a trend study as a slide show, and a panel study as a snapshot taken at two points in time.
Causation in Traffic Stops
What causes observed proportions of traffic stops by race of driver? That is the central question confronted by researchers and public officials who tried to sort out the problem of racial profiling. Studies in three states framed these questions somewhat differently, but each was concerned with one descriptive question, then a family of causal questions:

1. Are there racial or ethnic disparities in the traffic stops?
2. If yes, what causes these disparities?
   - Overt discrimination by police?
   - Something else?

The first question refers to covariation as a criterion for cause. It is necessary to show disparities, or disproportionality by race before determining that discrimination exists. In other words, if minority motorists are not being stopped or ticketed in greater proportion to their numbers in the population, then discrimination does not exist.

Regarding the second question, most news reports and horror stories assumed some form of discrimination was responsible for any observed disparities. In fact, some researchers seemed to believe that discrimination was the only possible reason for disparities in traffic stops (for example, Lamberth, 2003). Others recognized that different explanations could be at work. For example, Smith et al. wrote:

... discussions of racial disparity in policing require a good faith effort to account for non-discriminatory sources of racial disparity in stops associated with driving behavior before reaching a conclusion that a particular police force is guilty of racial profiling in traffic stops. (2003:39)

Maxfield and Kelling (2005:5–6) addressed a more narrow research question: Why was the proportion of minority drivers stopped substantially higher for one segment of the New Jersey Turnpike than for other segments? Their research considered five possible causal factors in addition to actual discrimination:

- **Deployment**: Do more troopers happen to be deployed in areas with high proportions of minority drivers?
- **Behavior**: In comparison to whites, do minority motorists disproportionately engage in illegal driving practices that bring them to the attention of the New Jersey State Police?
- **Instrumentation**: Are New Jersey State Police troopers now more accurately and/or more consistently reporting the race and ethnicity of traffic violators in comparison to past practices?
- **Evasion**: Are white traffic offenders on the turnpike more able to avoid detection by the New Jersey State Police than minority traffic offenders?
- **Interaction**: Is the continuity over time of racially disproportionate stops in the stop data caused by some mix of the above hypotheses?

Similar alternative explanations were considered by most other researchers in most other areas of the country. In later chapters we describe steps taken to address these questions.

**Units**

Studying traffic enforcement and the possibility of race-based discrimination involves several different units of analysis. The studies done in Pennsylvania and North Carolina collected data on several units, moving from the least to the greatest level of aggregation:
traffic stops: driver characteristics and reasons for stop
■ trooper: aggregate characteristics of all drivers and stops; plus data on individual trooper
■ small-area segment: aggregate characteristics of all troopers, drivers, and stops; plus data on road segment and individual area
■ station or county: aggregates from lower-level units; plus summary data on county residents

Then, of course, studies done in different states reflected the laws and other features of each state. For example, in Pennsylvania troopers are prohibited from stopping cars traveling less than 7 mph above the posted speed limit. New Jersey has no stated policy, but informal norms tolerate speeding up to about 15 mph over the posted limit on many expressways.

Time Dimension
The time dimension revealed certain trends in traffic enforcement. In North Carolina, researchers found a sharp decline in traffic enforcement during 1999, when the state legislature was considering laws to monitor the possibility of racial profiling (Smith et al. 2003:28–29). This documents what Engel and associates term “disengagement,” where police cut back on enforcement activities to avoid getting into trouble.

The time dimension was one factor triggering research in New Jersey. The figure below shows, over a 2-year period, what proportion of drivers stopped for moving violations was black. Two generally stable trend lines are presented. The lower one shows data for the entire New Jersey Turnpike, while the upper line represents the southern segment, about 50 miles long. You can see that the proportion of stopped drivers who were black is about twice as high for the southern segment as for the entire turnpike. The size and stability of this difference was one of the main reasons for launching the New Jersey study. The time dimension here played a role in documenting the persistence of race disparities over time.
motion picture. A cross-sectional study, like a snapshot, produces an image at one point in time. This can provide useful information about crime—burglary, for example—at a single time, perhaps in a single place. A trend study is akin to a slide show—a series of snapshots, in sequence over time. By viewing a slide show, we can tell how some indicator—change in burglary rates—varies over time. But a trend study is usually based on aggregate information. It can tell us something about aggregations of burglary over time, but not, for instance, whether the same people are committing burglaries at an increasing or decreasing rate, or whether there are more or fewer burglars with a relatively constant rate of crime commission. A panel study, like a motion picture, can capture moving images of the same individuals and give us information about individual rates of offending over time.

These, then, are some of the ways time figures in criminal justice research and some of the ways researchers have learned to cope with it. In designing any study, you need to look at both the explicit and implicit assumptions they are making about time. Are you interested in describing some process that occurs over time, such as whether mandatory jail sentences reduce drunk driving? Or are you simply going to describe how many people were arrested for drunk driving in the past year? If you want to describe a process that occurs over time, will you be able to make observations at different points in the process, or will you have to approximate such observations, drawing logical inferences from what you can observe now?

How to Design a Research Project

Designing research requires planning several stages, but the stages do not always occur in the same sequence.

We’ve now seen some of the options available to criminal justice researchers in designing projects, but what if you were to conduct your own research? Where would you start? Then where would you go? How would you begin planning your research?

Although research design occurs at the beginning of a research project, it involves all the steps of the subsequent project. The comments that follow, then, should give you some guidance on how to start a research project, as well as provide an overview of the topics addressed in subsequent chapters of the book.

Every project has a starting point, but it is important to think through later stages even at the beginning. Figure 4.3 presents a schematic view of the social science research process. We present this view reluctantly because it may suggest more of a “cookbook” approach to research than is the case in practice. Nonetheless, it’s important to have an overview of the whole process before we launch into the details of particular components of research.

The Research Process

At the top of the diagram in Figure 4.3 are interests, ideas, theories, and new programs—the possible beginning points for a line of research. The letters (A, B, X, Y, and so forth) represent variables or concepts such as deterrence or child abuse. Thus, you might have a general interest in finding out why the threat of punishment deters some, but not all, people from committing crimes, or you might want to investigate how burglars select their targets. Alternatively, your inquiry might begin with a specific idea about the way things are. You might have the idea that aggressive arrest policies deter drug use, for example. Question marks in the diagram indicate that you aren’t sure things are the way you suspect they are. We have represented a theory as a complex set of relationships among several variables (A, B, E, and F).

The research process might also begin with an idea for a new program. Imagine that you are the director of a probation services department, and you want to develop a new program to require weekly drug tests for people on probation.
Specify the meaning of the concepts and variables to be studied.

How will we actually measure the variables under study?

Experiments
Survey research
Field research
Content analysis
Existing data research
Comparative research
Evaluation research

Whom do we want to be able to draw conclusions about? Who will be observed for that purpose?

Collecting data for analysis and interpretation

Transforming the data collected into a form appropriate to manipulation and analysis

Analyzing data and drawing conclusions

Reporting results and assessing their implications

Figure 4.3 The Research Process
Because you have taken a course on criminal justice research methods, you decide to design an evaluation of the new program before trying it out. The research process begins with your idea for the new drug-testing program.

Notice the movement back and forth among these several possible beginnings. An initial interest may lead to the formulation of an idea, which may be fit into a larger theory, and the theory may produce new ideas and create new interests. Or your understanding of some theory may encourage you to consider new policies.

To make this discussion more concrete, let’s take a specific research example. Suppose you are concerned about the problem of crime on your campus, and you have a special interest in learning more about how other students view the issue and what they think should be done about it. Going a step further, let’s say you have the impression that students are especially concerned about violent crimes such as assault and robbery, and that many students feel the university should be doing more to prevent violent crime. The source of this idea might be your own interest after being a student for a couple of years. You might develop the idea while reading about theories of crime in a course you are taking. Perhaps you recently read stories about a “crime wave” on campus. Or maybe some combination of things makes you want to learn more about campus crime.

The purpose of your research will be mainly exploratory. You probably have descriptive and explanatory interests as well: How much of a problem is violent crime on campus? Are students especially concerned about crime in certain areas? Why are some students more worried about crime than others? What do students think would be effective changes to reduce campus crime problems?

At this point, you should begin to think about units of analysis and the time dimension. Your interest in violent crime might suggest a study of crimes reported to campus police in recent years. In this case, the units of analysis will be social artifacts (crime reports) in a longitudinal study (crime reports in recent years). Or, after thinking a bit more, you may be interested in current student attitudes and opinions about violent crime. Here, the units of analysis will be individuals (college students), and a cross-sectional study will suit your purposes nicely.

**Getting Started**

To begin pursuing your interest in student concerns about violent crime, you undoubtedly will want to read something about the issue. You might begin by finding out what research has been done on fear of crime and on the sorts of crime that concern people most. Your university library and various websites will be important sources for learning about past research. Newspaper stories should provide information on the violent crimes that occurred recently on campus. Perhaps your university posts crime reports on its website. In addition, you will probably want to talk to people—other students or campus police officers, for example. These activities will prepare you to handle the various research design decisions we are about to examine. As you review the research literature, you should make note of the designs used by other researchers, asking whether the same designs will meet your research objective.

What is your objective, by the way? It’s important that you are clear about that before you design your study. Do you plan to write a paper based on your research to satisfy a course requirement or as an honors thesis? Is your purpose to gain information that will support an argument for more police protection or better lighting on campus? Do you want to write an article for the campus newspaper or an academic journal?

Usually, your objective for undertaking research can be expressed in a report. Appendix C on the website for this book will help you with the organization of research reports, and we recommend that you outline such a report as the first step in the design of any project.
You should be clear about the kinds of statements you will want to make when the research is complete. Here are two examples of such statements: “x percent of State U students believe that sexual assault is a big problem on campus,” and “Female students living off campus are more likely than females living in dorms to feel that emergency phones should be installed near buildings where evening classes are held.” Although your final report may not look much like your initial image of it, outlining the planned report will help you make better decisions about research design.

Conceptualization

We often talk casually about criminal justice concepts such as deterrence, recidivism, crime prevention, community policing, and child abuse, but it’s necessary to specify what we mean by these concepts in order to do research on them. Chapter 5 will examine this process of conceptualization in depth. For now, let’s see what it might involve in our hypothetical example.

If you are going to study student concerns about violent crime, you must first specify what you mean by “concern about violent crime.” This ambiguous phrase can mean different things to different people. Campus police officers are concerned about violent crime because that is part of their job. On the one hand, students might be concerned about crime in much the same way they are concerned about other social problems, such as homelessness, animal rights, and the globalized economy. They recognize these issues as problems society must deal with, but they don’t feel that the issues affect them directly; we could specify this concept as “general concern about violent crime.” On the other hand, students may feel that the threat of violent crime does affect them directly, and they express some fear about the possibility of being a victim; let’s call this “fear for personal safety.”

Obviously, you need to specify what you mean by the term in your research, but this doesn’t necessarily mean you have to settle for a single definition. In fact, you might want to define the concept of “concern about violent crime” in more than one way and see how students feel about each.

Of course, you need to specify all the concepts you wish to study. If you want to study the possible effect of concern about crime on student behavior, you’ll have to decide whether you want to limit your focus to specific precautionary behavior such as keeping doors locked or general behavior such as going to classes, parties, and football games.

Choice of Research Method

A variety of methods are available to the criminal justice researcher. Each method has strengths and weaknesses, and certain concepts are more appropriately studied by some methods than by others.

A survey is the most appropriate method for studying both general concern and fear for personal safety. You might interview students directly or ask them to fill out a questionnaire. As we’ll see in Chapter 9, surveys are especially well suited to the study of individuals’ attitudes and opinions. Thus, if you wish to examine whether students who are afraid of crime are more likely to believe that campus lighting should be improved than students who are not afraid, a survey is a good method.

This is not to say that you can’t make good use of other methods presented in Part Three. Through content analysis (discussed in Chapter 11), for example, you might examine letters to the editor in your campus newspaper and analyze what the writers believe should be done to improve campus safety. Field research (see Chapter 10), in which you observe whether students tend to avoid dark areas of the campus, will help you understand student behavior in avoiding certain areas of the campus at night. Or you might study official complaints made to police and college administrators about crime problems on campus. As you read Part Three, you’ll see ways other research methods might be used to study this topic. The best
study design is often one that uses more than one research method, taking advantage of their different strengths.

**Operationalization**

Having specified the concepts to be studied and chosen the research method, you now must create concrete measurement techniques. **Operationalization**, discussed in detail in Chapter 5, refers to the concrete steps or operations used to measure specific concepts.

If you decide to study concern about violent crime by a survey, your operationalization will take the form of questionnaire items. You might operationalize fear for personal safety with the question “How safe do you feel alone on the campus after dark?” This could be followed by boxes indicating the possible answers “Safe” and “Unsafe.” Student attitudes about ways of improving campus safety could be operationalized with the item “Listed below are different actions that might be taken to reduce violent crime on campus. Beside each description, indicate whether you favor or oppose the actions described.” This could be followed by several different actions, with “Favor” and “Oppose” boxes beside each.

**Population and Sampling**

In addition to refining concepts and measurements, decisions must be made about whom or what to study. The population for a study is that group (usually of people) about whom we want to be able to draw conclusions. We are almost never able to study all the members of the population that interests us, however. In virtually every case, we must sample subjects for study. Chapter 8 describes methods for selecting samples that adequately reflect the whole population that interests us. Notice in Figure 4.3 that decisions about population and sampling are related to decisions about the research method to be used.

In the study of concern about violent crime, the relevant population is the student population of your college. As you’ll discover in Chapter 8, however, selecting a sample requires you to get more specific than that. Will you include part-time as well as full-time students? Only degree candidates or everyone? Students who live on campus, off campus, or both? There are many such questions, and each must be answered in terms of your research purpose. If your purpose is to study concern about sexual assault, you might consider limiting your population to female students. If hate crimes are of special interest, you will want to be sure that your study population included sufficient numbers of minorities and others who are thought to be particularly targeted by hate crimes.

**Observations**

Having decided what to study, among whom, and by what method, you are ready to make observations—to collect empirical data. The chapters of Part Three, which describe various research methods, discuss the different observation methods appropriate to each.

For a survey of concern about violent crime, you might prepare a web-based questionnaire distributed via e-mail to a sample selected from the student body. Or you could have a team of interviewers conduct the survey over the telephone. The relative advantages and disadvantages of these and other possibilities are discussed in Chapter 9.

**Data Processing and Analysis**

Finally, we manipulate the collected data for the purpose of drawing conclusions that reflect on the interests, ideas, and theories that initiated the inquiry. Chapter 13 describes a few of the many options available to you in analyzing data. Notice in Figure 4.3 that the results of your analyses feed back into your initial interests, ideas, and theories. In practice, this feedback may initiate another cycle of inquiry. In the study of student concern about violent crime, the analysis phase will have both descriptive and explanatory purposes. You might begin by calculating the percentage of students who feel afraid to use specific parking facilities
Chapter 4  General Issues in Research Design

109

after dark and the percentage who favor or oppose each of the different things that might be done to improve campus safety. Together, these percentages will provide a good picture of student opinion on the issue.

Moving beyond simple description, you might examine the opinions of different subsets of the student body: men versus women; freshmen, sophomores, juniors, seniors, and graduate students; and students who live in dorms versus off-campus apartments. You might then conduct some explanatory analysis to make the point that students who are enrolled in classes that meet in the evening hours are most in favor of improved campus lighting.

Application

The final stage of the research process involves using the research you’ve conducted and the conclusions you’ve reached. To start, you will probably want to communicate your findings so that others will know what you’ve learned. It may be appropriate to prepare—and even publish—a written report. Perhaps you will make oral presentations in class or at a professional meeting. Or you might prepare a web page that presents your results. Other students will be interested in hearing what you have learned about their concerns about violent crime on campus.

Your study might also be used to actually do something about campus safety. If you find that a large proportion of students you interviewed believe that a parking lot near the library is poorly lighted, university administrators could add more lights, or campus police might patrol the area more frequently. Crime prevention programs might be launched in dormitories if residents are more afraid of violent crime than students who live in other types of housing. Students in a Rutgers University class on crime prevention focused on car thefts and break-ins surrounding the campus in Newark, New Jersey. Their semester project presented specific recommendations on how university and city officials could reduce the problem.

Finally, you should consider what your research suggests with regard to further research on your subject. What mistakes should be corrected in future studies? What avenues, opened up slightly in your study, should be pursued further in later investigations?

Research Design in Review

As this overview shows, research design involves a set of decisions regarding what topic is to be studied, among what population, with what research methods, and for what purpose. Whereas the earlier sections of this chapter on cause and effect, units of analysis, and the time dimension aimed at broadening your perspective in all these regards, research design is the process of narrowing, of focusing, your perspective for purposes of a particular study.

If you are doing a research project for a course you are taking, many aspects of research design may be specified for you in advance. For a project in a course on experimental methods in criminal justice, the design for selecting subjects will be specified for you. For a project in a course on corrections policy, the research topic will be somewhat specified. Because it is not feasible for us to anticipate all such constraints, the following discussion will assume there are none.

In designing a research project, you will find it useful to begin by assessing three things: (1) your interests, (2) your abilities, and (3) the resources available to you. Each of these considerations will suggest a number of possible studies.

What are you interested in understanding? Surely, you have several questions about crime and possible policy responses. Why do some juvenile gangs sell drugs while others steal cars? Why do particular neighborhoods near campus seem to have higher rates of burglary? What types of community groups are more active in neighborhood anticrime programs? Do sentencing policies discriminate against minorities? Do cities with gun control laws have lower murder rates? Are sentences for rape more
severe in some states than in others? Are mandatory jail sentences more effective than license suspension in reducing repeat drunk-driving offenses? Think for a while about the kinds of questions that interest and concern you.

Once you have a few questions you are interested in answering, think about the kind of information you will need to answer them. What research units of analysis will provide the most relevant information: gangs, burglary victims, drunk drivers, households, community groups, police departments, cities, or states? This question should be inseparable from the question of research topics. Then ask which aspects of the units of analysis will provide the information you need to answer your research question.

Your next consideration is how to obtain that information. Are the relevant data likely to be already available somewhere (say, in a government publication), or will you have to collect them yourself? If you think you will have to collect them, how will you do that? Will it be necessary to observe juvenile gangs, interview a large number of burglary victims, or attend meetings of community crime prevention groups? Or will you have to design an experiment to study sentences for drunk driving?

As you answer these questions, you are well into the process of research design. Keep in mind, however, your own research abilities, the resources available to you, and the time required to complete your research project. Do not design the perfect study if you will be unable to carry it out in a reasonable time. You may want to try a research method you have not used before because research should be a learning experience in many ways, but you should not put yourself at too great a disadvantage.

Once you have a general idea of what you want to study and how, carefully review previous research in journals, books, and government reports to see how other researchers have addressed the topic and what they have learned about it. Your review of the literature may lead you to revise your research design; perhaps you will decide to use another researcher’s method or even replicate an earlier study. The independent replication of research projects is a standard procedure in the physical sciences, and it is just as important in criminal justice research. Or you might want to go beyond replication and study some aspect of the topic that you feel other researchers have overlooked.

Here’s another approach you might take. Suppose a topic has been studied previously using survey methods. Can you design an experimental study to test the findings of those earlier researchers? Or can you think of how a field study might supplement information gained from a survey? The use of several different research methods to test the same finding is sometimes called “triangulation,” and you should keep it in mind as a valuable research strategy. Because each research method has particular strengths and weaknesses, there is always a danger that research findings will reflect, at least in part, the method of inquiry. In the best of all worlds, your own research design should bring more than one research method to bear on the topic.

The Research Proposal

*Research proposals describe planned activities and include a budget and time line.*

If you undertake a research project—an assignment for this course, perhaps, or even a major study funded by the government or a research foundation—you will probably have to provide a research proposal describing what you intend to accomplish and how. We’ll conclude this chapter with a discussion of how you might prepare such a proposal.

Elements of a Research Proposal

Some funding agencies have specific requirements for a proposal’s elements, structure, or both. For example, in its research solicitation announcements for the 2007 fiscal year, the National Institute of Justice (NIJ) describes what should be included in research propos-
als on such topics as terrorism and elder abuse (http://www.ojp.usdoj.gov/nij/funding.htm). Your instructor may have certain requirements for a research proposal you are to prepare in this course. Here are some basic elements that should be included in almost any research proposal.

**Problem or Objective** What exactly do you want to study? Why is it worth studying? Does the proposed study contribute to our general understanding of crime or policy responses to crime? Does it have practical significance? If your proposal describes an evaluation study, then the problem, objective, or research questions may already be specified for you. For example, in its request for research on elder abuse issued in 2006, the NIJ required that proposals address certain specific items in describing the impact of proposed research:

1. Potential for significant advances in scientific or technical understanding of the problem.
2. Potential for significant advances in the field.

**Literature Review** What have others said about this topic? What theories address it, and what do they say? What research has been done? Are the findings consistent, or do past studies disagree? Are there flaws in the body of existing research that you feel you can remedy?

**Research Questions** What specific questions will your research try to answer? Given what others have found, as stated in your literature review, what new information do you expect to find? It’s useful to view research questions as a more specific version of the problem or objective described earlier. Then, of course, your specific questions should be framed in the context of what other research has found.

**Subjects for Study** Whom or what will you study in order to collect data? Identify the subjects in general terms, and then specifically identify who (or what) is available for study and how you will reach them. Is it appropriate to select a sample? If so, how will you do that? If there is any possibility that your research will have an impact on those you study, how will you ensure that they are not harmed by the research? Finally, if you will be interacting directly with human subjects, you will probably have to include a consent form (as described in Chapter 3) in an appendix to your proposal.

**Measurement** What are the key variables in your study? How will you define and measure them? Do your definitions and measurement methods duplicate (that’s okay, incidentally) or differ from those of previous research on this topic? If you have already developed your measurement device (a questionnaire, for example), or if you are using something developed by others, you should include a copy in an appendix to your proposal.

**Data Collection Methods** How will you actually collect the data for your study? Will you observe behavior directly or conduct a survey? Will you undertake field research, or will you focus on the reanalysis of data already collected by others? Criminal justice research often includes more than one such method.

**Analysis** Briefly describe the kind of analysis you plan to conduct. Spell out the purpose and logic of your analysis. Are you interested in precise description? Do you intend to explain why things are the way they are? Will you analyze the impact of a new program? What possible explanatory variables will your analysis consider, and how will you know whether you’ve explained program impact adequately?

**References** Be sure to include a list of all materials you consulted and cited in your proposal. Formats for citations vary. Your instructor may specify certain formats, or refer you to
specific style manuals for guidelines on how to cite books, articles, and web-based resources.

**Schedule**  It is often appropriate to provide a schedule for the various stages of research. Even if you don’t do this for the proposal, do it for yourself. Unless you have a time line for accomplishing the several stages of research and keeping track of how you’re doing, you may end up in trouble.

**Budget**  If you are asking someone to give you money to pay the costs of your research, you will need to provide a budget that specifies where the money will go. Large, expensive projects include budgetary categories such as personnel, equipment, supplies, and expenses (for example, travel, copying, and printing). Even for a more modest project you will pay for yourself, it’s a good idea to spend some time anticipating any expenses involved: office supplies, photocopying, computer disks, telephone calls, transportation, and so on.

If you are interested in conducting a criminal justice research project, it is a good idea to prepare a research proposal for your own purposes, even if you aren’t required to do so by your instructor or a funding agency. If you are going to invest your time and energy in such a project, you should do what you can to ensure a return on that investment.

As it happens, while working on the revisions for the book you are now reading, Maxfield worked with a colleague to write a research proposal submitted to the National Institute of Justice. You’ll find most elements of that proposal (not the budget) on websites for Chapter 4.

**Answers to the Units-of-Analysis Exercise**

1. Social artifacts (alcohol-related fatal crashes)
2. Groups (cities and counties)
3. Groups (cities)
4. Individuals (probationers)
5. Social artifacts (court cases)
6. Organizations (police substations)
7. Groups (countries)
8. Social artifacts (offense files)
9. Social artifacts (news items)
10. Individuals (burglars)

**Main Points**

- Explanatory scientific research centers on the notion of cause and effect.
- Most explanatory social research uses a probabilistic model of causation. $X$ may be said to cause $Y$ if it is seen to have some influence on $Y$.
- $X$ is a necessary cause of $Y$ if $Y$ cannot happen without $X$ having happened. $X$ is a sufficient cause of $Y$ if $Y$ always happens when $X$ happens.
- Three basic requirements determine a causal relationship in scientific research: (1) The independent variable must occur before the dependent variable, (2) the independent and dependent variables must be empirically related to each other, and (3) the observed relationship cannot be explained away as the effect of another variable.
- When scientists consider whether causal statements are true or false, they are concerned with the validity of causal inference.
- Four classes of threats to validity correspond to the types of questions researchers ask in trying to establish cause and effect. Threats to statistical conclusion validity and internal validity arise from bias. External and construct validity threats may limit our ability to generalize from an observed relationship.
- A scientific realist approach to examining mechanisms in context bridges idiographic and nomothetic approaches to causation.
- Units of analysis are the people or things whose characteristics researchers observe, describe, and explain. The unit of analysis in criminal justice research is often the individual person, but it may also be a group, organization, or social artifact.
- Researchers sometimes confuse units of analysis, resulting in the ecological fallacy or the individualistic fallacy.
- Cross-sectional studies are those based on observations made at one time. Although such studies are limited by this characteristic, inferences can often be made about processes that occur over time.
3. In describing different approaches to the time dimension, criminologist Lawrence Sherman (1995) claims that cross-sectional studies can show differences and that longitudinal studies can show change. How does this statement relate to the three criteria for inferring causation?

4. William Julius Wilson (1996:167) cites the following example of why it’s important to think carefully about units and time. Imagine a 13-bed hospital, in which 12 beds are occupied by the same person for 1 year. The other hospital bed is occupied by 52 people, each staying 1 week. At any given time, 92 percent of beds are occupied by long-term patients (12 out of 13); but over the entire year, 81 percent of patients are short-term patients (52 out of 64). Discuss the implications of a similar example, in which “jail cell” is substituted for “hospital bed.”

**Key Terms**

- Cohort study
- Conceptualization
- Construct validity
- Cross-sectional study
- Ecological fallacy
- External validity
- Internal validity
- Longitudinal study
- Operationalization
- Panel study
- Probabilistic

- Prospective
- Reductionism
- Retrospective
- Scientific realism
- Statistical conclusion
- Validity
- Trend study
- Units of analysis
- Validity
- Validity threats

**Online Sessions for Chapter 4**

Online Sessions present further examples of selected material covered in each chapter. Your instructor may choose to use these for in-class discussion (as I do with my students) or you can access them at home. To ensure that you always have access to live, correct links for the websites described here, we provide the necessary links on the companion website for Research Methods for Criminal Justice and Criminology at http://www.thomsonedu.com/criminaljustice/maxfield. Once at the companion website, select this specific chapter, click on “Chapter Resources,” then click on “Web Links.”

**Cause or Something Else?** The first two sentences of an article in the New York Times questioned the validity of auto safety ratings:

One way of reading the new report by the Insurance Institute for Highway Safety is that the Mercedes E-Class sedan has the safest design of any car or truck and the two-door Chevrolet Blazer the worst. Another way to read the report, to be released Tuesday, is that E-Class drivers tend to drive more carefully than Blazer drivers. (Hakim, Danny. 2005. “Is the Car Unsafe, or the Driver?” New York Times, 15 March, p. C1)

Certain makes and models were consistently rated especially safe or unsafe based on the frequency and value of insurance claims. But Hakim
suggests that the design, features, and marketing of different kinds of cars might attract buyers who are especially safe or unsafe drivers. This story is based on annual ratings of collision, injury, and theft claims compiled by the Highway Loss Data Institute (HLDI). Make a list of three vehicles you think might be rated high on safety, and three that you think might be rated low. Check the HLDI ratings and see how well you did. [http://www.iihs.org/brochures/ictl/ictl.html](http://www.iihs.org/brochures/ictl/ictl.html)

**Units of Analysis** The National Youth Gang Center compiles information about youth gangs, including the survey we discussed in Chapter 4. Visit the Center’s website to review various materials, including data from the survey. Look for information on the survey and learn more about how it’s conducted. The New Jersey Department of Law and Public Safety conducts its own gang survey. Check that out, and perhaps look for similar surveys in other states. [http://www.iir.com/NYGC/PublicationLinks.htm](http://www.iir.com/NYGC/PublicationLinks.htm) [http://www.nj.gov/oag/gangsurvey/index.html](http://www.nj.gov/oag/gangsurvey/index.html)

**Proposal to the National Institute of Justice** Maxfield was briefly diverted from working on the 5th edition of the text to collaborate on a research proposal that was submitted to the National Institute of Justice on 20 June 2006. A copy of the proposal and selected appendixes is presented on the companion website. This gives you a real life example of how a proposal is organized.

**Additional Readings**


Chapter 5

Concepts, Operationalization, and Measurement

It’s essential to specify exactly what we mean (and don’t mean) by the terms we use. This is the first step in the measurement process, and we’ll cover it in depth.

Introduction 116
Conceptions and Concepts 116
Conceptualization 118
Indicators and Dimensions 118
Confusion over Definitions and Reality 119
Creating Conceptual Order 119

WHAT IS RECIDIVISM? 122

Operationalization Choices 122

JAIL STAY 124
Measurement as “Scoring” 124
Exhaustive and Exclusive Measurement 126

Levels of Measurement 127
Implications of Levels of Measurement 128

Criteria for Measurement
Quality 130

Reliability 130
Validity 133

PUTTING IT ALL TOGETHER: MEASURING RACE 136

Composite Measures 136
Typologies 138
An Index of Disorder 139

Measurement Summary 140
Main Points 141
Introduction

Because measurement is difficult and imprecise, researchers try to describe the measurement process explicitly.

This chapter describes the progression from having a vague idea about what we want to study to being able to recognize it and measure it in the real world. We begin with the general issue of conceptualization, which sets up a foundation for our examination of operationalization and measurement. We then turn to different approaches to assessing measurement quality. The chapter concludes with an overview of strategies for combining individual measures into more complex indicators. Our examination of measurement continues in the next chapter, where we’ll focus on different strategies for measuring crime.

As you read this chapter, keep in mind a central theme: communication. Ultimately, criminal justice and social scientific research seeks to communicate findings to an audience—professors, classmates, journal readers, or coworkers in a probation services agency, for example. Moving from vague ideas and interests to a completed research report, as we described in Chapter 4, involves communication at every step—from general ideas to more precise definitions of critical terms. With more precise definitions, we can begin to develop measures to apply in the real world.

Conceptions and Concepts

Clarifying abstract mental images is an essential first step in measurement.

If you hear the word recidivism, what image comes to mind? You might think of someone who has served time for burglary and who breaks into a house soon after being released from prison. Or, in contrast to that rather specific image, you might have a more general image of habitual criminals. Someone who works in a criminal justice agency might have a different mental image. Police officers might think of a specific individual they have arrested repeatedly for a variety of offenses, and a judge might think of a defendant who has three prior convictions for theft.

Ultimately, recidivism is simply a term we use in communication—a word representing a collection of related phenomena that we have either observed or heard about somewhere. It’s as though we have file drawers in our minds containing thousands of sheets of paper, and each sheet has a label in the upper right-hand corner. One sheet of paper in your file drawer has the term recidivism on it, and the person who sits next to you in class has one, too.

The technical name for those mental images, those sheets of paper in our file drawers, is conception. Each sheet of paper is a conception—a subjective thought about things that we encounter in daily life. But those mental images cannot be communicated directly. There is no way we can directly reveal what’s written on our mental images. Therefore, we use the terms written in the upper right-hand corners as a way of communicating about our conceptions and the things we observe that are related to those conceptions.

For example, the word crime represents our conception about certain kinds of behavior. But individuals have different conceptions; they may think of different kinds of behavior when they hear the word crime. For example, police officers in most states would include possession of marijuana among their conceptions of crime, whereas members of the advocacy group NORML would not. Recent burglary victims might recall their own experiences in their conceptions of crime, whereas more fortunate neighbors might think about the murder story in yesterday’s newspaper.

Because conceptions are subjective and cannot be communicated directly, we use the words and symbols of language as a way of communicating about our conceptions and the things we observe that are related to those conceptions. Concepts are the words or symbols in language that we use to represent these mental
images. We use concepts to communicate with one another, to share our mental images. Although a common language enables us to communicate, it is important to recognize that the words and phrases we use represent abstractions. Concepts are abstract because they are independent of the labels we assign to them. Crime as a concept is abstract, meaning that in the English language this label represents mental images of illegal acts. Of course, actual crimes are real events, and our mental images of crime may be based on real events (or the stuff of TV drama). However, when we talk about crime, without being more specific, we are talking about an abstraction. Thus, for example, the concept of crime proposed by Michael Gottfredson and Travis Hirschi (1990:15)—using force or fraud in pursuit of self-interest—is abstract. “Crime” is the symbol or label they have assigned to this concept.

Let’s discuss a specific example. What is your conception of “serious crime”? What mental images come to mind? Most people agree that airplane hijacking, rape, bank robbery, and murder are serious crimes. What about a physical assault that results in a concussion and facial injuries? Many of us would classify it as a serious crime, but not if the incident took place in a boxing ring. Is burglary a serious crime? It doesn’t rank up there with drive-by shooting, but we would probably agree that it is more serious than shoplifting. What about drug use or drug dealing?

Our mental images of serious crime may vary depending on our backgrounds and experiences. If your home has ever been burglarized, you might be more inclined than someone who has not suffered that experience to rate it as a serious crime. If you have been both burglarized and robbed at gunpoint, you would probably think the burglary was less serious than the robbery.

Disagreement over the seriousness of drug use continues. Younger people, whether or not they have used drugs, may be less inclined to view drug use as a serious crime, whereas police and other public officials might rank drug use as very serious. California and Oregon are among states that have legalized the use of marijuana for medical purposes. However, as of 2006 the U.S. Department of Justice views all marijuana use as a crime, challenging state laws and raiding San Francisco medical marijuana dispensaries (Murphy, 2005).

“Serious crime” is an abstraction, a label we use to represent a concept. However, we must be careful to distinguish the label we use for a concept from the reality that the concept represents. There are real robberies, and robbery is a serious crime, but the concept of crime seriousness is not real. The use of marijuana for medical purposes is real, but the concept of its seriousness is abstract.

In order to link conceptions, concepts, and measurement, consider Abraham Kaplan’s (1964) discussion of three classes of things that scientists measure: direct observables, indirect observables, and constructs. The first class, direct observables, includes those things that we can observe simply and directly, such as the color of an apple or the words in a crime report. Indirect observables require “relatively more subtle, complex, or indirect observations” (1964:55). We note that a police officer has written “robbery” in the place for “offense type” on a crime report and has thus indirectly observed what crime has occurred. Newspaper stories, court transcripts, and criminal history records provide indirect observations of past actions. Finally, constructs are theoretical creations based on things that cannot be observed directly or indirectly. IQ is a good example: It is constructed mathematically from observations of the answers to questions on an IQ test.

Kaplan (1964:49) defined concept as a “family of conceptions.” A concept is, as Kaplan noted, a construct. The concept of serious crime, then, is a construct created from your conception of it, our conception of it, and the conceptions of all those who have ever used the term. The concept of serious crime cannot be observed directly or indirectly. We can, however, meaningfully
discuss the concept, observe examples of serious crime, and measure it indirectly.

Conceptualization
Day-to-day communication is made possible through general but often vague and unspoken agreements about the use of terms. Usually, other people do not understand exactly what we wish to communicate, but they get the general drift of our meaning. For example, although we may not fully agree about the meaning of the term serious crime, it’s safe to assume that the crime of bank robbery is more serious than the crime of bicycle theft. A wide range of misunderstandings is the price we pay for our imprecision, but somehow we muddle through. Science, however, aims at more than muddling, and it cannot operate in a context of such imprecision.

**Conceptualization** is the process by which we specify precisely what we mean when we use particular terms. Suppose we want to find out whether violent crime is more serious than nonviolent crime. Most of us would probably assume that is true, but it might be interesting to find out whether it’s really so. Notice that we can’t meaningfully study the issue, let alone agree on the answer, without some precise working agreements about the meanings of the terms we are using. They are working agreements in the sense that they allow us to work on the question.

We begin by clearly differentiating violent and nonviolent crime. In violent crimes, an offender uses force or threats of force against a victim. Nonviolent crimes either do not involve any direct contact between a victim and an offender or involve contact but no force. For example, pickpocketing involves direct contact but no force. In contrast, robbery involves at least the threat to use force on victims. Burglary, auto theft, shoplifting, and the theft of unattended personal property such as bicycles are examples of nonviolent crimes. Assault, rape, robbery, and murder are violent crimes.

**Indicators and Dimensions**
The end product of the conceptualization process is the specification of a set of indicators of what we have in mind, indicating the presence or absence of the concept we are studying. To illustrate this process, let’s discuss the more general concept of crime seriousness. This concept is more general than serious crime because it implies that some crimes are more serious than others.

One good indicator of crime seriousness is harm to the crime victim. Physical injury is an example of harm, and physical injury is certainly more likely to result from violent crime than from nonviolent crime. What about other kinds of harm? Burglary victims suffer economic harm from property loss and perhaps damage to their homes. Is the loss of $800 in a burglary an indicator of more serious crime than a $10 loss in a robbery in which the victim was not injured? Victims of both violent crime and nonviolent crime may suffer psychological harm. Charles Silberman (1978:18–19) described how people feel a sense of personal violation after discovering that their home has been burglarized. Other types of victim harm can be combined into groups and subgroups as well.

The technical term for such groupings is **dimension**—some specifiable aspect of a concept. Thus, we might speak of the “victim harm dimension” of crime seriousness. This dimension could include indicators of physical injury, economic loss, or psychological consequences. And we can easily think of other indicators and dimensions related to the general concept of crime seriousness. If we consider the theft of $20 from a poor person to be more serious than the theft of $2000 from a wealthy oil company CEO, victim wealth might be another dimension. Also consider a victim identity dimension. Killing a burglar in self-defense would not be as serious as threatening to kill the president of the United States.

Thus, it is possible to subdivide the concept of crime seriousness into several dimensions.
Specifying dimensions and identifying the various indicators for each of those dimensions are both parts of conceptualization.

Specifying the different dimensions of a concept often paves the way for a more sophisticated understanding of what we are studying. We might observe, for example, that fistfights among high school students result in thousands of injuries per year but that the annual costs of auto theft cause direct economic harm to hundreds of insurance companies and millions of auto insurance policyholders. Recognizing the many dimensions of crime seriousness, we cannot say that violent crime is more serious than nonviolent crime in all cases.

As it happens, defining and measuring crime seriousness is an important issue in criminal justice policy and research. Domestic assaults and acquaintance rape are examples of violent crimes that are often treated differently than physical or sexual violence among strangers, suggesting a victim–offender relationship dimension. For instance, the battered-woman defense has been used in trials of women who kill a spouse or lover following an extended period of violence (Williams, 1991). Such mitigating circumstances mean that some types of murder are treated differently than others, based in part on motivation and the relationship between victim and offender. The tendency of courts to punish rapists less harshly if they had some prior relationship with their victims has led to growing protests (Lopez, 1992). Studies have shown that death penalties were more often imposed on blacks convicted of killing whites, which indicates that victim race is related to sentencing and might therefore be considered an indicator of crime seriousness (Balduz, Pulaski, and Woodworth, 1983). Later in this chapter, we will discuss specific attempts to develop measures of crime seriousness.

Confusion over Definitions and Reality

To review briefly, our concepts are derived from the conceptions (mental images) that summarize collections of seemingly related observations and experiences. Although the observations and experiences are real, our concepts are mental creations. The terms associated with concepts are merely devices created for communication. The term crime seriousness is an example. Ultimately, that phrase is only a collection of letters and has no intrinsic meaning. We could have as easily and meaningfully created the term crime pettiness to serve the same purpose.

Often, however, we fall into the trap of believing that terms have real meanings. That danger seems to grow stronger when we begin to take terms seriously and attempt to use them precisely. And the danger is all the greater in the presence of experts who appear to know more about what the terms really mean. It’s easy to yield to the authority of experts in such a situation.

Once we have assumed (mistakenly) that terms have real meanings, we begin the task of discovering what those real meanings are and what constitutes a genuine measurement of them. Figure 5.1 illustrates this process. We make up conceptual summaries of real observations because the summaries are convenient—so convenient, however, that we begin to think they are real. The process of regarding as real things that are not is called reification. The reification of concepts in day-to-day life is very common.

Creating Conceptual Order

The design and execution of criminal justice research requires that we clear away the confusion over concepts and reality. To this end, logicians and scientists have found it useful to distinguish three kinds of definitions: real, conceptual, and operational. The first of these reflects the reification of terms, and as Carl G. Hempel (1952:6) has cautioned:

A “real” definition, according to traditional logic, is not a stipulation determining the meaning of some expression but a statement of the “essential nature” or the “essential attributes” of some entity. The notion of essential nature, however,
1. Many of our observations in life seem to have something in common. We get the sense that they represent something more general than the simple content of any single observation. We find it useful, moreover, to communicate about the general concept.

2. It is inconvenient to keep describing all the specific observations whenever we want to communicate about the general concept they seem to have in common, so we give a name to the general concept—to stand for whatever it is the specific observations have in common.

3. As we communicate about the general concept, using its term, we begin to think that the concept is some thing that really exists, not just a summary reference for several concrete observations in the world.

4. The belief that the concept itself is real results in irony. We now begin discussing and debating whether specific observations are “really” sufficient indicators of the concept.

Figure 5.1 Process of Conceptual Entrapment
is so vague as to render this characterization useless for the purposes of rigorous inquiry.

A “real” or “essential nature” definition is inherently subjective. The specification of concepts in scientific inquiry depends instead on conceptual and operational definitions. A **conceptual definition** is a working definition specifically assigned to a term. In the midst of disagreement and confusion over what a term really means, the scientist specifies a working definition for the purposes of the inquiry. Wishing to examine socioeconomic status (SES), for example, we may simply specify that we are going to treat it as a combination of income and educational attainment. With that definitional decision, we rule out many other possible aspects of SES: occupational status, money in the bank, property, lineage, lifestyle, and so forth.

The specification of conceptual definitions does two important things. First, it serves as a specific working definition we present so that readers will understand exactly what we mean by a concept. Second, it focuses our observational strategy. Notice that a conceptual definition does not directly produce observations; rather, it channels our efforts to develop actual measures.

As a next step, we must specify exactly what we will observe, how we will do it, and what interpretations we will place on various possible observations. These further specifications make up the **operational definition** of the concept—a definition that spells out precisely how the concept will be measured. Strictly speaking, an operational definition is a description of the “operations” undertaken in measuring a concept.

Pursuing the definition of SES, we might decide to ask the people we are studying three questions:

1. What was your total household income during the past 12 months?
2. How many persons are in your household?
3. What is the highest level of school you have completed?

Next, we need to specify a system for categorizing the answers people give us. For income, we might use the categories “under $25,000” and “$25,000–$35,000.” Educational attainment might be similarly grouped into categories, and we might simply count the number of people in each household. Finally, we need to specify a way to combine each person’s responses to these three questions to create a measure of SES.

The end result is a working and workable definition of SES. Others might disagree with our conceptualization and operationalization, but the definition has one essential scientific virtue: It is absolutely specific and unambiguous. Even if someone disagrees with our definition, that person will have a good idea of how to interpret our research results because what we mean by the term SES—reflected in our analyses and conclusions—is clear.

Here is a diagram showing the progression of measurement steps from our vague sense of what a term means to specific measurements in a scientific study:

```
Conceptualization
  ↓
Conceptual definition
  ↓
Operational definition
  ↓
Measurements in the real world
```

To test your understanding of these measurement steps, return to the beginning of the chapter, where we asked you what image comes to mind in connection with the word *recidivism*. Recall your own mental image, and compare it with Tony Fabelo’s (1995) discussion in the box titled “What Is Recidivism?”
Operationalization Choices

Describing how to obtain actual empirical measures begins with operationalization.

Recall from Chapter 4 that the research process is not usually a set of steps that proceed in order from first to last. This is especially true of operationalization, the process of developing operational definitions. Although we begin by conceptualizing what we wish to study, once we start to consider operationalization, we may revise our conceptual definition. Developing an operational definition also moves us closer to measurement, which requires that we think about selecting a data collection method as well. In other words, operationalization does not proceed according to a systematic checklist.

To illustrate this fluid process, let’s return to the issue of crime seriousness. Suppose we want to conduct a descriptive study that shows which crimes are more serious and which crimes are less serious.

One obvious dimension of crime seriousness is the penalties that are assigned to different crimes by law. Let’s begin with this conceptualization. Our conceptual definition of crime seriousness is therefore the level of punishment that a state criminal code authorizes for different crimes. Notice that this
definition has the distinct advantage of being unambiguous. We’re making progress, which leads us to an operational definition something like this:

Consult the Texas Criminal Code. (1) Those crimes that may be punished by death will be judged most serious. (2) Next will be crimes that may be punished by a prison sentence of more than 1 year. (3) The least serious crimes are those with jail sentences of less than a year and/or fines.

The operations undertaken to measure crime seriousness are specific. Our data collection strategy is also clear: Search the web or go to the library, make a list of crimes described in the Texas Code, and classify each crime into one of the three groups.

Note that we have produced rather narrow conceptual and operational definitions of crime seriousness. We might presume that penalties in the Texas Code take into account additional dimensions like victim harm, offender motivation, and other circumstances of individual crimes. However, the three groups of crimes include very different types of incidents and so do not tell us much about crime seriousness.

An alternative conceptualization of crime seriousness might center on what people

### Systemwide Recidivism Rates

Recidivism rates can be reported for all offenders in the system—for all offenders released from prison or for all offenders placed on probation. This I call systemwide recidivism rates. Approximately 48 percent of offenders released from prison on parole or mandatory supervision, or released from county jails on parole, in 1991 were reincarcerated by 1994 for a new offense or a parole violation.

For offenders released from prison in 1991 the reincarceration recidivism rate three years after release from prison by offense of conviction is listed below:

<table>
<thead>
<tr>
<th>Crime</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burglary</td>
<td>56%</td>
</tr>
<tr>
<td>Robbery</td>
<td>54%</td>
</tr>
<tr>
<td>Theft</td>
<td>52%</td>
</tr>
<tr>
<td>Drugs</td>
<td>43%</td>
</tr>
<tr>
<td>Assault</td>
<td>44%</td>
</tr>
<tr>
<td>Homicide</td>
<td>40%</td>
</tr>
<tr>
<td>Sexual assault</td>
<td>39%</td>
</tr>
<tr>
<td>Sex offense</td>
<td>34%</td>
</tr>
</tbody>
</table>

For the same group, the reincarceration recidivism rate three years after release by age group is listed below:

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>17–25</td>
<td>56%</td>
</tr>
<tr>
<td>26–30</td>
<td>52%</td>
</tr>
<tr>
<td>31–35</td>
<td>48%</td>
</tr>
<tr>
<td>36–40</td>
<td>46%</td>
</tr>
<tr>
<td>41 or older</td>
<td>35%</td>
</tr>
</tbody>
</table>

### The Meaning of Systemwide Recidivism Rates

The systemwide recidivism rate of prison releases should not be used to measure the performance of institutional programs. There are many socio-economic factors that can affect systemwide recidivism rates.

For example, the systemwide recidivism rate of offenders released from prison in 1995 declined because of changes in the characteristics of the population released from prison. Offenders are receiving and serving longer sentences, which will raise the average age at release. Therefore, “performance” in terms of systemwide recidivism will improve but not necessarily due to improvements in the delivery of services within the prison system.

On the other hand, the systemwide recidivism rate of felons released from state jail facilities should be expected to be relatively high, since state jail felons are property and drug offenders who tend to have high recidivism rates.
think of as serious crime. In this view, crime seriousness is based on people’s beliefs, which may reflect their perceptions of harm to victims, offender motivation, or other dimensions. Conceptualizing crime seriousness in this way suggests a different approach to operationalization: You will present descriptions of various crimes to other students in your class and ask them to indicate how serious they believe the crimes are. If crime seriousness is operationalized in this way, a questionnaire is the most appropriate data collection method.

We must make many other decisions before actually taking measurements of crime seriousness. How should crimes be described, and how should students indicate whether a crime is serious or not serious? Will crime descriptions simply be ranked from most to least serious, or will we try to make more precise measurements about how, for example, rape compares with smoking marijuana?

**Measurement as “Scoring”**

Operationalization involves describing how actual measurements will be made. The next step, of course, is making the measurements. Royce Singleton and associates (Singleton, Straits, and Straits, 2005:100) define *measurement* as “the process of assigning numbers or labels to units of analysis in order to represent conceptual properties. This process should be quite familiar to the reader even if the definition is not.”

Think of some examples of the process. Your instructor assigns number or letter grades to exams and papers to represent your mastery of course material. You count the number of pages in this week’s history assignment to represent how much time you will have to spend...
he got into trouble again. It did not take long. Two months after being released, Allan was arrested for robbery and returned to jail.

Now it gets complicated. A judge imposes the remaining 11 months of Allan’s suspended sentence. Allan is denied bail and must wait for his trial in jail. It is soon learned that Allan is wanted by police in Illinois for passing bad checks. Many people would be delighted to send Allan to Illinois; they tell officials in that state they can have him, pending resolution of the situation in Marion County. Allan is now in jail for three reasons: (1) serving his sentence for the original burglary, (2) awaiting trial on a robbery charge, and (3) waiting for transfer to Illinois.

Is this one jail stay or three? In a sense, it is one jail stay because one person, Allan, is occupying a jail cell. But let’s say Allan’s trial on the robbery charge is delayed until after he completes his sentence for the burglary. He stays in jail and begins a new jail stay. When he comes up for trial, the prosecutor asks to waive the robbery charges against Allan in hopes of exporting him to the neighboring state, and a new jail stay begins as Allan awaits his free trip to Illinois.

You may recognize this as a problem with units of analysis. Is the unit the person who stays in jail? Or are the separate reasons Allan is in jail—which are social artifacts—the units of analysis? After some thought, Maxfield decided that the social artifact was the more appropriate unit because he was interested in whether jail cells are more often occupied by people serving sentences or people awaiting trial. But that produced a new question of how to deal with people like Allan. Do we double-count the overlap in Allan’s jail stays, so that he accounts for two jail stays while serving his suspended sentence for burglary and waiting for the robbery trial? This seemed to make sense, but then Allan’s two jail stays would count the same as two other people with one jail stay each. In other words, Allan would appear to occupy two jail beds at the same time. This was neither true nor helpful in describing how long people stay in jail for different reasons.

studying. The American Bar Association rates nominees to the U.S. Supreme Court as qualified, highly qualified, or not qualified. You might rank last night’s date on the proverbial scale of 1 to 10, reflecting whatever conceptual properties are important to you.

Another way to think of measurement is in terms of scoring. Your instructor scores exams by counting the right answers and assigning some point value to each answer. Referees keep score at basketball games by counting the number of 1-point free throws and 2- and 3-point field goals for each team. Judges or juries score persons charged with crime by pronouncing “guilty” or “not guilty.” City murder rates are scored by counting the number of murder victims and dividing by the number of city residents.

Measurement is distinct from operationalization in that measurement involves actually making observations and assigning scores—numbers or other labels—to those observations. Making observations, of course, is related to the data collection method, which, in turn, is implied by operationalization. However, the measurement process begins much earlier, usually with conceptualization.

Many people consider measurement to be the most important and difficult phase of criminal justice research. It is difficult, in part, because so many basic concepts in criminal justice are not easy to define as specifically as we would like. Without being able to settle on a conceptual definition, we find operationalizing and measuring things challenging. This is illustrated by the box titled “Jail Stay.”

In addition to being challenging, different operationalization choices can produce different results. Here are three brief examples, developed with the excellent assistance of students.
First, in comparing a general-purpose survey of crime victimization with specialized studies of family conflict, Murray Straus (1999) reports that domestic assault rates are over 16 times higher in surveys of family conflict. He attributes this to an explicit focus on crime in crime surveys, whereas studies of family conflict do not cue respondents with the keyword crime. As a result, physical violence among family members appears to be much more prevalent in Straus’s research than in general crime surveys. The opposite is true of injuries—Straus reports that less than 4 percent of family violence incidents produced an injury, compared with 75 percent in the National Crime Victimization Survey (NCVS). These divergent findings make sense. Straus finds higher rates of less severe violence, whereas crime surveys reveal lower rates of more severe violence.

Second, according to the FBI Uniform Crime Report (UCR) about 7 million larcenies were reported in 2001 (Federal Bureau of Investigation, 2002), whereas the NCVS estimated over 14 million larcenies nationwide for the same year (Rennison, 2002). FBI data count only crimes reported to police, whereas the NCVS includes incidents not reported. Less than half of larcenies are reported, accounting for most of the difference. But for motor vehicle theft, the patterns are reversed: Over 1,226,000 vehicle thefts were recorded for the 2001 UCR, compared with fewer than 1,009,000 for the NCVS. A different operationalization rule is at work here: The NCVS measures crimes against households and cannot count thefts of vehicles owned by businesses or other organizations; these crimes do appear in UCR counts.

Third, in the box titled “What Is Recidivism?” Tony Fabelo argues that the at-risk period for comparing recidivism for different groups of offenders should be uniform. It’s possible to examine 1-, 2-, or 3-year rates, but comparisons should use standard at-risk periods. Varying the at-risk period produces, as we might expect, differences in recidivism rates. Evaluating a Texas program that provided drug abuse treatment, Michael Eisenberg (1999:8) reports rates for different at-risk periods:

<table>
<thead>
<tr>
<th></th>
<th>1-Year</th>
<th>2-Year</th>
<th>3-Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>All participants</td>
<td>14%</td>
<td>37%</td>
<td>42%</td>
</tr>
</tbody>
</table>

Note that the difference between 1- and 2-year rates is much larger than that between 2- and 3-year rates. Operationalizing “recidivism” as a 1-year failure rate would be much less accurate than operationalizing the concept as a 2-year rate, because recidivism rates seem to stabilize at the 2-year point.

**Exhaustive and Exclusive Measurement**

Briefly revisiting terms introduced in Chapter 1, an attribute is a characteristic or quality of something. “Female” is an example, as are “old” and “student.” Variables, in contrast, are logical sets of attributes. Thus, “gender” is a variable composed of the attributes “female” and “male.” The conceptualization and operationalization processes can be seen as the specification of variables and the attributes composing them. Thus, “employment status” is a variable that has the attributes “employed” and “unemployed,” or the list of attributes could be expanded to include other possibilities such as “employed part-time,” “employed full-time,” and “retired.”

Every variable should have two important qualities. First, the attributes composing it should be exhaustive. If the variable is to have any utility in research, researchers must be able to classify every observation in terms of one of the attributes composing the variable. We will run into trouble if we conceptualize the variable “sentence” in terms of the attributes “prison” and “fine.” After all, some convicted persons are assigned to probation, some have a portion of their prison sentence suspended, and others may receive a mix of prison terms, probation,
suspended sentences, or perhaps community service. Notice that we could make the list of attributes exhaustive by adding “other” and “combination.” Whatever approach we take, we must be able to classify every observation.

At the same time, attributes composing a variable must be mutually exclusive. Researchers must be able to classify every observation in terms of one and only one attribute. Thus, for example, we need to define “prison” and “fine” in such a way that nobody can be both at the same time. That means we must be able to handle the person whose sentence includes both a prison term and a fine. In this case, attributes could be defined more precisely by specifying “prison only,” “fine only,” and “both prison and fine.”

Levels of Measurement
Attributes composing any variable must be mutually exclusive and exhaustive. Attributes may be related in other ways as well. Of particular interest is that variables may represent different levels of measurement: nominal, ordinal, interval, and ratio. Levels of measurement tell us what sorts of information we can gain from the “scores” assigned to the values of a variable.

Nominal Measures Variables whose attributes have only the characteristics of exhaustiveness and mutual exclusiveness are nominal measures. Examples are gender, race, city of residence, college major, Social Security number, and marital status. Although the attributes composing each of these variables—“male” and “female” for the variable “gender”—are distinct from one another and more or less exhaust the possibilities among people, they have none of the additional structures mentioned later. Nominal measures merely offer names or labels for characteristics.

Imagine a group of people being characterized in terms of a nominal variable and physically grouped by the appropriate attributes. Suppose we are at a convention attended by hundreds of police chiefs. At a social function, we ask them to stand together in groups according to the states in which they live: all those from Vermont in one group, those from California in another, and so forth. (The variable is “state of residence”; the attributes are “live in Vermont,” “live in California,” and so on.) All the people standing in a given group have at least one thing in common; the people in any one group differ from the people in all other groups in that same regard. Where the individual groups are formed, how close they are to one another, and how they are arranged in the room is irrelevant. All that matters is that all the members of a given group share the same state of residence and that each group has a different shared state of residence.

Ordinal Measures Variables whose attributes may be logically rank ordered are ordinal measures. The different attributes represent relatively more or less of the variable. Examples of variables that can be ordered in some way are opinion of police, occupational status, crime seriousness, and fear of crime.

Continuing the example of police chiefs at a social gathering, imagine that we ask all those all those with a high school diploma (but who were not also college graduates) to stand in one group, those who have obtained bachelor degrees from college to stand in another group, and all those who have completed graduate or professional degrees to stand in a third group. This manner of grouping people satisfies the requirements for exhaustiveness and mutual exclusiveness. In addition, however, we might logically arrange the three groups in terms of their amount of formal education (the shared attribute). We might arrange the three groups in a row, ranging from least to most formal education. This arrangement provides a physical representation of an ordinal measure. If we know which groups two individuals are in, we can determine that one has more, less, or the same formal education as the other.

Note that in this example it is irrelevant how close or far apart the educational groups
are from one another. They might stand 5 feet apart or 500 feet apart; the college and high school groups could be 5 feet apart, and the less-than-high-school group might be 500 feet farther down the line. These actual distances have no meaning. The college group, however, should be between the high-school group and the graduate/professional degree group, or else the rank order is incorrect.

Interval Measures When the actual distance that separates the attributes composing some variables does have meaning, the variables are interval measures. The logical distance between attributes can then be expressed in meaningful standard intervals.

Interval measures commonly used in social scientific research are constructed measures such as standardized intelligence tests. The interval that separates IQ scores of 100 and 110 is the same as the interval that separates scores of 110 and 120 by virtue of the distribution of the observed scores of the many thousands of people who have taken the test over the years. Criminal justice researchers often combine individual nominal and ordinal measures to produce a composite interval measure. We’ll see examples at the end of this chapter.

Ratio Measures Most of the social scientific variables that meet the minimum requirements for interval measures also meet the requirements for ratio measures. In ratio measures, the attributes that compose a variable, besides having all the structural characteristics mentioned previously, are based on a true zero point. Examples from criminal justice research are age, dollar value of property loss from burglary, number of prior arrests, blood alcohol content, and length of incarceration.

Returning to the example of various ways to classify police chiefs, we might ask the chiefs to group themselves according to years of experience in their present position. All those new to their job would stand together, as would those with 1 year of experience, those with 2 years on the job, and so forth. The facts that members of each group share the same years of experience and that each group has a different shared length of time satisfy the minimum requirements for a nominal measure. Arranging the several groups in a line from those with the least to those with the most experience meets the additional requirements for an ordinal measure and permits us to determine whether one person is more experienced, is less experienced, or has the same level of experience as another. If we arrange the groups so that there is the same distance between each pair of adjacent groups, we satisfy the additional requirements of an interval measure and can say how much more experience one chief has than another. Finally, because one of the attributes included—experience—has a true zero point (police chiefs just appointed to their job), the phalanx of convention-goers also meets the requirements for a ratio measure, permitting us to say that one person is twice as experienced as another.

Implications of Levels of Measurement

To review this discussion and to understand why level of measurement may make a difference, consider Table 5.1. It presents information on crime seriousness adapted from a survey of crime severity conducted for the Bureau of Justice Statistics (Wolfgang, Figlio, Tracy, and Singer, 1985). The survey presented brief descriptions of more than 200 different crimes to a sample of 60,000 people. Respondents were asked to assign a score to each crime based on how serious they thought the crime was compared with bicycle theft (scored at 10).

The first column in Table 5.1 lists some of the crimes described. The second column shows a nominal measure that identifies the victim in the crime: home, person, business, or society. Type of victim is an attribute of each crime. The third column lists seriousness scores computed from survey results, ranging from 0.6 for trespassing to 35.7 for murder. These seriousness scores are interval measures because the distance between, for example, auto theft (at
8.0) and accepting a bribe (at 9.0) is the same as that between accepting a bribe (at 9.0) and obstructing justice (at 10.0). Seriousness scores are not ratio measures; there is no absolute zero point, and three instances of obstructing justice (at 10.0) do not equal one rape with injury (at 30.0).

The fourth column shows the ranking for each of the 17 crimes in the table; the most serious crime, murder, is ranked 1, followed by rape with injury, and so on. The rankings express only the order of seriousness, however, because the difference between murder (ranked 1) and rape (ranked 2) is smaller than the distance between rape and robbery with injury (ranked 3).

Finally, the crime descriptions presented to respondents indicated the value of property loss for each offense. This is a ratio measure with a true zero point, so that 10 burglaries with a loss of $1000 each have the same property value as one arson offense with a loss of $10,000.

Specific analytic techniques require variables that meet certain minimum levels of measurement. For example, we could compute the average property loss from the crimes listed in Table 5.1 by adding up the individual numbers in the fifth column and dividing by the number of crimes listed (17). However, we would not be able to compute the average victim type because that is a nominal variable. In that case, we could report the modal—the most common—victim type, which is society in Table 5.1.

You may treat some variables as representing different levels of measurement. Ratio measures are the highest level, followed by interval, ordinal, and nominal. A variable that represents a given level of measurement—say, ratio—may also be treated as representing a lower level of measurement—say, ordinal. For example, age is a ratio measure. If we wish to examine only the relationship between age and some ordinal level variable, such as delinquency involvement (high, medium, or low), we

<table>
<thead>
<tr>
<th>Crime</th>
<th>Victim</th>
<th>Seriousness Score</th>
<th>Rank</th>
<th>Value of Property Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accepting a bribe</td>
<td>Society</td>
<td>9.0</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Arson</td>
<td>Business</td>
<td>12.7</td>
<td>6</td>
<td>$10,000</td>
</tr>
<tr>
<td>Auto theft</td>
<td>Home</td>
<td>8.0</td>
<td>10</td>
<td>$12,000</td>
</tr>
<tr>
<td>Burglary</td>
<td>Business</td>
<td>15.5</td>
<td>5</td>
<td>$100,000</td>
</tr>
<tr>
<td>Burglary</td>
<td>Home</td>
<td>9.6</td>
<td>8</td>
<td>$1,000</td>
</tr>
<tr>
<td>Buying stolen property</td>
<td>Society</td>
<td>5.0</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>Heroin sales</td>
<td>Society</td>
<td>20.6</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Heroin use</td>
<td>Society</td>
<td>6.5</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>Murder</td>
<td>Person</td>
<td>35.7</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Obstructing justice</td>
<td>Society</td>
<td>10.0</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Public intoxication</td>
<td>Society</td>
<td>0.8</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Rape and injury</td>
<td>Person</td>
<td>30.0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Robbery and injury</td>
<td>Person</td>
<td>21.0</td>
<td>3</td>
<td>$1,000</td>
</tr>
<tr>
<td>Robbery attempt</td>
<td>Person</td>
<td>3.3</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>Robbery, no injury</td>
<td>Person</td>
<td>8.0</td>
<td>10</td>
<td>$1,000</td>
</tr>
<tr>
<td>Shoplifting</td>
<td>Business</td>
<td>2.2</td>
<td>14</td>
<td>$10</td>
</tr>
<tr>
<td>Trespassing</td>
<td>Home</td>
<td>0.6</td>
<td>16</td>
<td>0</td>
</tr>
</tbody>
</table>

Source: Adapted from Wolfgang, Figlio, Tracy, and Singer (1985).
might choose to treat age as an ordinal-level variable as well. We might characterize the subjects of our study as being young, middle-aged, and old, specifying the age range for each of those groupings. Finally, age might be used as a nominal-level variable for certain research purposes. Thus, people might be grouped together as baby boomers if they were born between 1945 and 1955.

The analytic uses planned for a given variable, then, should determine the level of measurement to be sought, with the realization that some variables are inherently limited to a certain level. If a variable is to be used in a variety of ways that require different levels of measurement, the study should be designed to achieve the highest level possible. Although ratio measures such as number of arrests can later be reduced to ordinal or nominal ones, it is not possible to convert a nominal or ordinal measure to a ratio one. More generally, you cannot convert a lower-level measure to a higher-level one. That is a one-way street worth remembering.

Criteria for Measurement Quality

The key standards for measurement quality are reliability and validity.

Measurements can be made with varying degrees of precision, which refers to the fineness of the distinctions made between the attributes that compose a variable. Saying that a woman is “43 years old” is more precise than that she is “in her forties.” Describing a felony sentence as “18 months” is more precise than “over 1 year.” As a general rule, precise measurements are superior to imprecise ones, as common sense would suggest. Precision is not always necessary or desirable, however. If knowing that a felony sentence is over 1 year is sufficient for your research purpose, then any additional effort invested in learning the precise sentence would be wasted. The operationalization of concepts, then, must be guided partly by an understanding of the degree of precision required. If your needs are not clear, be more precise rather than less.

Don’t confuse precision with accuracy. Describing someone as “born in Stowe, Vermont” is more precise than “born in New England,” but suppose the person in question was actually born in Boston? The less precise description, in this instance, is more accurate; it’s a better reflection of the real world. This is a point worth keeping in mind. Many criminal justice measures are imprecise, so reporting approximate values is often preferable.

Precision and accuracy are obviously important qualities in research measurement, and they probably need no further explanation. When criminal justice researchers construct and evaluate measurements, they pay special attention to two technical considerations: reliability and validity.

Reliability

In the abstract, reliability is a matter of whether a particular measurement technique, applied repeatedly to the same thing, will yield the same result each time. In other words, measurement reliability is roughly the same as measurement consistency or stability. Imagine, for example, a police officer standing on the street, guessing the speed of cars that pass by, and issuing speeding tickets based on that judgment. If you received a ticket from this officer and went to court to contest it, you would almost certainly win your case. The judge would no doubt reject this way of measuring speed, regardless of the police officer’s experience. The reliability or consistency of this method of measuring vehicle speed is questionable at best. If the same police officer used a radar speed detector, however, it is doubtful that you would be able to beat the ticket. The radar device is judged a much more reliable way of measuring speed.

Reliability, though, does not ensure accuracy any more than precision does. The speedometer in your car may be a reliable instru-
ment for measuring speed, but it is common for speedometers to be off by a few miles per hour, especially at higher speeds. If your speedometer shows 55 miles per hour when you are actually traveling at 60, it gives you a consistent but inaccurate reading that might attract the attention of police officers with more accurate radar guns.

Measurement reliability is often a problem with indicators used in criminal justice research. Numerous studies have shown that measures of crime based on police records often suffer from reliability problems. A classic example is the study by Richard Mc Cleary and associates (Mc Cleary, Nienstedt, and Erven, 1982:362), which analyzed changes in police records of burglary following a change in how burglary reports were investigated in a large city. Under the new system, detectives formally investigated burglaries that previously had been examined only by patrol officers. Burglaries declined sharply as soon as the new investigation procedures were implemented. The reason for the decline was that some patrol officers counted some crimes as burglaries that did not meet the official definition of burglary. When a smaller number of detectives began to investigate burglaries, they were more consistent in applying the official definition.

Other examples of reliability problems can be found in criminal justice research and policy settings. How accurate are the evaluations of judges in determining which defendants are at risk of fleeing before trial? A study by Sheila Royo Maxwell (1999) found that the factors used by judges in deciding whether to release defendants on their own recognizance were neither reliable nor effective predictors of the likelihood that defendants would fail to appear in court. Inconsistency in the administration of blood alcohol tests in certain states has forced researchers to search for new measures of drunk driving (Heeren, Smith, Morelock, and Hingson, 1985). Forensic DNA evidence is increasingly being used in violent crime cases. A National Research Council (1996) study found a variety of errors in laboratory procedures—including sample mishandling, evidence contamination, and analyst bias. These are measurement reliability problems that can lead to unwarranted exclusion of evidence or to the conviction of innocent people. For example, irregularities in DNA tests by a Texas crime lab led to the exoneration of at least one previously convicted defendant and prompted reviews of hundreds of additional cases (McVicker and Khanna, 2003).

Reliability problems crop up in many forms. Reliability is a concern every time a single observer is the source of data because we have no way to guard against that observer’s subjectivity. We can’t tell for sure how much of what’s reported represents true variation and how much is due to the observer’s unique perceptions.

Reliability can also be an issue when more than one observer makes measurements. Survey researchers have long known that different interviewers get different answers from respondents as a result of their own attitudes and demeanor. Suppose we want to conduct a study of editorial opinions about work release centers. We could create a team of coders to read hundreds of newspaper editorials and classify them in terms of their position on the issue, but different coders might code the same editorial differently. Or we may want to classify a few hundred community anticrime groups in terms of some standard coding scheme, such as a set of categories created by the National Institute of Justice. However, a police officer and a neighborhood activist are unlikely to code all those groups into the same categories.

These examples illustrate problems of reliability. Similar problems arise when we ask people for information about themselves. Sometimes we ask questions that people don’t know the answers to (“How many times have you seen a police officer in the last month?”). Sometimes we ask people about things that are totally irrelevant to them (“Are you satisfied with the FBI’s guidelines on the purchase of office supplies?”). And sometimes we ask
questions that are so complicated that a person who has a clear opinion on the matter might have a different interpretation of the question when asked a second time.

How do we create reliable measures? Because the problem of reliability is a basic one in criminal justice measurement, researchers have developed a number of techniques for dealing with it.

**The Test–Retest Method** Sometimes it is appropriate to make the same measurement more than once. If there is no reason to expect the information to change, we should expect the same response both times. If answers vary, however, then the measurement method is, to the extent of that variation, unreliable. Here’s an illustration.

In their classic research on delinquency in England, Donald West and David Farrington (1977) interviewed a sample of 411 males from a working-class area of London at age 16 and again at age 18. The subjects were asked to describe a variety of aspects of their lives, including educational and work history, leisure pursuits, drinking and smoking habits, delinquent activities, and experience with police and courts.

Because many of these topics involve illegal or at least antisocial activity, West and Farrington were concerned about the accuracy of information obtained in their interviews. They assessed reliability in several ways. One was to compare responses from the interview at age 18 with those from the interview at age 16. For example, in each interview, the youths were asked at what age they left school. In most cases, there were few discrepancies in stated age from one interview to the next, which led the authors to conclude: “There was therefore no systematic tendency for youths either to increase or lessen their claimed period of school attendance as they grew older, as might have occurred if they had wanted either to exaggerate or to underplay their educational attainments” (1977:76–77). If West and Farrington had found less consistency in answers to this and other items, they would have had good reason to doubt the truthfulness of responses to more sensitive questions. The test–retest method suggested to the authors that memory lapses were the most common source of minor differences.

Although this method can be a useful reliability check, it is limited in some respects. Faulty memory may produce inconsistent responses if there is a lengthy gap between the initial interview and the retest. A different problem can arise in trying to use the test–retest method to check the reliability of attitude or opinion measures. If the test–retest interval is short, then answers given in the second interview may be affected by earlier responses if subjects try to be consistent.

**Interrater Reliability** It is also possible for measurement unreliability to be generated by research workers—for example, interviewers and coders. To guard against interviewer unreliability, it is common practice in surveys to have a supervisor call a subsample of the respondents on the telephone and verify selected information. West and Farrington (1977:173) checked interrater reliability in their study of London youths and found few significant differences in results obtained from different interviewers.

Comparing measurements from different raters works in other situations as well. For example, Michael Geerken (1994) presents an important discussion of reliability problems that researchers are likely to encounter in measuring prior arrests through police “rap sheets.” Duplicate entries, the use of aliases, and the need to transform official crime categories into a smaller number of categories for analysis are among the problems that Geerken cites. One way to increase consistency in translating official records into research measures, a process often referred to as coding, is to have more than one person code a sample of records and then compare the consistency of coding decisions made by each person. This approach was used by Michael Maxfield and Cathy Spatz Widom...
In their analysis of adult arrests of child abuse victims.

In general, whenever researchers are concerned that measures obtained through coding may not be classified reliably, they should have each independently coded by different people. In the hypothetical study of newspaper editorials about a proposed work release center, coding decisions that generate disagreement should be evaluated more carefully and resolved. If we find a great deal of disagreement, our operational definitions of how to code newspaper editorials should be carefully reviewed and made more specific.

**Split-Half Method**  As a general rule, it is always a good idea to make more than one measurement of any subtle or complex social concept, such as prejudice or fear of crime. This procedure lays the groundwork for another check on reliability. Suppose you’ve created a questionnaire that contains 10 items you believe measure prejudicial beliefs about African Americans and delinquency. Using the split-half technique, you randomly assign those 10 items to two sets of five items. Each set should provide a good measure of prejudice, and the sets should agree in the way they classify the respondents. If the two sets of items measure people differently, then that, again, points to a problem in the reliability of how you are measuring the variable.

In their longitudinal study of links between childhood experiences and adolescent delinquency, Woodward and Fergusson (2000) used interview data that included something like an abbreviated IQ test. IQ tests are usually quite lengthy, reflecting the need to include a large number of test items to measure multiple dimensions of intelligence. One consequence is that it’s not possible to use standard IQ tests in a survey—split-half. Then scores on each subset of items are computed, and compared across the split-halves. In their study, Woodward and Fergusson report a split-half reliability of value .93, roughly indicating that 93 percent of scores between the two halves were consistent.

The reliability of measurements is a fundamental issue in criminal justice research, and we’ll return to it in the chapters to come. For now, however, we hasten to point out that even total reliability doesn’t ensure that our measures actually measure what we think they measure. That brings us to the issue of validity.

**Validity**

In conventional usage, the term **validity** means that an empirical measure adequately reflects the meaning of the concept under consideration. Put another way, measurement validity involves whether you are really measuring what you say you are measuring. Recall that an operational definition specifies the operations you will perform to measure a concept. Does your operational definition accurately reflect the concept you are interested in? If the answer is yes, you have a valid measure. A radar gun is a valid measure of vehicle speed, but a wind velocity indicator is not because the volume of air displaced by a slow-moving truck will register higher than that displaced by a fast-moving sports car.

Although methods for assessing reliability are relatively straightforward, it is more difficult to demonstrate that individual measures are valid. Because concepts are not real, but abstract, we cannot directly demonstrate that measures, which are real, are actually measuring an abstract concept. Nevertheless, researchers have some ways of dealing with the issue of validity.

**Face Validity**  First, there’s something called **face validity.** Particular empirical measures may or may not jibe with our common agreements and our individual mental images about a particular concept. We might debate the adequacy of measuring satisfaction with police services by counting the number of citizen
complaints registered by the mayor’s office, but we’d surely agree that the number of citizen complaints has something to do with levels of satisfaction. If someone suggested that we measure satisfaction with police by finding out whether people like to watch police dramas on TV, we would probably agree that the measure has no face validity; it simply does not make sense.

Second, there are many concrete agreements among researchers about how to measure certain basic concepts. The Census Bureau, for example, has created operational definitions of such concepts as family, household, and employment status that seem to have a workable validity in most studies using those concepts.

**Criterion-Related Validity**

Criterion-related validity involves comparing a measure with some external criterion. A measure can be validated by showing that it predicts scores on another measure that is generally accepted as valid; this is sometimes referred to as convergent validity. The validity of College Board exams, for example, is shown in their ability to predict the success of students in college.

Timothy Heeren and associates (Heeren, Smith, Morelock, and Hingson, 1985) offer a good example of criterion-related validity in their efforts to validate a measure of alcohol-related auto fatalities. Of course, conducting a blood alcohol laboratory test on everyone killed in auto accidents would be a valid measure. Not all states regularly do this, however, so Heeren and colleagues tested the validity of an alternate measure: single-vehicle fatal accidents involving male drivers occurring between 8:00 P.M. and 3:00 A.M. The validity of this measure was shown by comparing it with the blood alcohol test results for all drivers killed in states that reliably conducted such tests in fatal accidents. Because the two measures agreed closely, Heeren and associates claimed that the proxy or surrogate measure would be valid in other states.

Another approach to criterion-related validity is to show that our measure of a concept is different from measures of similar but distinct concepts. This is called discriminant validity, meaning that measures can discriminate different concepts. For example, Wesley Skogan and Michael Maxfield (1981:56–57) described how “crime” may sometimes be considered a code word for “race” among white Americans. Measures of fear of crime for some people may therefore be contaminated by racist fears about African Americans and crime. They established the discriminant validity of their measure by showing that it was not related to an index of racial intolerance.

Sometimes, it is difficult to find behavioral criteria that can be used to validate measures as directly as in the examples described here. In those instances, however, we can often approximate such criteria by considering how the variable in question ought, theoretically, to relate to other variables.

**Construct Validity**

Construct validity is based on the logical relationships among variables. Let’s suppose, for example, that we are interested in studying fear of crime—its sources and consequences. As part of our research, we develop a measure of fear of crime, and we want to assess its validity.

In addition to our measure, we will also develop certain theoretical expectations about the way the variable “fear of crime” relates to other variables. For instance, it’s reasonable to conclude that people who are afraid of crime are less likely to leave their homes at night for entertainment than people who are not afraid of crime. If our measure of fear of crime relates to how often people go out at night in the expected fashion, that constitutes evidence of our measure’s construct validity. However, if people who are afraid of crime are just as likely to go out at night as people who are not afraid, that challenges the validity of our measure.

Tests of construct validity, then, can offer a weight of evidence that our measure either does or doesn’t tap the quality we want it to measure, without providing definitive proof. We have
suggested here that tests of construct validity are less compelling than tests of criterion validity. However, there is room for disagreement about which kind of test should be done in a given situation. It is less important to distinguish these two types than to understand the logic of validation that they have in common: If we are successful in measuring some variable, then those measurements should relate to other measures in some logical fashion.

**Content Validity** Finally, **content validity** refers to the degree to which a measure covers the range of meanings included within the concept. For instance, this question has frequently been used in surveys to measure fear of crime:

How safe do you feel (or would you feel) walking alone in this area after dark? Would you say very safe, fairly safe, fairly unsafe, or very unsafe?

Although this question may provide a valid measure of fear in connection with street crime, it is not a good measure of fear of burglary, auto theft, or airplane hijacking. The concept of fear of crime is broader than the concept represented in the question.

**Multiple Measures** Another approach to validation of an individual measure is to compare it with alternative measures of the same concept. The use of multiple measures is similar to establishing criterion validity. However, the use of multiple measures does not necessarily assume that the criterion measure is always more accurate. For example, many crimes that are committed never result in an arrest, so arrests are not good measures of how many crimes are committed by individuals. Self-report surveys have often been used to measure delinquency and criminality. But how valid are survey questions that ask people how many crimes they have committed?

The approach used by West and Farrington (and by others) is to ask, for example, how many times someone has committed robbery and how many times someone has been arrested for that crime. Those who admit having been arrested for robbery are asked when and where the arrest occurred. Self-reports can then be validated by checking police arrest records. This works two ways: (1) It is possible to validate individual reports of being arrested for robbery, and (2) researchers can check police records for all persons interviewed to see if there are any records of robbery arrests that subjects do not disclose to interviewers.

Figure 5.2 illustrates the difference between validity and reliability. Think of measurement as analogous to hitting the bull's-eye on a target. A reliable measure produces a “tight pattern,” regardless of where it hits, because reliability is a function of consistency. Validity,
in contrast, relates to the arrangement of shots around the bull’s-eye. The failure of reliability in the figure can be seen as a random error; the failure of validity is a systematic error. Notice that neither an unreliable nor an invalid measure is likely to be very useful.

This chapter’s installment of our running example discusses reliability and validity of different approaches to measuring the race of drivers.

**Composite Measures**

*Combining individual measures often produces more valid and reliable indicators.*

Sometimes it is possible to construct a single measure that captures the variable of interest. For example, asking auto owners whether their car has been stolen in the previous 6 months is a straightforward way to measure auto theft victimization. But other variables may be bet-
Composite measures are frequently used in criminal justice research for three reasons. First, despite carefully designing studies to provide valid and reliable measurements of variables, the researcher is often unable to develop single indicators of complex concepts. That is especially true with regard to attitudes and opinions that are measured through surveys. For example, we saw that measuring fear of crime through a question that asks about feelings of

<table>
<thead>
<tr>
<th></th>
<th>Tollbooth Survey</th>
<th>Police stop data</th>
<th>Photo coding, speeding only</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>66%</td>
<td>52%</td>
<td>58%</td>
</tr>
<tr>
<td>Black</td>
<td>15</td>
<td>29</td>
<td>26</td>
</tr>
<tr>
<td>Hispanic</td>
<td>11</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Other</td>
<td>9</td>
<td>10</td>
<td>9</td>
</tr>
</tbody>
</table>

The last column presents race breakdown of drivers for cars that were further classified as speeding, in this case exceeding the posted speed limit by 15 or more miles per hour. We will have more to say about how speed was measured in the next chapter.

**Reliability and Validity**

Interrater reliability is well-suited to coding race by observation with multiple observers. In Pennsylvania, Engel et al. required that the two observers had to agree on race classification of drivers. If they did not agree, the driver’s race was recorded as unknown. But this occurred in only 2.7 percent of over 160,000 drivers observed in the Pennsylvania study.

In one sense, reliability in New Jersey was enhanced by having digital photos; they could be reviewed at length. However, taking photos with a stationary camera through the windshield of a car traveling at high speed presented technical challenges. The quality of many images was poor; 24 percent of the 38,747 images were judged to be unusable. Three research assistants independently viewed the remaining images on high-resolution computer monitors. It was required that two of three coders agree on the race of driver. This criterion was met for 89 percent of images that remained after discarding those that were unusable.

We have already touched on the question of validity. If we are interested in whether police decisions to stop drivers are based on a driver’s race, then a measure of the appearance of race is more valid than self-identified race. This is not true for measuring Hispanic ethnicity.

---

1Beginning in 2004, the FBI crime index was discontinued. The reasons underlying this change are a good illustration of what affects the quality of an index. See FBI (2005:5–6) for more information.
safety on neighborhood streets measures some dimensions of fear but certainly not all of them. This leads us to question the validity of using that single question to measure fear of crime.

Second, we may wish to use a rather refined ordinal measure of a variable, arranging cases in several ordinal categories from very low to very high on a variable such as degree of parental supervision. A single data item might not have enough categories to provide the desired range of variation, but an index or scale formed from several items would.

Finally, indexes and scales are efficient devices for data analysis. If a single data item gives only a rough indication of a given variable, considering several data items may give us a more comprehensive and more accurate indication. For example, the results of a single drug test would give us some indication of drug use by a probationer. Examining results from several drug tests would give us a better indication, but the manipulation of several data items simultaneously can be very complicated. In contrast, composite measures are efficient data reduction devices. Several indicators may be summarized in a single numerical score, even while perhaps very nearly maintaining the specific details of all the individual indicators.

**Typologies**

Researchers combine variables in different ways to produce different composite measures. The simplest of these is a *typology*, sometimes called a “taxonomy.” Typologies are produced by the intersection of two or more variables to create a set of categories or types. We may, for example, wish to classify people according to the range of their experience in criminal court. Assume we have asked a sample of people whether they have ever served as a juror and whether they have ever testified as a witness in criminal court. Table 5.2 shows how the yes/no responses to these two questions can be combined into a typology of experience in court.

Typologies can be more complex—combining scores on three or more measures or combining scores on two measures that take many different values. For an example of a complex typology, consider research by Rolf Loeber and associates (Loeber, Stouthamer-Loeber, von Kammen, and Farrington, 1991) on patterns of delinquency over time. The researchers used a longitudinal design in which a sample of boys was selected from Pittsburgh public schools and interviewed many times. Some questions asked about their involvement in delinquency and criminal offending. This approach made it possible to distinguish boys who reported different types of offending at different times.

Loeber and associates first classified delinquent and criminal acts into these ordinal seriousness categories (1991:44):

- **None**: No self-reported delinquency
- **Minor**: Theft of items worth less than $5, vandalism, fare evasion
- **Moderate**: Theft over $5, gang fighting, carrying weapons
- **Serious**: Car theft, breaking and entering, forced sex, selling drugs

Next, to measure changes in delinquency over time, the researchers compared reports of delinquency from the first screening interview with reports from follow-up interviews. These two measures—delinquency at time 1 and delinquency at time 2—formed the typology, which they referred to as a “dynamic classification of offenders” (1991:44). Table 5.3 summarizes this typology.

<table>
<thead>
<tr>
<th>Serve on Jury?</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testify as Witness?</td>
<td>No</td>
<td>A</td>
</tr>
<tr>
<td>Yes</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

**Table 5.2** Typology of Court Experience

Typology:
- A: No experience with court
- B: Experience as juror only
- C: Experience as witness only
- D: Experience as juror and witness

<table>
<thead>
<tr>
<th>Serve on Jury?</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testify as Witness?</td>
<td>No</td>
<td>A</td>
</tr>
<tr>
<td>Yes</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>
The first category in the table, “nondelinquent,” includes those boys who reported committing no offenses at both the screening and follow-up interviews. “Starters” reported no offenses at screening and then minor, moderate, or serious delinquency at follow-up, whereas “desistors” were just the opposite. Those who committed the same types of offenses at both times were labeled “stable”; “de-escalators” reported committing less serious offenses at follow-up; and “escalators” moved on to more serious offenses.

Table 5.3 Typology of Change in Juvenile Offending

<table>
<thead>
<tr>
<th>Typology</th>
<th>Screening (Time 1)</th>
<th>Follow-Up (Time 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Nondelinquent</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B. Starter</td>
<td>0</td>
<td>1, 2, or 3</td>
</tr>
<tr>
<td>C. Desistor</td>
<td>1, 2, or 3</td>
<td>0</td>
</tr>
<tr>
<td>D. Stable</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D. Stable</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>D. Stable</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>E. Deescalator</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>E. Deescalator</td>
<td>2 or 3</td>
<td>1</td>
</tr>
<tr>
<td>F. Escalator</td>
<td>1</td>
<td>2 or 3</td>
</tr>
<tr>
<td>F. Escalator</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Juvenile Offending Typology
0: None
1: Minor
2: Moderate
3: Serious

Source: Adapted from Loeber and associates (1991:43–46).

The first category in the table, “nondelinquent,” includes those boys who reported committing no offenses at both the screening and follow-up interviews. “Starters” reported no offenses at screening and then minor, moderate, or serious delinquency at follow-up, whereas “desistors” were just the opposite. Those who committed the same types of offenses at both times were labeled “stable”; “de-escalators” reported committing less serious offenses at follow-up; and “escalators” moved on to more serious offenses.

Notice the efficiency of this typology. Two variables (delinquency at screening and follow-up) with four categories each are reduced to a single variable with six categories. Furthermore, the two measures of delinquency are themselves composite measures, produced by summarizing self-reports of a large number of individual offenses. Finally, notice also how this efficiency is reflected in the clear meaning of the new composite measure. This dynamic typology summarizes information about time, offending, and offense seriousness in a single measure.

An Index of Disorder
“What is disorder, and what isn’t?” asks Wesley Skogan (1990:4) in his book on the links between crime, fear, and social problems such as public drinking, drug use, litter, prostitution, panhandling, dilapidated buildings, and groups of boisterous youths. In an influential article titled “Broken Windows,” James Wilson and George Kelling (1982) describe disorder as a sign of crime that may contribute independently to fear and crime itself. The argument goes something like this: Disorder is a symbol of urban decay that people associate with crime. Signs of disorder can produce two related problems. First, disorder may contribute to fear of crime, as urban residents believe that physical decay and “undesirables” are symbols of crime. Second, potential offenders may interpret evidence of disorder as a signal that informal social control mechanisms in a neighborhood have broken down and that the area is fair game for mayhem and predation.

We all have some sort of mental image (conception) of disorder, but to paraphrase Skogan’s question: How do we measure it? Let’s begin by distinguishing two conceptions of disorder. First, we can focus on the physical presence of disorder—whether litter, public drinking, public drug use, and the like are actually evident in an urban neighborhood. We might measure the physical presence of disorder through a series of systematic observations. This is the approach used by Robert Sampson and Stephen Raudenbush (1999) in their study of links between disorder and crime in Chicago. Unfortunately, these authors observed very few examples of disorder and altogether ignored the question of whether such behaviors were perceived as problematic by residents of Chicago neighborhoods.

That brings us to the second conception, one focusing on the perception of disorder. Thus, some people might view public drinking as disorderly, whereas others (New Orleans...
residents, for example) consider public drinking to be perfectly acceptable. Questionnaires and survey methods are the best suited for measuring perceived disorder.

Having settled on perceptions of disorder that we will measure through a survey, we must make some more decisions. Consider two versions of a question asking about people loitering on the street. The first is from a series of surveys conducted in three U.S. cities; the second is from a nationwide survey conducted in England and Wales in 1984. Each question is paraphrased from the original questionnaire.

1. Are groups of loiterers hanging out on the streets a big problem, some problem, or almost no problem in your neighborhood? (See Skogan and Maxfield, 1981)
2. In your area, how common are loiterers hanging around on the street: very common, fairly common, not very common, or not at all common? (See Maxfield, 1987)

The first question asks about perceptions of loiterers hanging out as a problem; the second question asks about perceptions of the frequency of people hanging out. Notice also that the first question requires two rather different things of respondents: They must perceive loiterers hanging around, and they must judge that to be a problem. Skogan used the first formulation of the question, reasoning that perception of behavior as a problem is required for it to be perceived as disorder.

Is this a good measure of disorder? The belief that loiterers hanging around is a problem surely has something to do with perceptions of disorder, but there is more to it than that. As it stands, we have a measure of one dimension of disorder, but our measure is quite narrow. In order to represent the concept of disorder more completely, we should measure additional behaviors or characteristics that represent other examples of disorder.

Skogan used questions about nine different examples of disorder and classified them into two groups representing what he calls social and physical disorder (Skogan, 1990:51, 191):

<table>
<thead>
<tr>
<th>Social Disorder</th>
<th>Physical Disorder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups of loiterers</td>
<td>Abandoned buildings</td>
</tr>
<tr>
<td>Drug use and sales</td>
<td>Garbage and litter</td>
</tr>
<tr>
<td>Vandalism</td>
<td>Junk in vacant lots</td>
</tr>
<tr>
<td>Gang activity</td>
<td>Public drinking</td>
</tr>
<tr>
<td>Street harassment</td>
<td></td>
</tr>
</tbody>
</table>

Questions corresponding to each of these examples of disorder asked respondents to rate them as big problems (scored 2), some problem (scored 1), or almost no problem (scored 0) in their neighborhood. Together, these nine items measure different types of disorder and appear to have reasonable face validity. However, examining the relationship between each individual item and respondents’ fear of crime or experience as a crime victim would be unwieldy at best. So Skogan created two indexes, one for social disorder and one for physical disorder, by adding up the scores for each item and dividing by the number of items in each group. Figure 5.3 shows a hypothetical sample questionnaire for these nine items, together with the scores that would be produced for each index.

This example illustrates how several related variables can be combined to produce an index that has three desirable properties. First, a composite index is a more valid measure of disorder than is a single question. Second, computing and averaging across all items in a category create more variation in the index than we could obtain in any single item. Finally, two indexes are more parsimonious than nine individual variables; data analysis and interpretation can be more efficient.

**Measurement Summary**

We have covered substantial ground in this chapter but still have introduced only the important and often complex issue of measurement in criminal justice research. The box “Measuring Race” presents an illustration from our running example. More than a step in the research process, measurement involves continuous thinking about what concep-
Introduction:

Now I’m going to read you a list of crime-related problems that may be found in some parts of the city. For each one, please tell me how much of a problem it is in your neighborhood. Is it a big problem, some problem, or almost no problem?

<table>
<thead>
<tr>
<th>Problem</th>
<th>Big problem</th>
<th>Some problem</th>
<th>No problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups of people loitering</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>People using or selling drugs</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Abandoned buildings</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Vandalism</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Garbage and litter on street</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>People drinking in public</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Junk in vacant lots</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>People making rude or insulting remarks</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Social = 2 + 1 + 1 + 0 + 2 + 1 = 7</td>
<td>Index score = $\frac{7}{6} = 1.16$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical = 0 + 2 + 0 = 2</td>
<td>Index score = $\frac{2}{3} = 0.67$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.3  Index of Disorder

- Higher levels of measurements specify categories that have ranked order or more complex numerical properties.
- A given variable can sometimes be measured at different levels of measurement. The most appropriate level of measurement used depends on the purpose of the measurement.
- Precision refers to the exactness of the measure used in an observation or description of an attribute.
- Reliability and validity are criteria for measurement quality. Valid measures are truly indicators of underlying concepts. A reliable measure is consistent.
- The creation of specific, reliable measures often seems to diminish the richness of meaning our general concepts have. A good solution is to use multiple measures, each of which taps different aspects of the concept.
- Composite measures, formed by combining two or more variables, are often more valid measures of complex criminal justice concepts.

**Main Points**

- Concepts are mental images we use as summary devices for bringing together observations and experiences that seem to have something in common.
- Our concepts do not exist in the real world, so they can’t be measured directly.
- In operationalization, we specify concrete empirical procedures that will result in measurements of variables.
- Operationalization begins in study design and continues throughout the research project, including the analysis of data.
- Categories in a measure must be mutually exclusive and exhaustive.

**Key Terms**

- Concept
- Conception
- Conceptual definition
- Conceptualization
- Construct validity
- Content validity
- Criterion-related validity
- Dimension
- Face validity
- Interval measures
Nominal measures  Reification
Operational definition  Reliability
Ordinal measures  Validity
Ratio measures

✪ Review Questions and Exercises

1. Review the box titled “What Is Recidivism?” From that discussion, write conceptual and operational definitions for recidivism. Summarize how Fabelo proposes to measure the concept. Finally, discuss possible reliability and validity issues associated with Fabelo’s proposed measure.

2. Each year publications such as the Places Rated Almanac attract praise from residents of highly rated cities and protests from those who live in lower-rated areas. The 1997 edition rated Orange County, California (where Babbie lives), first and Newark, New Jersey (where Maxfield works), last. Find a recent edition of this or a similar rating guide, and discuss the measurement procedures used for developing the ratings. If appropriate, criticize how various dimensions of the ratings have been operationalized. Can you spot and explain any apparent anomalies in the ratings? For example, the 1997 edition rated Gary, Indiana, higher than Santa Fe, New Mexico.

3. We all have some sort of mental image of the pace of life. In a fascinating book titled A Geography of Time, Robert Levine (1997) operationalized the pace of life in cities around the world with a composite measure of the following:
   a. How long it took a single pedestrian to walk 60 feet on an uncrowded sidewalk.
   b. What percentage of public clocks displayed the correct time.
   c. How long it took to purchase the equivalent of a first-class postage stamp with the equivalent of a $5 bill.

Discuss possible reliability and validity issues with these indicators of pace of life. Be sure to specify a conceptual definition for pace of life.

✪ Online Sessions for Chapter 5

Online Sessions present further examples of selected material covered in each chapter. Your instructor may choose to use these for in-class discussion (as I do with my students) or you can access them at home. To ensure that you always have access to live, correct links for the websites described here, we provide the necessary links on the companion website for Research Methods for Criminal Justice and Criminology at http://www.thomsonedu.com/criminaljustice/maxfield. Once at the companion website, select this specific chapter, click on “Chapter Resources,” then click on “Web Links.”

Race as a Social Construct  Our running example for this chapter centers on measuring race. Race is a concept that has been categorized in many ways over the years. The U.S. census in 2000 dramatically altered how race is quantified and measured for purposes of counting the population of the United States. Various technical reports issued by Census Bureau staff discuss the scope of problems in measuring race, especially among people of Hispanic or Latino heritage. Two papers by Elizabeth Martin and Eleanor Gerber, released in 2005 and 2006, respectively, are of particular interest. Find these and other reports that may be of interest on a web page maintained by the Survey Research Division of the Census Bureau. http://www.census.gov/srd/www/byyear.html

Performance Measures in New York  For several years, staff in the New York City Office of the Mayor have compiled detailed performance measures for city agencies. Beginning in 2002, very detailed statistics became available for neighborhood areas throughout New York. This provides a fascinating way to learn more about measures while exploring New York City. Begin at the following page: http://www.nyc.gov/html/ops/html/mmr/mmr_sub.shtml

In July 2006 this page included four scroll-down links. Notice the second, “Indicator definitions.” Scroll down different departments to learn how performance indicators are defined. For example, click on the Department of Health and Mental Hygiene; scroll down to see how various rodent control performance measures are defined. Then go back one page, click on the Mayor’s Management Report scroll-down, click next on Health and Mental Hygiene, and see how rodent control performed last year.

Go back and click next on “My Neighborhood Statistics.” You will eventually come to a screen asking for a street address or intersection. Filling in these fields brings up a map of the area together with tabbed tables that show area measures for different city services; “Public Safety and Legal Affairs” brings up reported crime for the area. For
example, Maxfield occasionally presents guest lectures at the John Jay College of Criminal Justice, located at 10th Ave. and 58th St. in Manhattan. If you type in the address, then click on the public safety link, you’ll find crime reports for the 18th Precinct. Think of the tabs containing statistics for different services as measures of different dimensions of life in New York.

Next try the city Department of Parks and Recreation. The address below brings up a page of detailed information on parks performance measures, included a 42-word definition of “litter.”

http://www.nycgovparks.org/sub_about/parks_numbers/ratings.html

**Consumer Price Index** The Consumer Price Index (CPI) is a real index. In trying to measure the complex construct of the cost of living, it includes data on prices for a wide variety of consumer goods. This is because the cost of living cannot be measured by prices for any single item or small number of items. The website below includes links to a large number of publications. Scroll down near the bottom for a list of “fact sheets” for different dimensions of things people spend money on.

http://www.bls.gov/cpi/#publications

Then find the section on “relative importance of items in the Consumer Price Indexes.” It’s not obvious, so look around; the link below worked in November 2006:


This shows the hundreds of individual items (measures) that make up the CPI, together with weights assigned to represent how important each item is in the total index. You might find it interesting to compare how important different items are in the index to your own pattern of consumer spending.

**Additional Readings**

Best, Joel, *Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activists* (Berkeley: University of California Press, 2001). Despite the title, much of this entertaining and informative book describes problems with measurement. For example, page 45 tells us: “Measuring involves deciding how to go about counting.” Best emphasizes how ambiguity in measures of social problems makes it easy for advocates to exaggerate the frequency of such problems. Mass media often report and perpetuate inaccurate measures. What results, Best informs us, are *mutant statistics.*

Bureau of Justice Statistics, *Performance Measures for the Criminal Justice System* (Washington, DC: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, 1993). This collection of essays by prominent criminal justice researchers focuses on developing measures for evaluation uses. The discussion of general measurement issues as encountered in different types of justice agencies is uncommonly thoughtful. You will find this a provocative discussion of how to measure important constructs in corrections, trial courts, and policing. See especially the general essays by John DiIulio and James Q. Wilson.

Gaes, Gerald G., Camp, Scott D., Nelson, Julianne B., and Saylor, William G., *Measuring Prison Performance: Government Privatization and Accountability* (Walnut Creek, CA: AltaMira Press, 2004). This book stemmed partly from the BJS report, in an effort to expand how to measure the various dimensions of prisons. Another stated goal of the authors is to devise a system for comparing the performance of public and private correctional facilities. This is an excellent resource for anyone interested in corrections.

Langworthy, Robert (ed.), *Measuring What Matters: Proceedings from the Policing Research Institute Meetings* (Washington, DC: U.S. Department of Justice, Office of Justice Programs, National Institute of Justice, 1999). With the spread of community policing, researchers and officials alike have struggled with the question of how to measure police performance. Most people agree that simply counting crimes is not enough, but no one knows how to measure other dimensions of police performance. This document presents papers and discussions from a series of meetings in which police, researchers, reporters, and others discussed what matters in policing and how to measure it.

# Measuring Crime

How do you measure crime? How much crime is there? Researchers and policy makers have developed several different approaches to measuring crime. Nevertheless, there’s no definitive answer to the second question. We’ll examine a variety of strategies for measuring crime and discuss the strengths and weaknesses of each.

## Introduction 145

### General Issues in Measuring Crime 145

What Offenses? 145

What Units of Analysis? 146

What Purpose? 147

### Crimes Known to Police 148

Uniform Crime Reports 149

The UCR and Criteria for Measurement Quality 151

Incident-Based Police Records 152

The National Incident-Based Reporting System 152

### NIBRS and Criteria for Measurement Quality 154

### Measuring Crime Through Victim Surveys 155

The National Crime Victimization Survey 155

NCVS Redesign 157

Community Victimization Surveys 160

Comparing Victim Surveys and Crimes Known to Police 160

### Surveys of Offending 161

National Survey on Drug Use and Health 162

Monitoring the Future 163

(continued)
Introduction

Measures of crime are important for many criminal justice research purposes.

Having discussed the principles of measurement and measurement quality, we now turn to a basic task for criminal justice researchers: measuring crime. Crime is a fundamental variable in criminal justice and criminology. Descriptive and exploratory studies often seek to count how much crime there is in some specific area, a question of obvious concern to criminal justice officials as well as researchers. Explanatory studies, viewing crime as a dependent variable, seek to learn what causes crime; applied studies often focus on what actions might be effective in reducing crime.

Crime can also be an independent variable—for example, in a study of how crime affects fear or other attitudes, or whether people who live in high-crime areas are more likely than others to favor long prison sentences for drug dealers. Sometimes crime can be both an independent and a dependent variable, as in a study about the relationship between drug use and other offenses.

Whatever our research purpose, and whether we’re interested in what causes crime or what crime causes, measuring crime clearly is important. It’s also difficult; how to measure crime has long been a key research issue in criminology and criminal justice.

We have three objectives in this chapter. First, we’ll use the challenge of measuring crime to illustrate the more general measurement issues we considered in Chapter 5. Second, we’ll examine a range of available measures of crime and compare the strength and weaknesses of each. Finally, we’ll briefly consider some independent measures of crime developed for specific research or policy purposes.

General Issues in Measuring Crime

Researchers must decide on offenses, units of analysis, and purposes before specifying measures of crime.

At the outset, we must acknowledge some broad questions that influence all measures of crime: (1) What offenses should be measured? (2) What units of analysis should be used? and (3) What is the research or policy purpose in measuring crime?

What Offenses?

Let’s begin by proposing a conceptual definition of crime—one that will enable us to decide what specific types of crime we’ll measure.
Recall a definition from Michael Gottfredson and Travis Hirschi (1990:15), mentioned in Chapter 5: “Acts of force or fraud undertaken in pursuit of self-interest.” This is an interesting definition, but it is better suited to an extended discussion of theories of crime than to our purposes in this chapter. For example, we would have to clarify what was meant by self-interest, a term that has engaged philosophers and social scientists for centuries.

James Q. Wilson and Richard Herrnstein (1985:22) propose a different definition that should get us started: “A crime is any act committed in violation of a law that prohibits it and authorizes punishment for its commission.” Although other criminologists (Gottfredson and Hirschi, for example) might not agree with this conceptual definition, it has the advantage of being reasonably specific. We could be even more specific by consulting a state or federal code and listing the type of acts for which the law provides punishment.

Our list would be very long. For example, the Indiana Code (IC) includes harvesting ginseng root out of season (IC 14-31-3-16) and selling a switchblade knife (IC 35-47-5-2) as acts punishable by 6 months’ incarceration and a $1000 fine. Taking Indiana ginseng out of the state without permission (IC 14-31-3-20) and assault resulting in nonserious bodily injury to an adult (IC 35-42-2-1) are equivalent crimes that could bring a year’s incarceration and a fine of $5000. However we decide to measure crime, we certainly want to distinguish acts of violence and the sale of illegal weapons from irregularities concerning rare herbs.

What Units of Analysis?
Recall that units of analysis are the specific entities researchers collect information about. Chapter 4 considered individuals, groups, social artifacts, and other units of analysis. Deciding how to measure crime requires that we once again think about these units.

Crimes involve four elements that are often easier to recognize in the abstract than they are to actually measure. The most basic of these elements is the offender. Without an offender, there’s no crime, so a crime must, at a minimum, involve an offender. The offender is therefore one possible unit of analysis. We might decide to study burglars, auto thieves, bank robbers, child molesters, drug dealers, or people who have committed many different types of offenses.

Crimes also require some sort of victim, the second possible unit of analysis. We could study victims of burglary, auto theft, bank robbery, or assault. Notice that this list of victims includes different types of units: households or businesses for burglary, car owners for auto theft, banks for bank robbery, and individuals for assault. Some of these units are organizations (banks, businesses), some are individual people, some are abstractions (households), and some are ambiguous (individuals or organizations can own automobiles).

What about so-called victimless crimes like drug use, bookmaking, or prostitution? In a legal sense, victimless crimes do not exist because crimes are acts that injure society, organizations, and/or individuals. But studying crimes in which society is the principal victim—prostitution, for example—presents special
challenges, and specialized techniques have been developed to measure certain types of victimless crimes. In any event, it’s crucial to think about units of analysis in advance. Later, we will examine surveying victims as one approach to counting crime. But surveying individuals, organizations, and society involves fundamentally different tasks.

The final two elements of crimes—offense and incident—are closely intertwined and so will be discussed together. An offense is defined as an individual act of burglary, auto theft, bank robbery, and so on. The Federal Bureau of Investigation (FBI) defines incident as “one or more offenses committed by the same offender, or group of offenders acting in concert, at the same time and place” (Federal Bureau of Investigation, 2000:17; emphasis in original).

Think about the difference between offense and incident for a moment. A single incident can include multiple offenses, but it’s not possible to have one offense and multiple incidents. Of course, a single incident could include multiple victims.

To illustrate the different units of analysis—offenders, victims, offenses, and incidents—consider the examples in the box titled “Units of Analysis and Measuring Crime.” These examples help distinguish units from each other and illustrate the links among different units.

Notice that we have said nothing about aggregate units of analysis, a topic we examined in Chapter 4. We have considered only individual units, even though measures of crime are often reported for aggregate units of analysis—neighborhoods, cities, counties, states, and so on.

What Purpose?
Different strategies for measuring crime can be distinguished by their general purpose. Measuring crime has at least one of three general purposes: (1) monitoring, (2) agency accountability, and (3) research.

We measure crime for the purpose of monitoring in much the same way that we measure consumer prices, stock market activity, traffic fatalities, birthrates, population, unemployment, and HIV infection rates. This reflects more than merely a compulsion for counting things, however. We measure a variety of social, economic, demographic, and public health indicators to keep track of social and economic conditions, the size and age distribution of the population, and threats to public health. By the same token, one purpose for measuring crime is to monitor potential threats to public safety and security.

At the national level, two series of crime measures seek to assess “the magnitude, nature, and impact of crime in the Nation [sic]” (U.S. Department of Justice, 2004:1). We’ll examine the Uniform Crime Reports (UCR) and the National Crime Victimization Survey (NCVS) in detail later in this chapter. Here, we point out that the fundamental purpose for these two measures is monitoring. In the field of public health, this is referred to as a surveillance system (Wilt and Gabrel, 1998). Just as a variety of statistical series administered by the U.S. Public Health Service monitor the incidence of disease and death rates from various causes, the U.S. Department of Justice oversees two nationwide surveillance systems for measuring crime.

The second measurement purpose is agency accountability. Government agencies are obliged to keep records that document their actions and areas of responsibility. Such accountability is a basic principle of the U.S. version of democratic government and is one reason individual police departments measure crime. “You can’t manage what you can’t measure,” states the title page on Idaho’s 2004 state crime report, underscoring the significance of accountability (Idaho State Police, 2005).

The final purpose is research; measures of crime are made for research purposes that are distinct from the purpose of surveillance or accountability. We emphasize these different purposes for a reason we’ll encounter throughout this and subsequent chapters. Criminal justice research often uses measures of crime that are collected for surveillance or accountability purposes, not for research. It’s worth keeping that
point in mind when planning a research project that will measure crime with one or more existing data series.

**Crimes Known to Police**

*Police-based crime measures are the most widely used, but they are subject to certain types of error.*

The most widely used measures of crime are based on police records and are commonly referred to as “crimes known to police.” This phrase has important implications for understanding what police records do and do not measure. One obvious implication is that crimes not known to police cannot be measured by consulting police records. We can better understand the significance of this by considering the two ways police come to know about crime: observation and reports from other people.

Certain types of crimes are detected almost exclusively by observation—for example, traffic offenses and so-called victimless crimes. Police count drug sales because they observe the transaction; they count incidents of prosti-

---

### Units of Analysis and Measuring Crime

Figuring out the different units of analysis in counting crimes can be difficult and confusing at first. Much of the problem comes from the possibility of what database designers call one-to-many and many-to-many relationships. The same incident can have multiple offenses, offenders, and victims, or just one of each. Fortunately, thinking through some examples usually clarifies the matter. Our two examples are adapted from an FBI publication (2000:18).

<table>
<thead>
<tr>
<th>Example 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two males entered a bar. The bartender was forced at gunpoint to hand over all money from the cash register. The offenders also took money and jewelry from three customers. One of the offenders used his handgun to beat one of the customers, thereby causing serious injury. Both offenders fled on foot.</td>
</tr>
<tr>
<td>1 incident</td>
</tr>
<tr>
<td>1 robbery offense</td>
</tr>
<tr>
<td>2 offenders</td>
</tr>
<tr>
<td>4 victims (bar cash, 3 patrons)</td>
</tr>
<tr>
<td>1 aggravated assault offense</td>
</tr>
<tr>
<td>2 offenders</td>
</tr>
<tr>
<td>1 victim</td>
</tr>
</tbody>
</table>

Even though only one offender actually assaulted the bar patron, the other offender would be charged with assisting in the offense because he prevented others from coming to the aid of the assault victim.

<table>
<thead>
<tr>
<th>Example 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two males entered a bar. The bartender was forced at gunpoint to hand over all money from the cash register. The offenders also took money and jewelry from two customers. One of the offenders, in searching for more people to rob, found a female customer in a back room and raped her there, outside of the view of the other offender. When the rapist returned, both offenders fled on foot.</td>
</tr>
<tr>
<td>This example includes two incidents because the rape occurred in a different place and the offenders were not acting in concert. And because they were not acting in concert in the same place, only one offender was associated with the rape incident.</td>
</tr>
<tr>
<td>Incident 1</td>
</tr>
<tr>
<td>1 robbery offense</td>
</tr>
<tr>
<td>2 offenders</td>
</tr>
<tr>
<td>3 victims (bar cash, 2 patrons)</td>
</tr>
<tr>
<td>Incident 2</td>
</tr>
<tr>
<td>1 rape offense</td>
</tr>
<tr>
<td>1 offender</td>
</tr>
<tr>
<td>1 victim</td>
</tr>
</tbody>
</table>
tution because they witness solicitations. Ob-
viously, police detect traffic offenses through
observation. Most other crimes, however, are
detected and counted because they are reported
to police by other people—victims or witnesses.
Victims report burglaries, robberies, assaults,
purse snatchings, and other offenses to police;
witnesses also report crimes.

Recognizing that police measure crime in
these two ways—observation and reports by
others—we can readily think of crimes that
are not well measured by police records. Con-
sider shoplifting, for example. Many instances
of shoplifting are observed neither by police
nor by other people who might report them
to police. Those instances are not detected
and therefore not measured. Shoplifting cer-
tainly is included in our conceptual defi ni-
tion of an act committed in violation of a law
that prohibits it and that authorizes punish-
ment. However, measuring shoplifting by count-
ing crimes known to police would omit many
offenses.

Thinking about crime measured almost
exclusively by police observation—victimless
 crimes and traffic offenses—should make you
realize that crimes known to police are not a
good measure of these types of offenses either.
Most of us have committed traffic offenses and
not been caught. Similarly, most instances of
drug sales, not to mention drug possession, are
not detected.

The other way police measure crime is also
imperfect. Many crimes are not reported to po-
lace, especially minor thefts and certain types of
assaults. People don’t report crimes for several
reasons, which tend to vary by type of crime. At-
tempted thefts may not be reported because no
property was lost. Many minor assaults or other
personal crimes are considered by victims to be
private matters that they will settle themselves
without involving the police. Other victims
may believe that minor losses are not important
enough to trouble police or that reporting a
crime would make no difference because police
could neither capture the offender nor recover
the stolen property (Bureau of Justice Statistics,
2005: Table 102).

Another problem with police measurement
of crime undermines the meaning of the phrase
“crimes known to police.” Research has shown
what some people may have personally expe-
rienced: Police do not always make official re-
cords of crimes they observe or crimes reported
to them. Donald Black (1970) and Albert Reiss
(1971) described a number of factors that influ-
eince police decisions on whether to officially
record crimes they know about. Assaults, for
example, between people who know each other
well or are related to each other are less likely to
be recorded as assaults than are fights between
strangers. If a victim urges a police officer not to
arrest someone or not to press charges against
an offender, the officer is less likely to treat the
incident as a crime. Black also found that po-
lice more often made official crime reports for
incidents that involved victims of higher so-
cioeconomic status. In cases where someone is
repeatedly victimized, usually by the same of-
fender, police less often record later incidents.
Research in England (Laycock, 2001) and in
New Jersey (Mele, 2003) finds nonrecording of
repeat incidents for domestic violence.

Finally, Richard and Carolyn Rebecca Block
(1980) found that robberies in Chicago were
less often recorded if no weapon was used, if
victims resisted the offender, or if the attempt
was unsuccessful. They concluded that robber-
ies are counted by police if they are “based on a
stereotypical idea of robbery—a helpless victim
attacked by a gun-wielding thug. The more an
incident approximates this ideal robbery, the
more likely it will become a robbery statistic”

**Uniform Crime Reports**

Police measures of crime form the basis for the
FBI’s Uniform Crime Reports (UCR), a data se-
ries that has been collected since 1930 and has
been widely used by criminal justice research-
ers. However, certain characteristics and proce-
dures related to the UCR affect its suitability
as a measure of crime. Most of our discussion highlights shortcomings in this regard, but keep in mind that the UCR is still a very useful measure for researchers and public officials.

Because UCR data are based on crimes reported to police, they share the measurement problems mentioned earlier. However, the FBI crime counts include three additional sources of measurement error.

First, the UCR does not even try to count all crimes reported to police. What are referred to as Part I offenses are counted if these offenses are reported to police (and recorded by police). Part I offenses include murder and non-negligent manslaughter, forcible rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft (Federal Bureau of Investigation, 1999). Other offenses, referred to as Part II crimes, are counted only if a person has been arrested and charged with a crime. The UCR therefore does not include such offenses as shoplifting, drug sale or use, fraud, prostitution, simple assault, vandalism, receiving stolen property, and all other nontraffic offenses unless someone is arrested. This means that a large number of crimes reported to police are not measured in the UCR.

Second, one of the reasons Part II offenses are counted only if an arrest is made is that individual states have different definitions of crimes. The operational definition of crime can vary from state to state, and this introduces another source of measurement error into the FBI data. For example, the state of Louisiana includes verbal threats in its counts of assaults, but most other states do not (Justice Research and Statistics Association, 1996:20).

The FBI compiles its UCR figures from data submitted by individual states or local law enforcement agencies. In some states, local police and sheriffs’ departments send their crime reports to a state agency, which forwards the data to the FBI. In other states, local law enforcement agencies send crime data directly to the FBI. However, not all local police and sheriffs’ departments send complete crime report data to either their state agency or the FBI. Inconsistency also exists in the quality of data submitted. In other words, individual states, cities, and counties vary in the quality and completeness of crime data sent to the FBI and reported in the annual UCR publication, Crime in the United States. Just as the decennial census cannot count everyone who lives in the United States, the FBI is not able to reliably count all crimes—either Part I or Part II offenses—that occur in the United States. In some cases, especially since the early 1990s, only a small fraction of law enforcement agencies in a state report under the UCR program. Michael Maltz (1999) describes the scope of this problem (greater than most people realize) and efforts to estimate the amount of missing reports.

UCR data can also suffer from clerical, data processing, and, in some cases, political problems. For example, Henry Brownstein (1996) describes his experience as a senior analyst in the New York Division of Criminal Justice Services, a state agency that compiles local crime reports for submission under the UCR program. The accuracy of data from cities and counties in New York is affected by staff shortages that undermine efforts to verify local reports, maintenance of an aging computer program that compiles UCR reports, and what Brownstein calls the “New York City reconciliation.” Here is how he describes that problem (1996:22–23):

As localities and agencies within localities compete amongst themselves for a greater share of State resources, they all compete to show that they are responsible for a greater share of the problem that the resources will be used to solve. Consequently, everyone wants credit for reported crimes and arrests. In New York City, where there are so many competing jurisdictions and agencies, this translates as a problem of duplicate reporting. So every summer a senior data entry clerk conducts the reconciliation, separating out by hand and by assumption the duplicate reports of the
same crimes and arrests submitted by different jurisdictions and agencies.

The third source of measurement error in the UCR is produced by the hierarchy rule used by police agencies and the FBI to classify crimes. Under the hierarchy rule, if multiple crimes are committed in a single incident, only the most serious is counted in the UCR. For example, if a burglar breaks into a home, rapes one of the occupants, and flees in the homeowner’s car, at least three crimes are committed—burglary, rape, and vehicle theft. Under the FBI hierarchy rule, however, only the most serious crime, rape, is counted in the UCR, even though if captured, the offender could be charged with all three offenses. In the examples described in the box “Units of Analysis and Measuring Crime,” the UCR would count one offense in each incident: a single robbery in the first example and rape in the second.

The UCR and Criteria for Measurement Quality

Let’s now consider how using the UCR to operationalize and measure crime satisfies the criteria for measurement quality we discussed in Chapter 5. Are the UCR data exclusive, exhaustive, valid, and reliable? First, the UCR clearly is neither an exclusive nor an exhaustive measure. Many crimes are not counted (nonexhaustive), and the hierarchy rule means that crime definitions are not strictly exclusive, because only the most serious crime in an incident in which multiple crimes are committed is counted. It does not help us if, for example, we are especially interested in burglary because burglaries are not counted if a rape, robbery, or murder is committed in the same incident.

Because the UCR does not count all crimes, we can rightly question its validity. The UCR does not really measure the concept of crime as we have defined it: any act committed in violation of a law that prohibits it and authorizes punishment for its commission.

Finally, is the UCR a reliable measure? Not all law enforcement agencies submit complete reports to the FBI, and the quality of the data submitted varies. Inconsistencies in reporting, such as those Brownstein describes for New York, produce problems with the reliability of UCR data. Also recall the Blocks’ study and other research findings that police do not always make records of crimes that come to their attention. When individual police officers exercise discretion in making crime reports, the reliability of measuring crime through the UCR is further undermined.

Recognizing these and other problems with UCR data, the FBI has been pursuing two initiatives to enhance the validity and reliability of crimes known to police. First, under its Quality Assurance Review program, staff from the FBI Criminal Justice Information Services Division audit crime data submitted by samples of reporting agencies. Results from this data quality review are shared with staff from state agencies; 11 states participated in audits conducted in 2001 (Federal Bureau of Investigation, 2002:5). The second FBI initiative is actually a from-the-ground-up overhaul of crime reporting, a topic we will examine in the next section.

Before we move on to other approaches to measuring crime, consider how units of analysis figure into UCR data. The UCR system produces what is referred to as a summary-based measure of crime. This means that UCR data include summary, or total, crime counts for reporting agencies; in most cases these are cities or counties. UCR data therefore represent groups as units of analysis. Crime reports are available for cities or counties, and these may be aggregated upward to measure crime for states or regions of the United States. But UCR data available from the FBI cannot represent individual crimes, offenders, or victims as units.

Recall that it is possible to aggregate units of analysis to higher levels, but it is not possible to disaggregate grouped data to the individual level. Because UCR data are aggregates,
they cannot be used in descriptive or explanatory studies that focus on individual crimes, offenders, or victims. UCR data are therefore restricted to the analysis of such units as cities, counties, states, or regions.

**Incident-Based Police Records**

The U.S. Department of Justice sponsors two series of crime measures that are based on incidents as units of analysis. The first of these incident-based measures, Supplementary Homicide Reports (SHR), was begun in 1961 and is a spin-off from the UCR program, as implied by the word supplementary.

Local law enforcement agencies submit detailed information about individual homicide incidents under the SHR program. This includes information about victims and, if known, offenders (age, gender, race); the relationship between victim and offender; the weapon used; the location of the incident; and the circumstances surrounding the killing.

Notice how the SHR relates to our discussion of units of analysis. Incidents are the basic unit and can include one or more victims and offenders; because the series is restricted to homicides, offense is held constant.

Because the SHR is an incident-based system, investigators can use SHR data to conduct descriptive and explanatory studies of individual events. For example, it’s possible to compare the relationship between victim and offender for male victims and female victims or to compare the types of weapons used in killings by strangers and killings by nonstrangers. Such analyses are not possible if we are studying homicide using UCR summary data. If our unit of analysis is jurisdiction—city or county, for example—we can examine only the aggregate number of homicides in each jurisdiction for a given year; it will not be possible to say anything about individual homicide incidents.

Crime measures based on incidents as units of analysis therefore have several advantages over summary measures. It’s important to keep in mind, however, that SHR data still represent crimes known to police and recorded by police. Of course, records of homicides will be better represented in police records than will, say, records of shoplifting, but clerical and other errors can still be a factor. Brownstein (1996) describes some of these problems, and Maxfield (1989) discusses some validity concerns with respect to the SHR. Most potential errors in the SHR are due to recording and record-keeping practices, topics we will discuss in Chapter 11.

**The National Incident-Based Reporting System**

The most recent development in police-based measures at the national level is the ongoing effort by the FBI and the Bureau of Justice Statistics (BJS) to convert the UCR to a National Incident-Based Reporting System (NIBRS, pronounced ny’-bers). Planning for replacement of the UCR began in the mid-1980s, but because NIBRS represents major changes, law enforcement agencies have shifted only gradually to the new system.

Briefly, NIBRS is a Very Big Deal. We’ll first put things in perspective by comparing NIBRS and the UCR for a single state, Idaho, which currently reports incident-level measures. Then we’ll discuss some of the specific features of NIBRS, concluding with comments on progress in implementing the system.

About 17,000 law enforcement agencies report UCR summary data each year; that’s 17,000 annual observations, one for each reporting agency. In 2004, 106 agencies in Idaho reported UCR data, so Idaho submitted a maximum of 106 observations for 2004. Under NIBRS, Idaho reported over 95,000 incidents in 2004. So, for Idaho, shifting from the summary UCR system for measuring crime to the incident-based NIBRS system meant shifting from 106 units (UCR reporting jurisdictions) to more than 95,000 units. In other words, rather than reporting 106 summary crime counts for eight UCR Part I offenses, Idaho reported detailed information on 95,522 individual in-
cidents. And this is Idaho, which ranked 39th among the states in 2000 resident population!

These numbers illustrate the main difference between NIBRS and the UCR system: reporting each crime incident rather than reporting the total number of certain crimes for each law enforcement agency. But the significance of shifting from reporting aggregate numbers to reporting individual incidents lies in the type of information that is available about each incident. In essence, NIBRS measures many features of each incident, and each of these features is reported individually. Table 6.1 lists most of the “segments” or categories of information recorded for each incident, together with examples of the information recorded within each segment.

As Table 6.1 shows, NIBRS includes much more detailed information about individual incidents, together with the offenses, offenders, and victims within each incident. Referring back to our earlier discussion of the elements of crime—incidents, offenders, offenses, and victims—notice how the information in Table 6.1 is organized around each incident. Each incident can include one or more offenses, offenders, and victims, as we described in the box on units of analysis.

In addition, NIBRS guidelines call for gathering information about a much broader array of offenses. Whereas the UCR reports information about eight Part I offenses, NIBRS is designed to collect detailed information on 46 “Group A” offenses. Table 6.2 demonstrates the significance of this by showing NIBRS crime data for Idaho in 2004. Compare the top part of the table, reporting crime counts for UCR index offenses, to the bottom part. Additional NIBRS Group A offenses more than doubles the number of crimes “known to police” in Idaho (43,611 UCR Part I, plus 51,911 additional Group A). Simple assault and vandalism are by far the most common of these additional offenses, but drug violations accounted for almost 13,000 offenses in 2004 (drug violations plus drug equipment violations).

Collecting detailed information on each incident for each offense, victim, and offender, and doing so for a large number of offense types, represents the most significant changes in NIBRS compared with the UCR. Dropping the hierarchy rule is also a major change, but that is a consequence of incident-based reporting. NIBRS incorporates additional important changes (Federal Bureau of Investigation, 2000:12–15):

1. **Victim type.** Table 6.1 lists most categories for victim type; the most notable of these is “society/public,” a category that has the effect of annulling the phrase “victimless crime.” Whereas UCR summary data combine all types of victims (individuals, businesses, others) into one summary measure, NIBRS makes it possible to distinguish different categories of victims.

2. **Attempted/completed.** UCR summary reports include both attempted and completed offenses, but it’s not possible to distinguish them. NIBRS adds a category to indicate
whether each offense within each incident was attempted or completed.

3. **Drug-related offenses.** Because the 1980s witnessed a growing public concern with illegal drug use, NIBRS includes provisions to assess offender drug use in nondrug offenses and to record whether drugs or drug paraphernalia were seized.

4. **Computers and crime.** The 1980s also witnessed an explosion of computer use, including computers as instruments and targets of crime. Categories to record the involvement of computers are included.

### NIBRS and Criteria for Measurement Quality

You now have some idea about the potential wealth of information available from NIBRS and the ways in which this new system represents major changes from the summary-based UCR. How does incident-based reporting fare on our criteria for measurement quality? After considering NIBRS in light of our earlier comments on the UCR, some improvements are clearly evident. Eliminating the hierarchy rule means offense classifications are mutually exclusive. Is NIBRS exhaustive? Table 6.2 lists more offenses compared with the UCR, but additional offenses are recorded only for crimes that result in arrest; not all crimes are counted.

In at least one sense, NIBRS data hold the promise of being more reliable. The FBI has produced very thorough documentation on how to record and classify incidents and their component records. Creating auditing standards and requiring submission of data on computer readable media also enhance reliability. Finally, the FBI requires that state records systems be certified before the state can submit incident based reports.

The production of NIBRS data is, however, still a selective process: Crimes are selectively reported to police and selectively recorded by police. And researchers are likely to encounter three practical problems in using NIBRS data to measure crime.

First, law enforcement agencies using NIBRS report huge amounts of data. For example, data files available for analysis include over 3.4 million incidents for the year 2002. And those 3.4 million incidents represent less than 20 percent of all reported crime in the United States. Researchers accustomed to working with a few hundred summary reports must

### Table 6.2 Crime in Idaho, 2004

<table>
<thead>
<tr>
<th>UCR Part I Offenses</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Murder, non-negligent manslaughter</td>
<td>35</td>
</tr>
<tr>
<td>Rape</td>
<td>577</td>
</tr>
<tr>
<td>Robbery</td>
<td>247</td>
</tr>
<tr>
<td>Aggravated assault</td>
<td>2594</td>
</tr>
<tr>
<td>Burglary</td>
<td>7700</td>
</tr>
<tr>
<td>Larceny</td>
<td>29,442</td>
</tr>
<tr>
<td>Motor vehicle theft</td>
<td>2696</td>
</tr>
<tr>
<td>Arson</td>
<td>320</td>
</tr>
<tr>
<td><strong>Subtotal</strong></td>
<td><strong>43,611</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Additional NIBRS Group A Offenses</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple assault</td>
<td>14,192</td>
</tr>
<tr>
<td>Intimidation</td>
<td>1766</td>
</tr>
<tr>
<td>Bribery</td>
<td>7</td>
</tr>
<tr>
<td>Counterfeit/forgery</td>
<td>1982</td>
</tr>
<tr>
<td>Destruction of property</td>
<td>14,516</td>
</tr>
<tr>
<td>Drug violations</td>
<td>6667</td>
</tr>
<tr>
<td>Drug equipment violations</td>
<td>6329</td>
</tr>
<tr>
<td>Embezzlement</td>
<td>295</td>
</tr>
<tr>
<td>Extortion/blackmail</td>
<td>12</td>
</tr>
<tr>
<td>Fraud</td>
<td>2426</td>
</tr>
<tr>
<td>Gambling</td>
<td>9</td>
</tr>
<tr>
<td>Kidnapping/abduction</td>
<td>236</td>
</tr>
<tr>
<td>Pornography/obscene material</td>
<td>33</td>
</tr>
<tr>
<td>Prostitution</td>
<td>10</td>
</tr>
<tr>
<td>Forcible sex offenses</td>
<td>1200</td>
</tr>
<tr>
<td>Nonforcible sex offenses</td>
<td>215</td>
</tr>
<tr>
<td>Stolen property</td>
<td>631</td>
</tr>
<tr>
<td>Weapons violations</td>
<td>1385</td>
</tr>
<tr>
<td><strong>Subtotal</strong></td>
<td><strong>57,911</strong></td>
</tr>
<tr>
<td><strong>Group A Total</strong></td>
<td><strong>95,522</strong></td>
</tr>
</tbody>
</table>

now analyze hundreds of thousands of incident reports.

Second, implementing NIBRS requires enormous effort by individual law enforcement agencies to comply with the FBI’s detailed specifications on record keeping and reporting. Because it is a computer-based system, individual agencies must either develop the necessary data processing capability or adapt existing systems to NIBRS specifications. The FBI and BJS have promoted the shift to NIBRS with technical and financial assistance for state and local agencies. Somewhat paradoxically, this has meant that smaller states and smaller police and sheriffs’ departments within states have been more readily able to develop NIBRS-compliant systems from the ground up (Maxfield, 1999). Most larger agencies developed customized records management systems some years ago, and adapting those systems to NIBRS standards proved to be difficult and costly (MEGG Associates, 1996).

Third, like UCR reporting, NIBRS is voluntary; no agency is required to submit crime reports to the FBI in any form. Given the major expense and effort involved in shifting to NIBRS, the conversion process has been slow. In 1988, the FBI published the first extensive guidelines on developing and implementing NIBRS. At the end of 2005, 29 states were certified to report NIBRS data to the FBI (Justice Research and Statistics Association, 2006). Though this is over half of all states, only some agencies in each state reported NIBRS data, and all reporting agencies represent about 22 percent of the U.S. population.

In the future, incident-based police records will become more readily available and will cover a larger number of law enforcement agencies. In fact, many agencies have developed their own incident-based records systems independent of NIBRS, largely because of major advances in computing technology (Maxfield, 1999). Furthermore, researchers are beginning to analyze NIBRS data, something that is certain to prompt other researchers to do the same. For example, see studies of child abuse (Finkelhor and Ormond, 2004; Snyder, 2000), hate crimes (Nolan, Akiyama, and Berhanu, 2002), and domestic violence (Vazquez, Stohr, and Perkiss, 2005).

Measuring Crime Through Victim Surveys

Victim surveys are alternative measures of crime but are still subject to error.

Recognizing the shortcomings associated with using measures of crime known to police, we now consider alternative approaches. Conducting a victim survey that asks people whether they have been the victim of a crime is one alternative. Survey research methods will be described in detail in Chapter 9. For now, we assume that you have a general understanding of what a survey involves—presenting a sample of people with carefully worded questions and recording their responses.

In principle, measuring crime through surveys has several advantages. Surveys can obtain information on crimes that were not reported to police. Asking people about victimizations can also measure incidents that police may not have officially recorded as crimes. Finally, asking people about crimes that may have happened to them provides data on victims and offenders (individuals) and on the incidents themselves (social artifacts). Like an incident-based reporting system, a survey can therefore provide more disaggregated units of analysis. When conducted in a rigorous, systematic fashion, surveys can yield reliable measures.

The National Crime Victimization Survey

Since 1972, the U.S. Census Bureau has conducted national surveys of crime and victimization, currently known as the National Crime Victimization Survey (NCVS). An earlier version of the NCVS was launched following pilot studies in the mid-1960s by President Lyndon Johnson’s Commission on Law Enforcement
and Administration of Justice, commonly known as the President’s Crime Commission. One of the primary reasons for conducting crime surveys was to illuminate what came to be referred to as the “dark figure of unreported crime.” The NCVS is based on a nationally representative sample of households and uses uniform procedures to select and interview respondents, which enhances the reliability of crime measures. Because individual people living in households are interviewed, the NCVS can be used in studies in which individuals or households are the unit of analysis.

The NCVS cannot measure all crimes, however, in part because of the procedures used to select victims. Because the survey is based on a sample of households, it cannot count crimes in which businesses or commercial establishments are the victims. Bank robberies, liquor store holdups, shoplifting, embezzlement, political corruption, and securities fraud are examples of crimes that cannot be systematically counted by interviewing household members. Samples of banks, gas stations, retail stores, business establishments, elected officials, or stockbrokers would be needed to measure those crimes. In much the same fashion, crimes directed at homeless victims cannot be counted by surveys of households like the NCVS.

What about “victimless” crimes? For example, think about how you would respond to a Census Bureau interviewer who asked whether you had been the victim of a drug sale. If you have bought drugs from a dealer, you might think of yourself as a customer rather than as a victim. Or if you lived near a park where drug sales were common, you might think of yourself as a victim even though you did not participate in a drug transaction. The point is that victim surveys are not good measures of victimless crimes because the respondents can’t easily be conceived as victims.

Measuring certain forms of delinquency through victim surveys presents similar problems. Status offenses such as truancy and curfew violations do not have identifiable victims who can be included in samples based on households. Homicide and manslaughter are other crimes that are not well measured by victim surveys, for obvious reasons.

Because the NCVS, by design, excludes many types of crimes, you should recognize potential validity problems. But what about the reliability of crime surveys in general and the NCVS in particular? Because it is a survey, the NCVS is subject to the errors and shortcomings associated with that method of measuring concepts. In the following paragraphs, we will consider some reliability problems that sometimes emerge when surveys are used to count crime.

For many years, NCVS interviewers began the section of the survey on victimization with this introduction: “Now I’d like to ask some questions about crime. They refer only to the last six months—between [date 6 months before interview] and [date of interview]. During the last six months, did anyone [crime description]?” (Bureau of Justice Statistics, 1994: 122). Respondents were then asked a series of screening questions to determine whether they might have been a crime victim in the previous 6 months. Those who answered yes to screening questions were next presented with detailed questions about the incident.

Asking people about crime in this way brings up the possibility of different types of recall error. First, respondents may simply not remember some incidents. This problem is particularly acute for minor crimes such as theft and for people who may have been the victim of more than one crime in the 6-month recall period. The second recall problem, known as telescoping, involves respondents not accurately recalling when an incident occurred. Forward telescoping means that people may respond to questions by mentioning crimes that occurred more than 6 months ago, thus bringing past incidents forward into the recall period. In contrast, backward telescoping means that respondents inaccurately recall recent crimes as occurring in the more distant past. Because the NCVS tries to count crimes that occur every 6 months, forward or backward telescoping can produce unreliable counts. However, the NCVS
is designed to reduce the possibility of telescoping, in part by specifying the relatively short recall window of 6 months.

A different type of recall problem affects people who have been victimized several times during the 6-month reference period. The Bureau of Justice Statistics (BJS) describes this as a problem with *series victimizations*, which are defined as six or more similar but separate crimes that the respondent cannot describe individually to an interviewer (Bureau of Justice Statistics, 2005). Series victimizations constitute a problem for surveys because it is not clear how they should be counted or combined with individual crimes that the respondent can describe separately. The potential impact of series victimizations is substantial, especially for violent crimes. For example, Callie Marie Rennison (2001) estimates that series crimes account for 6–7 percent of all violent crimes measured by the NCVS.

Finally, the NCVS potentially underestimates incidents in which the victim and offender know each other—domestic violence or other assaults involving friends or acquaintances, for example. Respondents may not tell interviewers about nonstranger crimes for various reasons. Some domestic violence victims view the assaults as a personal problem, not as a crime, and so may not mention the incidents. Other persons—victims of rape or domestic violence, for example—may feel shame or embarrassment and not wish to talk about their experiences. In addition, respondents who have been victimized by a family member may fear some further assault if they discuss a past incident.

**NCVS Redesign**

To address these and other concerns about the ability of the NCVS to measure different types of crime, the BJS substantially revised the survey in the 1990s. Victimization estimates from interviews conducted in 1993 reflect the full scope of the redesign effort after various changes in survey procedures were gradually incorporated. Even though the redesign took effect over 10 years ago, we discuss it at some length here because it illustrates how features of survey design and questionnaire construction affect measures of crime. In addition, researchers continue to examine how changes in the NCVS affected the quality of crime data produced (Lauritsen, 2005; Cantor and Lynch, 2005).

For the most part, the redesign effort focused on obtaining better measures of domestic violence and sexual assault, together with methods to help respondents recall a broader range of incidents. Here is a brief summary of major changes undertaken to produce better measures of victimization (Bureau of Justice Statistics, 1996; U.S. Bureau of the Census, 1994):

- Revised screening questions and added cues throughout the interview to help respondents recall and distinguish minor incidents
- More direct questions on rape and other sexual crimes, reflecting the belief that people’s willingness to discuss such incidents has increased in recent years
- Greater attention to measuring victimizations by someone the respondent knows, including incidents of domestic violence
- Gradual increase in the use of telephone interviews to replace in-person interviews
- An increase in the threshold for series victimizations from three to six, consistent with other efforts to help respondents distinguish individual victimizations

Before taking a look at how the redesigned NCVS affected victimization rates, let’s consider the changes in survey screening rates. These are summarized in Table 6.3. The redesigned NCVS presents more specific questions and cues, in effect encouraging respondents to think about specific types of incidents. Notice in particular the explicit reference to “forced or unwanted sexual acts” in the redesigned survey; the old NCVS made no direct reference to rape or other sexual assaults, implicitly including them in the general categories of “attack you” and “try to attack you.” Redesigned
NCVS questions also refer more directly to offenses committed by someone known to the respondent:

People often don’t think of incidents committed by someone they know. Did you have something stolen from you OR were you attacked or threatened by-

(a) Someone at work or at school-
(b) A neighbor or friend-
(c) A relative or family member-
(d) Any other person you’ve met or known?

Table 6.4 compares estimates of victimization rates from the first years of the redesigned NCVS with rates from pre-redesign versions.

Table 6.3 Comparison of Old and Redesigned NCVS Screening Questions

<table>
<thead>
<tr>
<th>General Screening Question</th>
<th>Old NCVS</th>
<th>Redesigned NCVS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Was anything stolen from you while you were away from home—for instance, at work, in a theater or restaurant, or while traveling?</td>
<td></td>
<td>Were you attacked or threatened OR did you have something stolen from you</td>
</tr>
<tr>
<td></td>
<td></td>
<td>a. At home including the yard or porch</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b. At or near a friend’s or neighbor’s home</td>
</tr>
<tr>
<td></td>
<td></td>
<td>c. At work or school</td>
</tr>
<tr>
<td></td>
<td></td>
<td>d. In a place such as a shopping mall, laundry room, restaurant, bank, or airport</td>
</tr>
<tr>
<td></td>
<td></td>
<td>e. While riding in any vehicle</td>
</tr>
<tr>
<td></td>
<td></td>
<td>f. On the street or in a parking lot</td>
</tr>
<tr>
<td></td>
<td></td>
<td>g. At such places as a party, theater, gym, picnic area, or while fishing or hunting OR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>h. Did anyone ATTEMPT to attack or attempt to steal anything belonging to you from any of these places?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Screening Questions for Violent Crimes</th>
<th>Old NCVS</th>
<th>Redesigned NCVS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did anyone take something from you by using force, such as by a stickup, mugging, or threat?</td>
<td></td>
<td>Has anyone attacked or threatened you in any of these ways</td>
</tr>
<tr>
<td>Did anyone TRY to rob you by using force or threatening to harm you?</td>
<td></td>
<td>a. With any weapon such as a gun or knife</td>
</tr>
<tr>
<td>Did anyone beat you up, attack you, or hit you with something such as a rock or bottle?</td>
<td></td>
<td>b. With anything like a baseball bat, frying pan, scissors, or stick</td>
</tr>
<tr>
<td>Were you knifed, shot at, or attacked with some other weapon by anyone at all?</td>
<td></td>
<td>c. By something thrown, such as a rock</td>
</tr>
<tr>
<td>Did anyone THREATEN to beat you up or threaten you with a knife, gun, or some other weapon, NOT including telephone threats?</td>
<td></td>
<td>d. Include any grabbing, punching, or kicking</td>
</tr>
<tr>
<td>Did anyone TRY to attack you in some other way?</td>
<td></td>
<td>e. Any rape, attempted rape, or other type of sexual assault</td>
</tr>
<tr>
<td></td>
<td></td>
<td>f. Any face-to-face threats OR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>g. Any attack or threat or use of force by anyone at all? Please mention it even if you were not certain it was a crime.</td>
</tr>
</tbody>
</table>

Incidents involving forced or unwanted sexual acts are often difficult to talk about. Have you ever been forced or coerced to engage in unwanted sexual activity by-

(a) Someone you didn’t know before |
(b) A casual acquaintance |
(c) Someone you know well?

Source: Adapted from Bachman and Saltzman (1995:8).
Panel A, reproduced from a study examining violence against women (Bachman and Saltzman, 1995), shows rates of violent crime victimization by victim–offender relationship for males and females. Notice that the redesigned survey yields higher rates for all categories of relationship but that the increase tends to be greater for nonstranger offenses. The category “intimate” includes both married persons and those in a quasi-marital or other intimate relationship. Redesigned survey results indicate that just less than 1 percent (9.3 per 1000) of female respondents reported being the victim of a violent offense by an intimate partner or former partner. It therefore appears that efforts to uncover more incidents of violence perpetrated by non-strangers have been successful.

Similarly, Panel B in Table 6.4 compares victimization rates for specific violent offenses; estimates of rape and sexual assault victimization rates for females were about 2.5 times higher in 1994 than in 1991. Notice, however, that the redesigned NCVS includes “sexual assault other than rape,” so the estimates are not directly comparable. Males are more likely to be victims of both aggravated and simple assault, but as the redesigned NCVS shows, estimates of assault victimization rates increased more for females than for males. This is consistent with redesign efforts to obtain better estimates of violent offenses.

At this point, we want to underscore two general points about the redesign in particular and the NCVS in general. First, what we learn about crime from the NCVS or any other survey depends on what we ask and how we ask it. Keep in mind that measures of crime are affected in part by the procedures we use to make those measures. This is another example of an important point from Chapter 5: Changing how we operationalize measures can change the values we obtain for those measures. Presenting respondents with rather general questions about being attacked by someone discloses fewer crimes of violence than asking questions about specific kinds of violence. Likewise, when we specifically ask respondents to think of acts committed by someone they know, we will uncover more offenses than if we do not include such cues and prompts. David Cantor and James Lynch (2005) describe other examples of how the redesigned survey affected victimization measures for different subgroups of respondents.

The second point is a consequence of the first: Because of the redesigned NCVS, any effort to compare trends and changes in crime over time must take account of changes in measurement. BJS analysts are attentive to this and caution readers of their reports to be aware of changes in survey methods: “Data

### Table 6.4  Comparison of Violent Victimization Rate, Old and Redesigned NCVS

<table>
<thead>
<tr>
<th>Victim–Offender Relationship</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old NCVS (1987–91)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intimate</td>
<td>5.4</td>
<td>0.5</td>
</tr>
<tr>
<td>Other relative</td>
<td>1.1</td>
<td>0.7</td>
</tr>
<tr>
<td>Acquaintance/friend</td>
<td>7.6</td>
<td>13.0</td>
</tr>
<tr>
<td>Stranger</td>
<td>5.4</td>
<td>12.2</td>
</tr>
<tr>
<td>Redesigned NCVS (1992–93)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intimate</td>
<td>9.3</td>
<td>1.4</td>
</tr>
<tr>
<td>Other relative</td>
<td>2.8</td>
<td>1.2</td>
</tr>
<tr>
<td>Acquaintance/friend</td>
<td>12.9</td>
<td>17.2</td>
</tr>
<tr>
<td>Stranger</td>
<td>7.4</td>
<td>19.0</td>
</tr>
</tbody>
</table>

**Source:** Bachman and Saltzman (1995:8).

### Panel B: Rate of Violent Victimization per 1000 Persons

<table>
<thead>
<tr>
<th>Offense</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991 NCVS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rape</td>
<td>1.4</td>
<td>0.2</td>
</tr>
<tr>
<td>Robbery</td>
<td>3.5</td>
<td>7.8</td>
</tr>
<tr>
<td>Aggravated assault</td>
<td>4.4</td>
<td>11.5</td>
</tr>
<tr>
<td>Simple assault</td>
<td>13.4</td>
<td>20.9</td>
</tr>
<tr>
<td>1994 NCVS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rape and sexual assault</td>
<td>3.7</td>
<td>0.2</td>
</tr>
<tr>
<td>Robbery</td>
<td>4.1</td>
<td>8.1</td>
</tr>
<tr>
<td>Aggravated assault</td>
<td>8.1</td>
<td>15.3</td>
</tr>
<tr>
<td>Simple assault</td>
<td>26.6</td>
<td>35.9</td>
</tr>
</tbody>
</table>

based on the redesign are not comparable to data before 1993. . . . A number of fundamental changes were introduced when the survey was redesigned. These changes were phased into the sample over several years” (Bureau of Justice Statistics, 1996:8).

We emphasize these two points because beginning researchers are often less critical users of data such as the NCVS than are more experienced researchers; after all, the survey is sponsored by the U.S. Department of Justice and conducted by the Census Bureau. Examining Table 6.4 without being aware of survey changes would suggest that violent victimization rates had increased sharply from 1991 to 1994, a conclusion that would be misleading.

Community Victimization Surveys
Following the initial development of victim survey methods in the late 1960s, the Census Bureau completed a series of city-level surveys. These were discontinued for a variety of reasons, but researchers and officials in the BJS occasionally conducted city-level victim surveys in specific communities. In 1998, the BJS and the Office of Community Oriented Policing Services (COPS) launched pilot surveys in 12 large and medium-sized cities (Smith, Steadman, Minton, and Townsend, 1999).

The city-level initiative underscores one of the chief advantages of measuring crime through victim surveys—obtaining counts of incidents not reported to police. In large part, city-level surveys were promoted by BJS and COPS to enable local law enforcement agencies to better understand the scope of crime—reported and unreported—in their communities. Notice also the title of the first report: “Criminal Victimization and Perceptions of Community Safety in 12 Cities, 1998.” We emphasize perceptions to illustrate that city-level surveys can be valuable tools for implementing community policing, a key component of the 1994 Crime Bill that provided billions of dollars to hire new police officers nationwide. It is significant that the Department of Justice recognized the potential value of survey measures of crime and perceptions of community safety to develop and evaluate community policing.

The initial BJS/COPS effort was a pilot test of new methods for conducting city-level surveys. These bureaus jointly developed a guidebook and software so that local law enforcement agencies and other groups can conduct their own community surveys (Weisel, 1999). These tools also promise to be useful for researchers who wish to study local patterns of crime and individual responses. Although the community survey initiative lapsed after George Bush became president, the BJS continues to update software and make it available on its website http://www.ojp.usdoj.gov/bjs/abstract/cvs.htm.

Comparing Victim Surveys and Crimes Known to Police
Before moving on to the next section, let’s briefly compare the different ways of measuring crime we have discussed so far.

Researchers have devoted special attention to comparing data from the UCR and the NCVS in efforts to determine how the two measures differ and what the strengths and weaknesses of each method are for measuring crime. An early study by Wesley Skogan (1974) recognized that crime surveys and police data take fundamentally different approaches to measuring crime but that UCR and NCVS counts of robbery and auto theft are moderately related. More recently, Michael Rand and Callie Rennison (2002) demonstrate that adjusting NCVS and UCR data for basic design differences produces very similar estimates of violent crime. An even more ambitious effort by David Farrington and associates (2004) compares crime measures from police records and victim surveys for eight countries.

Whereas the UCR provides only summary measures of aggregate units, the NCVS yields more disaggregated data on individual victims, offenders, and incidents. This means that the NCVS is better suited to studies of individual
factors in the types of incidents covered by the survey.

Because SHR and NIBRS data are incident-based systems, they can also be used to study individual incidents, victims, and offenders. However, the SHR measures one type of crime only. NIBRS holds great promise for the future, but its limited coverage means that it cannot yet serve as a measure of nationwide incidents known to police. For the present, NIBRS data can be used only for those states that currently participate in the program.

By the way, this feature of NIBRS—its availability from only certain local agencies and states—highlights another characteristic of the NCVS worth emphasizing: It is not possible to use survey data to study victimization at the local level. This is because the National Crime Victimization Survey is just that: a national survey. It is designed to represent nationwide levels of crime (subject to the limitations we have discussed), but it cannot provide statistically reliable estimates of crime for most cities, counties, or states. Michael Maltz (1998) reports the little-known fact that the NCVS can be used to estimate victimization for more common incidents in the 10 largest states. And a report by Janet Lauritsen and Robin Schaum (2005) examines NCVS data for the nation’s three largest cities—New York, Los Angeles, and Chicago. These three cities were also among those where BJS/COPS community victim surveys were conducted, as described by Smith and associates (1999).

Despite the limited coverage of NIBRS, incident-based measures have great potential for criminal justice research. Like the NCVS, NIBRS yields details about individual incidents. Unlike the NCVS, NIBRS data can be examined for specific geographic areas at the local level and, eventually, for all states. A series of articles in the Journal of Quantitative Criminology presents the first published analysis of NIBRS data. Michael Maxfield (1999) summarizes many of the additional applications of NIBRS data for research and policy. We conclude this section by briefly discussing one area in which NIBRS data measure a type of offense not counted by other annual data series at the national level.

Despite improvements following the NCVS redesign, its sampling plan is still restricted to household residents ages 12 and over. We have discussed the significance of a household-based sample. What kinds of offenses might the NCVS miss by not counting victimizations for individuals younger than 12?

Howard Snyder’s (2000) analysis of NIBRS sexual assault data from 12 states provides a clue. About 34 percent of some 61,000 victims of sexual assault were under age 12; that means that over 20,000 sexual assault victims would not have been counted by the NCVS (Snyder, 2000:1–2). Furthermore, the expanded information available under NIBRS reveals that over 40 percent of sexual assault victims under age 12 were assaulted by a family member; an additional 50 percent of the offenders were known by the victims. Only about 5 percent of female victims under age 12 were assaulted by strangers.

These figures should make you think of child abuse, or at least child sexual abuse, as an offense type not measured by the NCVS. Of course, child abuse is undercounted by the UCR and NIBRS as well because much child abuse is not reported to police. NIBRS data represent new sources of information about this type of crime and other offenses that affect young victims.

Surveys of Offending

Delinquency, “victimless” crimes, and crimes rarely observed may be measured by self-report surveys.

Just as survey techniques can measure crime by asking people to describe their experiences as victims, self-report surveys ask people about crimes they may have committed. We might initially be skeptical of this technique: How truthful are people when asked about crimes they may have committed? Our concern is jus-
tified. Many people do not wish to disclose illegal behavior to interviewers even if they are assured of confidentiality. Others might deliberately lie to interviewers and exaggerate the number of offenses they have committed.

Self-report surveys, however, are the best method available for trying to measure certain crimes that are poorly represented by other techniques. Thinking about the other methods we have discussed—crimes known to police and victimization surveys—suggests several examples. Crimes such as prostitution and drug abuse are excluded from victimization surveys and underestimated by police records of people arrested for these offenses. Public order crimes and delinquency are other examples. A third class of offenses that might be better counted by self-report surveys are crimes that are rarely reported to or observed by police. Shoplifting and drunk driving are examples.

Think of it this way: As we saw earlier, all crimes require an offender. Not all crimes have clearly identifiable victims who can be interviewed, however, and not all crimes are readily observed by police, victims, or witnesses. If we can’t observe the offense and can’t interview a victim, what’s the next logical step?

There are no nationwide efforts in the United States to systematically collect self-report measures for a variety of crime types. Instead, periodic surveys yield information either on specific types of crime or on crimes committed by a specific target population. We will consider two ongoing self-report surveys here and comment on the validity and reliability of this method for measuring crime.

**National Survey on Drug Use and Health**2

Like the NCVS, the National Survey on Drug Use and Health (NSDUH) is based on a national sample of households. Both surveys are designed to monitor nationwide patterns. Unlike the victimization survey, however, the central purpose of the NSDUH is to obtain self-reports of drug use.

The survey has been conducted since 1971, with sampling and questioning procedures having been revised several times. The survey is currently sponsored by the Substance Abuse and Mental Health Services Administration in the U.S. Department of Health and Human Services. Persons ages 12 and over who live in households are the target population. In the 2004 sample, about 68,000 individuals responded to questions regarding their use of illegal drugs, alcohol, and tobacco (Substance Abuse and Mental Health Services Administration, 2005). Because it has been conducted for over 3 decades, the NSDUH provides information on trends and changes in drug use among respondents. The 2004 survey was designed to obtain statistically reliable samples from the eight largest states in addition to the overall national sample.

Think for a moment about what sorts of questions we would ask to learn about people’s experience in using illegal drugs. Among other things, we would probably want to distinguish someone who tried marijuana once from daily users. The drug use survey does this by including questions to distinguish lifetime use (ever used) of different drugs from current use (used within the past month). You may or may not agree that, for example, use in the past month represents current use, but it is the standard used in regular reports on NSDUH results. That’s the operational definition of “current use.”

Two potential sources of measurement problems with the NSDUH should come to mind. We have already touched on the first concern: Do people tell the truth when asked about drug use? The NSDUH incorporates certain procedures to encourage candid, truthful responses from individuals. After obtaining basic demographic information about all household residents, interviewers conduct the rest of the interviews in a private area away from other household members. Interviewers

---

2Descriptive information about the National Survey on Drug Use and Health was obtained from Substance Abuse and Mental Health Services Administration (2005) website: http://oas.samhsa.gov/nsduh.htm.
do not directly ask about drug use. Instead, these questions, and questions about other illegal behaviors, are administered through computer-assisted self-interviewing. Respondents read questions on a computer screen or listen to questions through earphones and key in their responses. Beginning in 1999, the NSDUH instituted major changes in the question format and questionnaire design in an effort to improve the validity of self-report measures.

Although these procedures produce better measures of drug use than do interviews conducted by telephone, the NSDUH still underreports drug use (Gfroerer, Eyerman, and Chromy, 2002). In fact, the circumstances of the NSDUH interview appear to affect reporting by adolescents. Joseph Gfroerer (1993) found higher reported rates of drug use by youths from surveys administered in the classroom compared with surveys administered at home. We will discuss an example of school-based surveys shortly, but let’s first briefly consider the second source of concern in using the NSDUH as a measure of drug use.

Recall the earlier point about how the household sample design of the NCVS measures only victimizations that affect households or household residents. Commercial establishments, people in institutional quarters—military housing or work-release community corrections centers, for example—and homeless people are not eligible for inclusion in a household sample. Although the NSDUH makes special efforts to include people living in such facilities as shelters or migrant worker camps, a household survey on drug use excludes many people who do not live in traditional households.

Monitoring the Future

Our second example is different in two respects: (1) It targets a specific population, and (2) it asks sampled respondents a broader variety of questions.

Since 1975, the National Institute on Drug Abuse has sponsored an annual survey of high school seniors, Monitoring the Future: A Continuing Study of the Lifestyles and Values of Youth, or Monitoring the Future (MTF) for short. As its long title implies, the MTF survey is intended to monitor the behaviors, attitudes, and values of young people. In recent years, researchers and policy makers have become especially interested in drug, alcohol, and tobacco use among youths, and the MTF has been used as something of a sentinel to measure such behaviors.

The MTF actually includes several samples of high school students and others groups, totaling about 49,500 respondents in 2004 (Johnston et al., 2005). We’ll begin with the high school sample and then briefly mention the others.

Each spring, between 120 and about 140 high schools are sampled within particular geographic areas. In larger high schools, samples of up to 350 seniors are selected; in smaller schools, all seniors may participate. The 2004 sample yielded responses from about 15,000 high school seniors. The students fill out computer scan sheets containing batteries of questions that include self-reported use of alcohol, tobacco, and illegal drugs. In most cases, students record their answers in classrooms during normal school hours, although in some circumstances students complete the questionnaires in larger groups.

The core sample of the MTF—surveys of high school seniors—thus provides a cross section for measuring annual drug use and other illegal acts. In addition, the MTF has expanded its samples over the years to include public school students in the 8th and 10th grades (since 1991). A subset of about 2400 MTF respondents from the high school samples is selected each year to receive a follow-up questionnaire in the mail. The follow-up samples provide MTF data from college students—those high school seniors who went on to college—and from adults. Data from the oldest follow-up group were first collected in 1976, so, assuming most high school

---

Now recall our discussion of the time dimension in Chapter 4. Each year, both the MTF and the NSDUH measure drug use for a cross section of high school seniors and adults in households, thus providing a snapshot of annual rates of self-reported drug use. Examining annual results from the MTF and the NSDUH over time provides a time series or trend study that enables researchers and policy makers to detect changes in drug use among high school seniors, college students, and adults. Finally, the follow-up samples of MTF respondents constitute a series of panel studies whereby changes in drug use among individual respondents can be studied over time.

Because MTF measures for all samples are collected through self-reports, these data share the potential problems we mentioned in connection with the NSDUH. We noted that at least one study (Gfroerer, 1993) suggests that young people self-report drug use more frequently in school settings than in their homes. However, Mieczkowski (1996:378) points out that students are asked to provide their names to survey staff, something that’s necessary for later follow-ups but that might undermine the truthfulness of their responses.

What about sampling procedures for the MTF? In one sense, selecting schools and then students within schools is a sound procedure for sampling high school seniors. But you should be able to think of at least one problem with this approach. Students who are absent on the day the survey is administered are not included and are not eligible for the follow-up sample. We might reasonably expect that students with poor attendance records are absent more often and thus less represented in the sample. And because our interest in the MTF is as a measure of drug use and offending, we might suspect that students with poor attendance records might have higher rates of drug use and offending.

Validity and Reliability of Self-Report Measures

This chapter cannot supply the final word on efforts to validate self-report measures. On the one hand, the final word has not yet been written; on the other, a small number of researchers have examined the issue in some detail. In most cases, the latter studies compare self-reported offending with other measures—usually, records of offending from law enforcement and juvenile justice agencies. You may recognize this as an example of convergent validity, a topic from Chapter 5.

For example, in their study of London delinquents (mentioned in Chapter 5), Donald West and David Farrington (1977:22) compared official criminal records to self-reported delinquent convictions disclosed in interviews with their subjects. Only a small proportion of subjects failed to mention one or more delinquent acts that appeared in criminal records. After further study, West and Farrington concluded that these omissions were due more to memory lapses than to untruthful responses, because most inconsistencies between the sources occurred for high-rate delinquents. A recent follow-up study of these subjects confirmed the validity of self-reports relative to police measures of adult offending (Kazemian and Farrington, 2005).

In another longitudinal study of a sample of Pittsburgh youths, Farrington and associates (1996) examined convergent validity by again comparing self-reported offending and arrests with official records of arrests and juvenile petitions. Because it was a longitudinal panel study, interviewing the sample at multiple time points, the authors were able to estimate predictive validity by comparing self-reported delinquency at one time period with arrests and juvenile petitions at later times. As you might expect, the relationship is imperfect, but the researchers did find that subjects who self-reported higher levels of more serious offenses at time 1 were much more likely to have official arrest or delinquency records at time 2.
Three additional studies of self-report validity lead us into a summary of this method of measuring crime. Joseph Weis (1986) presents a thorough discussion of the strengths and weaknesses of self-report and official arrest records as measures of offending. Validity and reliability of self-reports are discussed in unusual detail by David Huizinga and Delbert Elliott (1986) as they draw on data from a long-term panel study. Among other things, Huizinga and Elliott make some important comments about the self-interests of researchers in too readily accepting the validity of whatever measure of crime they happen to be using. Finally, Michael Maxfield and associates (Maxfield, Weiler, and Widom, 2000) show that the convergence of self-reported and official records of arrests varies by gender, ethnicity, and type of offense. In a study of the long-term consequences of child abuse, the authors show that female and non-white subjects less often report known arrests. All subjects were more likely to self-report offenses that are more common—most notably, drug use.

Self-Report Surveys Summarized
Researchers and policy makers are best advised to be critical users of measures obtained from self-report surveys. Self-reports can and should be used to measure offending, but researchers, public officials, and others who use such measures should be aware of their strengths and limitations. For example, because MTF and NSDUH sampling and interviewing procedures have remained relatively constant, the surveys provide reasonably consistent information on trends in drug use or offending over time. These two surveys are better as measures of change than as measures of absolute levels.

Our consideration of these two surveys also highlighted the importance of sampling procedures. Samples based on households may not be readily generalized to other populations, and data obtained from high school seniors in class may be different from data obtained from dropouts or chronic truants.

Also consider that alternative measures of offenses like drug use and delinquency are not readily available for general populations. This leaves three alternatives: (1) carefully using imperfect measures, (2) having no measures of certain types of offenses, or (3) developing new measures. Our strong preference is to pursue both the first and the third alternatives. Having read Chapter 5 and worked your way through most of this chapter, you are becoming better equipped to carefully interpret measures of all kinds. Becoming more familiar with self-report methods used in the NSDUH and in self-report surveys routinely conducted in England will help you if the need arises to develop new measures.

Regarding the latter, in 2003 the British Home began a national survey of self-reported offending and drug use. Though results cannot be generalized to the United States, findings from Offending Crime and Justice Survey (OCJS) offer valuable insights into patterns of offending and drug use. OCJS methods also serve as models for obtaining self-report data from general samples (adults) and samples of groups believed at higher risk of offending and drug use (those aged 14–25). Tracey Budd and associates report results from the first two OCJS waves. David Farrington (2003) describes much of the rationale for launching the OCJS as a regular data series.

Drug Surveillance Systems
Surveillance systems have been developed to obtain alternative measures of drug use.

The challenge of developing reliable and valid measures of such offenses as drug use has prompted researchers and policy makers to search for alternative approaches. This has been something of a fitful effort, as various data systems come and go. In this section we briefly consider examples of focused efforts to monitor drug use and its consequences.
Arrestee Drug Abuse Monitoring
The National Institute of Justice (NIJ) conducted programs to measure drug use among samples of persons arrested for several years. The most recent version was the Arrestee Drug Abuse Monitoring (ADAM) program, which last collected data in the year 2003. Even though this program is no longer active, researchers and policymakers continue to use data from ADAM and its predecessor, Drug Use Forecasting (DUF). Because of that, and because ADAM-DUF illustrates some interesting principles of research methods, we briefly describe it here.

Four times per year, cities that participated in ADAM selected samples of persons arrested for a variety of offenses. Anonymous interviews and urine specimens were obtained from those who agreed to take part in the voluntary study. In 2003, more than 22,000 arrestees were interviewed in 39 participating jurisdictions (Zhang, 2004).

The main purpose of ADAM was to provide an ongoing assessment of the prevalence of drug use among persons arrested for criminal offenses. Results for 2003 showed that over 70 percent of adult arrestees tested positive for at least one drug of nine drugs (Zhang, 2004). ADAM results vary somewhat by city, with positive test results ranging from 42 percent of adult males in Woodbury, Iowa, to 86 percent in Chicago.

One of the most interesting aspects of ADAM was that it combined urinalysis and self-report measures of drug use. Interviews with arrestees include self-report items, so responses can be compared with urine test results. Beginning in 2000, a variety of questions were added to obtain information on drug-related policy issues such as perceived dependency and treatment needs, along with respondents’ assessment of local drug markets. Charles Katz and associates used interview data from ADAM to examine the links between gang member and drug use, finding that active gang members used drugs more extensively than did other arrestees (Katz, Webb, and Decker, 2005).

Researchers comparing urinalysis and self-report measures have reached different conclusions about how well self-reports stand up to urinalysis results. Richard Rosenfeld and Scott Decker (1993) concluded that subjects substantially underreport drug use. More recently, Julie Yun Soo Kim and associates (Kim, Fendrich, and Wislar, 2000) found that juveniles who tested positive for marijuana use more often self-reported lifetime marijuana use or use in the past 30 days than in the past 3 days. And juveniles were much more likely to self-report marijuana use than cocaine use. Together, these findings suggest that something called “social desirability” is at work: Using marijuana is seen as more socially desirable than using cocaine, and using either drug in the more distant past is more desirable than having used it in the past 3 days.

Any measure of crime is selective; neither all crimes nor all people are included. In what ways is the ADAM program selective as a measure of drug use? Based on what we have covered so far, three ways should be evident. First, ADAM operated in a relatively small number of cities—39 in 2003. Because participation requires a substantial effort from police departments, we can’t assume that participating cities are representative of other cities. Second, and perhaps most importantly, ADAM included only arrested persons. It’s obvious that we cannot generalize from people who are arrested to the population at large. It’s less obvious that arrest is a selective process. Police exercise discretion in deciding whether to make an arrest, and their priorities can change over time. Third, ADAM interviews and testing were voluntary, although this did not have much impact on participation. Perhaps surprisingly, a large proportion of 1998 arrestees agreed to participate: More than 80 percent consented to the interview, and between 75 and 98 percent provided a urine specimen (National Institute of Justice, 2003:9).

Even though ADAM is limited as a measure of drug use, it represented an imaginative attempt to obtain a different type of measure.
Data from ADAM provided information about gross trends in drug use among arrestees. Additional data collected through the interview can be used to analyze drug use by offense, age, gender, and other variables. Finally, even though ADAM was restricted to samples of persons arrested, it did offer useful information about the co-occurrence of drug use and crime among a group of people at high risk of both. Researchers believe it was discontinued largely because the NIJ has shifted its focus somewhat away from research and toward trying to serve justice more narrowly.

The Drug Abuse Warning Network

The ADAM system implicitly assumes that at least some drug users will be involved in other offenses. Our next specialized measure assumes that drug use can produce acute health problems that cause users to seek treatment in hospital emergency rooms. The Drug Abuse Warning Network (DAWN) collects emergency medical treatment reports for “drug episodes” from samples of hospitals and medical examiners nationwide. In 2004, the DAWN sample included 417 hospitals in 15 major metropolitan areas. Participating hospitals submitted information on almost 280,000 drug abuse episodes for the year 2004 (Substance Abuse and Mental Health Services Administration, 2006).

Drug episodes are defined as visits to a hospital emergency room that are produced by or directly related to use of illegal drugs or nonmedical use of legal drugs. Included under this definition are direct effects of drug ingestion (overdoses, other physical or psychic reactions), as well as injuries or deaths in which drug intoxication was a contributing factor. Nonfatal episodes are drawn from emergency rooms, whereas drug-related deaths are recorded by medical examiners. Notice that DAWN is based on units of analysis that are only indirectly linked to criminal offenses. The concept of drug episodes is most relevant for studies of public health, and, in fact, DAWN was designed as a data system for health surveillance.

Because it was begun in the early 1970s, DAWN is a comparatively long-term time series that monitors the most serious medical consequences of drug use. DAWN is best suited to measuring trends. For example, learning that the participating hospitals in Baltimore mentioned opiates (heroin or morphine) in 1667 drug episodes for 1990 is not especially meaningful. Comparing Baltimore with San Francisco, a city with about the same population in 1990, helps a little. San Francisco had 3954 opiate drug episodes in 1990, which suggests that heroin and morphine were a greater medical problem in that California city. In 1994, however, DAWN recorded 3123 opiate drug episodes for San Francisco and 8882 for Baltimore (McCaig and Greenblatt, 1996). By 2001, both Baltimore and San Francisco were among the top four cities in heroin use.

Like ADAM, DAWN records include demographic and other information about the individuals whose drug use brings them to hospitals and morgues. But the unusual unit of analysis for DAWN means that one individual can account for multiple drug episodes. This makes it difficult to use DAWN data for research in which individual people are the unit of analysis.

Mieczkowski (1996:387) points out that DAWN data for a single metropolitan area might serve as indicators of the impact of anti-drug programs. For example, examining trends in drug episodes involving teenagers might be used in an evaluation of high school education and prevention programs. Or measures of police cocaine seizures might be compared with changes in cocaine drug emergencies.

Measuring Crime for Specific Purposes

Sometimes, alternative measures of crime are collected for research or policy purposes.

Each of the crime measures discussed so far can be used for a variety of research purposes—exploration, description, explanation,
and applied research. However, each has the primary purpose of providing some type of crime count—crimes known to police, victimizations of households and people who live in households, self-reported drug use and other offending, drug use among arrestees and emergency room patients, or qualitative assessments of drug use and availability in specific urban areas. Such measures are useful for criminal justice and public health professionals. Researchers have also made extensive use of data from standard series such as the UCR, SHR, and NCVS.

Often, however, none of these regular series of crime measures meets the needs of researchers, public officials, or others looking for specialized information about crime. At this point, we want to call your attention to examples of crime measures developed for specific research and policy purposes.

Local Crime and Self-Report Surveys
We have seen that victimization surveys have certain advantages over other measures of crime. Surveys are especially useful for learning about crimes not reported to police. Because the NCVS is a national survey, one drawback is that it cannot be used to estimate victimizations for individual cities or for all but the largest states. Because of their potential strengths, victimization surveys are more often being conducted in specific states, cities, or even neighborhoods. Sometimes, local surveys are conducted by researchers; other times, local governments, police departments, and even community organizations conduct them. Cities or counties that have shifted to community policing often do victimization surveys to learn more about crime problems that affect people in particular areas.

Such surveys are rarely up to the standards required by researchers or those that guide such efforts as the NCVS. They usually do not employ probability samples (we’ll discuss these in Chapter 8), and questionnaires can be casual affairs at best. Nevertheless, local crime surveys are useful tools for public officials and citizens in many communities.

A variety of publications have been issued to guide such survey applications. A book by John Eck and Nancy LaVigne (1994) and an excellent handbook by the Bureau of Justice Assistance (1993) are examples of publications prepared for use by law enforcement agencies. As described earlier in this chapter, the BJS and COPS recently collaborated to produce guidelines and software for conducting community victimization surveys (Weisel, 1999). This tool has the potential to improve the quality and expand the use of local crime surveys. A website maintained by the Center for Problem-Oriented Policing (http://www.popcenter.org) includes links to these and other tools that can guide local agencies in developing crime surveys.

Incident-Based Crime Records
We mentioned earlier that larger jurisdictions have been slow to convert from UCR reporting to NIBRS, largely because police departments in large cities have tailored data management systems to their own particular needs. Advances in geographic information systems and the development of computer software to produce maps of crime patterns have spurred the creation of highly sophisticated incident-based crime data systems in some agencies.

Chicago offers one of the best examples among U.S. cities. Under the leadership of researchers at the Illinois Criminal Justice Information Authority and Loyola University of Chicago, detailed information about crimes reported to police can be linked to a variety of other information to reveal space-based patterns of crime (C. Block, 1995; R. Block, 1995). Chicago’s system is powerful, timely (updated daily), and easy to use. Community police officers in the city regularly produce maps of crime patterns for their own use and for neighborhood residents. Researchers have also used these data to examine, for example, the clustering of crime around Chicago mass transit stations (Davis, 1996).

The Newark, New Jersey, Police Department has developed a very detailed incident-based data system for tracking vehicle theft. Among
other things, data from Newark illustrate that most areas of the city “import” stolen cars—autos are stolen from other cities then recovered in Newark. Here’s another example. The average age of all cars stolen in Newark is about 6 years. But that varies substantially by vehicle type. Dodge models less than 3 years old are at great risk, whereas theft rates are higher for 10-year-old Toyotas. Having this kind of detailed information has contributed to theft prevention initiatives in Newark (Maxfield, Fujita, and Ambrosio, 2006).

Cities and other jurisdictions vary in the quality of their records systems, but many police departments are making rapid advances. Local measures of crimes known to police may be more current, complete, and detailed than measures submitted to the FBI under its UCR, SHR, or NIBRS data series.

**Observing Crime**

We have seen that police learn about certain types of crime primarily through observation—for example, drug use and sales, prostitution, public order offenses, and drunk driving. Researchers may face situations in which they need to make independent observations of crime.

For example, we have pointed out that shoplifting is poorly measured by police records. Victim surveys might reveal some instances of shoplifting, but only if we sampled shops, and then we would learn only about those incidents detected by shop staff. Self-report surveys could tell us something about shoplifters, but it would be difficult to use this method to measure shoplifting incidents. Terry Baumer and Dennis Rosenbaum (1982) conducted systematic observations of a large department store in Chicago for descriptive and applied purposes. Using ingenious methods, they sought to estimate the frequency of shoplifting and to evaluate the effectiveness of different store security measures. Chapter 10, on field research, describes this study and its findings in detail.

What about assault? Again, police measure assaults that are reported to them, usually by victims or witnesses. The NCVS was redesigned in part to get better counts of assault, but for some research purposes, independent observations may yield better measures.

Ross Homel and associates (Homel and Clark, 1994; Macintyre and Homel, 1997) were interested in the links between drinking and violence in bars and nightclubs. Specifically, they sought to learn what sorts of physical features and situations in bars tended to either discourage or facilitate violence. Their explanatory studies required selecting samples of public drinking establishments in Sydney, Australia, and dispatching pairs of observers to each site. Observers recorded information about violent incidents they witnessed, together with details about each bar’s physical layout, entertainment, and procedures for controlling access and regulating conduct. Researchers were able to make general explanatory statements about how the physical environment and management practices were related to violence.

Police normally learn about drug use or sales only if they witness the acts. Constraints on the ability of police to make such observations are often cited as obstacles to the enforcement of drug laws. In many communities, however, neighborhood organizations supplement law enforcement through patrols and targeted surveillance of drug activity in specific neighborhoods. If such citizen patrols witness suspected drug activity, they inform police, community prosecutors, or other justice officials who may launch further investigations. Neighborhood organizations in Kansas City, Albuquerque, Philadelphia, and many other cities use directed observations in this way (Maxfield, 2001).

These examples of directed observation have three common characteristics. First, each has a fairly specific research or policy purpose. Baumer and Rosenbaum wanted to obtain estimates of shoplifting frequency and to evaluate the effectiveness of certain security measures; Homel and associates wished to learn more about the association between public drinking and violence; members of community
organizations participate in neighborhood surveillance to take action against drug and other problems. Second, the three examples focus on relatively small areas—a single department store, a sample of bars and nightclubs, or a specific neighborhood. Finally, the expected density of incidents made observation an appropriate way to measure crime. Shoplifting happens in shops, and large department stores offer opportunities to observe many incidents; it’s well known that public violence is relatively common in bars, and police records can be used to select bars with a history of violence; finally, citizen patrols and neighborhood surveillance programs are launched because residents are troubled by local drug problems.

Measuring Crime Summary

Each measure of crime has strengths and weaknesses.

Table 6.5 summarizes some of what we have considered in this chapter by comparing different measures of crime. Each method has its

PUTTING IT ALL TOGETHER

MEASURING TRAFFIC VIOLATIONS

Background
The frequency and distribution of certain kinds of offenses underlie many research questions about racial profiling. In the first place, many people trace racial profiling to the use of traffic enforcement to interdict drug and weapons smuggling on the nation’s highways. Operation Pipeline, established by the Drug Enforcement Administration in 1986, trained state and local law enforcement officers to recognize features of vehicles that signal possible drug trafficking activity (General Accounting Office, 2000; Verniero and Zoubek, 1999). In many areas it was alleged that race was routinely used as a criterion to make “pretextual” traffic stops. Police were said to use traffic violations as a pretext to stop drivers they suspected of drug trafficking, and this disproportionately targeted nonwhite drivers.

Responses by police and other government agencies varied, ranging from outright denial that race influenced traffic stops, to promises that any racial bias in traffic enforcement would not be tolerated. Whereas some acknowledged interpreting Operation Pipeline as an invitation to target black drivers, many police agencies claimed that drivers were stopped only because of traffic violations, not race.

Traffic Violations
This raised the question of how to measure the actual frequency of traffic violations. You will recognize this as a challenging task. Slightly restating two questions presented at the very beginning of this chapter: How do you measure traffic violations? How many traffic violations are there? Just as it is virtually impossible to measure all crimes, it is not feasible to measure all traffic violations for two related reasons: Virtually everyone commits some sort of violation at least occasionally; because of this, police have to be selective in deciding which vehicles to stop. Relatively little is known about how police decide which traffic violations to target. Andresen (2005) describes how different troopers in New Jersey tend to concentrate on different kinds of violations.

Speeding
Researchers have focused on speeding in an effort to measure the distribution of traffic violation rates. Engel et al. (2004:10) sum up reasons for this in their Pennsylvania study. First, recent national surveys of the general population show that among those who self-report being stopped by police, speeding is the most common reason—55 percent in the NCVS, 65 percent in a U.S.
Department of Transportation survey (Durose et al., 2005; Royal, 2003). Second, they found that over 75 percent of stops reported by Pennsylvania State Police were for speeding infractions. Finally, speeding can be measured more easily and reliably than other kinds of driving violations. Just as police can use radar or laser detectors to reliably measure speed, researchers can similarly obtain measures that are accurate and consistent. The Pennsylvania research team used radar equipment issued by state police and measured the speed of cars passing sampled locations. New Jersey researchers placed radar detectors inside unmarked vans together with the digital cameras mentioned in the previous chapter (Lange et al., 2005). North Carolina combined speed metering with the mobile observation of race mentioned in the last chapter (Smith et al., 2003). A van was driven at the posted speed limit, researchers used stop watches to time vehicles passing them, then computed the passing vehicles’ speeds.

Of course, police records are also sources of information about speeding and other traffic violations. In all cases researchers examined police records in addition to some other source of data. As we will see in the next chapter, comparing police records to other measures was an important part of efforts to explain race disparity in traffic stops.

The survey data mentioned above can be viewed as something of a self-report survey of offenders. If you think about the rationale for self-report surveys discussed elsewhere in this chapter, you should be able to see why this method is useful to estimate traffic violation rates. Like drug use and delinquency, traffic violations are not likely to be reported to police. Because speeding and traffic violations are so common, it’s not possible for police to detect them all, and the ability of people to evade detection by police may vary. Just as police might not make records of all crimes that come to their attention, their records may not reflect all traffic stops. So we ask people about traffic violating behavior and about their experiences in being stopped by police. Among other things, these measures can be compared to measures from other sources.

Smith et al. (2003) included self-report items in different surveys of North Carolina residents as part of their effort to assess drivers’ perceptions of police and experience of being stopped. The researchers also incorporated some clever devices to assess the quality of self-report items. In addition to the general population, they selected some respondents who, according to state police records, had received traffic citations in the previous year. Administering self-report questionnaires to “known offenders” offered a way to estimate the reliability of the survey.

So, research on racial profiling used speeding as a measure of traffic violation, and measured that in a variety of ways, just as we have seen in this chapter that a variety of techniques are used to measure different kinds of crime.

own strengths and weaknesses. The UCR and SHR provide the best counts for murder and crimes in which the victim is either a business or a commercial establishment. Crimes against persons or households that are not reported to police are best counted by the NCVS. Usually, these are less serious crimes, many of them UCR Part II incidents that are counted only if a suspect is arrested. Recent changes in NCVS procedures have increased the counts of sexual assault and other violent victimizations. Compared with the UCR, NIBRS potentially provides much greater detail for a broader range of offenses. NIBRS complements the NCVS by including disaggregated incident-based reports for state and local areas and by recording detailed information on crimes against children under age 12.

Self-report surveys are best at measuring crimes that do not have readily identifiable victims and that are less often observed by or reported to police. The two self-report surveys listed in Table 6.5 sample different populations and use different interview procedures.
Sentinel measures target more narrowly defined populations for the purposes of monitoring and are best seen as measures of change. The two series in Table 6.5 measure drug use in the context of medical and legal crises.

Be aware that Table 6.5 and our discussions throughout this chapter include only a partial listing of crime measures used by researchers and public officials. Each local and state law enforcement agency maintains its own records, which often provide the best measures for research in specific geographic areas. Finally, a specific research purpose may require collecting independent measures of crime.

Don’t forget that all crime measures are selective, so it’s important to understand the selection process. Despite their various flaws, the measures of crime available to you can serve many research purposes. Researchers are best advised to be critical and careful users of whatever measure of crime best suits their research purpose. This advice is illustrated in this chapter’s running example, where researchers tackle the problem of measuring traffic violations.

**Main Points**

- Crime is a fundamental concept in criminal justice research. Different approaches to measuring crime illustrate general principles of conceptualization, operationalization, and measurement.
- Before using any measure of crime, researchers should understand what types of offenses it does and does not include.
• Different measures of crime are based on different units of analysis. The UCR is a summary measure that reports totals for individual agencies. Other measures use offenders, victims, incidents, or offenses as the units of analysis.
• Crime data are collected for one or more general purposes: monitoring, agency accountability, and research.
• Crimes known to police have been the most widely used measures. UCR data have been available since the early twentieth century; more detailed information about homicides was added to the UCR in 1961. Most recently, the FBI has developed an incident-based reporting system that is gradually being adopted.
• Surveys of victims reveal information about crimes that are not reported to police. The NCVS includes very detailed information about personal and household incidents but does not count crimes against businesses or individual victims under age 12. Although the NCVS is a nationally representative measure, it cannot estimate victimizations for states or local areas.
• Self-report surveys were developed to measure crimes with unclear victims that are less often detected by police. Two such surveys estimate drug use among high school seniors and adults. Self-report surveys do not measure all drug use because of incomplete reporting by respondents and procedures for selecting survey respondents.
• ADAM and DAWN provide measures of drug use among special populations but are best suited to monitoring changes in drug use.
• Different measures of crime are also developed for specific research and policy purposes. Many police departments do crime analysis with their own incident-based records. The NVAWS measures specific types of crime for specific purposes.
• We have many different measures of crime because each measure is imperfect. Each measure has its own strengths and weaknesses.

🌟 **Review Questions and Exercises**

1. Los Angeles police consider a murder to be gang related if either the victim or the offender is known to be a gang member, whereas Chicago police record a murder as gang related only if the killing is directly related to gang activities (Spergel, 1990). Describe how these different operational definitions illustrate general points about measuring crime discussed in this chapter.
2. Measuring gang-related crime is an example of trying to measure a particular dimension of crime: motive. Other examples are hate crimes, terrorist acts, and drug-related crimes. Specify conceptual and operational definitions for at least one of these types. Find one newspaper story and one research report that presents an example.
3. How would you measure crime in a specific city if you wanted to evaluate a community policing program that encourages neighborhood residents to report incidents to police?

🌟 **Online Sessions for Chapter 6**

Online Sessions present further examples of selected material covered in each chapter. Your instructor may choose to use these for in-class discussion (as I do with my students) or you can access them at home. To ensure that you always have access to live, correct links for the websites described here, we provide the necessary links on the companion website for *Research Methods for Criminal Justice and Criminology* at [http://www.thomsonedu.com/criminaljustice/maxfield](http://www.thomsonedu.com/criminaljustice/maxfield). Once at the companion website, select this specific chapter, click on “Chapter Resources,” then click on “Web Links.”

**Major Series of Crime Data** Published reports and actual data from various measures of crime we have discussed are increasingly available on the web. Visit websites for one of the following: Monitoring the Future (MTF), National Survey on Drug Use and Health (NSDUH), New York City Police Department, Drug Abuse Warning Network (DAWN), National Crime Victimization Survey (NCVS), or Uniform Crime Reports (UCR). Using one of these links, download data on at least two offenses (or, in the case of the MTF and the NSDUH, data on drug use for two specific drugs).
in two cities over a 5-year period. Bring your data to class and discuss what you have found.

MTF:  http://www.monitoringthefuture.org/new.html
NSDUH:  http://www.drugabusestatistics.samhsa.gov/nsduh.htm
DAWN: http://dawninfo.samhsa.gov/
NCVS: http://www.ojp.usdoj.gov/bjs/cvictgen.htm
UCR:  http://www.fbi.gov/ucr/ucr.htm

30 Years of Violent Crime Data  In this chapter and Chapter 5, we pointed out that different ways of operationalizing concepts can produce different results. The Bureau of Justice Statistics website presents another example of this in a graph displaying trends for four measures of violent crime over a 30-year period. Visit the BJS website, and review the graph and accompanying explanations. Then discuss how the four series of data are different and how they are similar. What do you think is the most important single factor explaining why the measures are different?  http://www.ojp.usdoj.gov/bjs/glance/cv2.htm

Help with NIBRS The Justice Research and Statistics Association (JRSA) assists state-level statistical analysis centers in using a variety of criminal justice data and in conducting evaluations. The JRSA has established a web page to help states shift to NIBRS. Visit the JRSA site, and then find a web page that presents state-level NIBRS data. Bring the results to class.  http://www.jrsa.org/ibrrc/background-status/state_projects.shtml

Why Federal Agencies Collect Data The National Research Council (2005) has published a report on principles for federal agencies that compile and report statistical series. It includes the Bureau of Justice Statistics, but also covers many more agencies. This reflects counting crime for monitoring and agency accountability purposes, as discussed in this chapter.  http://newton.nap.edu/catalog/11252.html

Drugs: A View from the White House Reflecting the difficulty of measuring drug crimes, the White House Office of National Drug Control Policy has compiled a variety of information on different dimensions of drug problems. Although the website includes quite a lot of self-promoting hype, some useful resources can be found. Among these are links to what’s known about the scope of use and circumstances of use for different drugs. There’s also a link to other federal sources of information about drug use and policies to combat illegal drug use. This is somewhat dated. For example, it includes a link to the ADAM program, for which data were last collected in the year 2003. Select two drugs and review the fact sheets, and facts and figures reports for each.  http://www.whitehousedrugpolicy.gov/drugfact/index.html

Additional Readings

Farrington, David P., Langan, Patrick A., and Tonry, Michael, eds.  Cross-National Studies in Crime and Justice  (Washington, DC: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, 2004) (http://www.ojp.usdoj.gov/bjs/abstract/cnscj.htm). This is an ambitious and unique attempt to compare crime measures for eight nations: Australia, Canada, England, the Netherlands, Scotland, Sweden, Switzerland, and the United States. Each country had conducted recent victim surveys, so victimization data could be compared to data from police records. The introductory essay summarizes this study, then individual chapters are presented for each country.

Fendrich, Michael, Johnson, Timothy P., Sudman, Seymour, Wislar, Joseph S., and Spiehler, Vina, “Validity of Drug Use Reporting in a High-Risk Community Sample: Comparison of Cocaine and Heroin Survey Reports with Hair Tests,”  American Journal of Epidemiology, Vol. 149, no. 10 (1999), pp. 955–62. Tackling the question of how valid self-reports are, the authors compare survey measures with laboratory tests of drug use from hair samples. Among other things, tests of hair samples can show drug use over time, making it possible to validate self-reported use over a period of several months.

Maxfield, Michael G., “The National Incident-Based Reporting System: Research and Policy Implications,”  Journal of Quantitative Criminology, Vol. 15, no. 2 (1999), pp. 119–49. This article describes NIBRS in detail and compares this source of crime data with the UCR and the NCVS. See also other articles in this issue.
for the first published examples of NIBRS data being used in research and policy analysis.

National Research Council, *Informing America’s Policy on Illegal Drugs: What We Don’t Know Keeps Hurting Us* (Washington, DC: National Academy Press, 2001). The title of this volume underscores the need for better measures of drug use. This report describes and compares four sources of information on drug use: ADAM, DAWN, the MTF, and the NSDUH.

Chapter 7

Experimental and Quasi-Experimental Designs

We’ll learn about the experimental approach to social science research. We’ll consider a wide variety of experimental and other designs available to criminal justice researchers.

Introduction 177

The Classical Experiment 178

Independent and Dependent Variables 178

Pretesting and Posttesting 179

Experimental and Control Groups 179

Double-Blind Experiments 181

Selecting Subjects 182

Randomization 182

Experiments and Causal Inference 183

Experiments and Threats to Validity 183

Threats to Internal Validity 184

Ruling Out Threats to Internal Validity 187

Generalizability and Threats to Validity 189

Threats to Construct Validity 189

Threats to External Validity 190

Threats to Statistical Conclusion Validity 191

Variations in the Classical Experimental Design 191

Quasi-Experimental Designs 193

Nonequivalent-Groups Designs 193

Cohort Designs 196

Time-Series Designs 197

(continued)
Introduction

Experimentation is an approach to research best suited for explanation and evaluation.

Research design in the most general sense involves devising a strategy for finding out something. We'll first discuss the experiment as a mode of scientific observation in criminal justice research. Abraham Kaplan (1964:144) described experimentation as “a process of observation, to be carried out in a situation expressly brought about for that purpose.” At base, experiments involve (1) taking action and (2) observing the consequences of that action. Social science researchers typically select a group of subjects, do something to them, and observe the effect of what was done.

It is worth noting at the outset that experiments are often used in nonscientific human inquiry as well. We experiment copiously in our attempts to develop a more generalized understanding about the world we live in. We learn many skills through experimentation: riding a bicycle, driving a car, swimming, and so forth. Students discover how much studying is required for academic success through experimentation. Professors learn how much preparation is required for successful lectures through experimentation.

Returning to social science applications, experiments are especially well suited to research projects that involve relatively well-defined concepts and propositions. A further requirement is the ability to control the conditions under which research is conducted. The traditional model of science, discussed in Chapter 2, and the experimental model are closely related.

Experimentation is best suited for hypothesis testing in explanatory or evaluation studies. Suppose we are interested in studying alcohol abuse among college students and in discovering ways to reduce it. We might hypothesize that acquiring an understanding about the health consequences of binge drinking and long-term alcohol use will have the effect of reducing alcohol abuse. We can test this hypothesis experimentally. To begin, we might ask a group of experimental subjects how much beer, wine, or spirits they drank on the previous day and how frequently, in an average week, they consume alcohol for the specific purpose of getting drunk. Next, we might show these subjects a video depicting the various physiological effects of chronic drinking and binge drinking. Finally—say, 1 month later—we might again ask the subjects about their use of alcohol in the previous week to determine whether watching the video actually reduced alcohol use.

Because experiments are best suited for hypothesis testing, they may also be appropriate in the study of criminal justice policy. In Chapter 2, we discussed the logical similarity between hypotheses and criminal justice policies, noting that evaluation research is conceptually
equivalent to hypothesis testing. The experimental model therefore can be a useful design for evaluating criminal justice policy.

You might typically think of experiments as being conducted in laboratories under carefully controlled conditions. Although this may be true in the natural sciences, few social scientific experiments take place in laboratory settings. The most notable exception to this occurs in the discipline of psychology, in which laboratory experiments are common. Criminal justice experiments are almost always conducted in field settings outside the laboratory.

The Classical Experiment

Variables, time order, measures, and groups are the central features of the classical experiment.

Like much of the vocabulary of research, the word experiment has acquired both a general and a specialized meaning. So far, we have referred to the general meaning, defined by David Farrington, Lloyd Ohlin, and James Q. Wilson (1986:65) as “a systematic attempt to test a causal hypothesis about the effect of variations in one factor (the independent variable) on another (the dependent variable). . . . The defining feature of an experiment lies in the control of the independent variable by the experimenter.” In a narrower sense, the term experiment refers to a specific way of structuring research, usually called the classical experiment. In this section, we examine the requirements and components of the classical experiment. Later in the chapter, we will consider designs that can be used when some of the requirements for classical experiments cannot be met.

The most conventional type of experiment, in the natural and the social sciences, involves three major pairs of components: (1) independent and dependent variables, (2) pretesting and posttesting, and (3) experimental and control groups. We will now consider each of those components and the way they are put together in the execution of an experiment.

Independent and Dependent Variables

Essentially, an experiment examines the effect of an independent variable on a dependent variable. Typically, the independent variable takes the form of an experimental stimulus that is either present or absent—that is, a dichotomous variable, having two attributes. That need not be the case, however, as subsequent sections of this chapter will show. In the example concerning alcohol abuse, how often subjects used alcohol is the dependent variable, and exposure to a video about alcohol’s effects is the independent variable. The researcher’s hypothesis suggests that levels of alcohol use depend, in part, on understanding its physiological and health effects. The purpose of the experiment is to test the validity of this hypothesis.

The independent and dependent variables appropriate to experimentation are nearly limitless. It should be noted, moreover, that a given variable might serve as an independent variable in one experiment and as a dependent variable in another. For example, alcohol abuse is the dependent variable in our example, but it might be the independent variable in an experiment that examines the effects of alcohol abuse on academic performance.

In the terms of our discussion of cause and effect in Chapter 4, the independent variable is the cause and the dependent variable is the effect. Thus, we might say that watching the video causes a change in alcohol use or that reduced alcohol use is an effect of watching the video.

It is essential that both independent and dependent variables be operationally defined for the purposes of experimentation. Such operational definitions might involve a variety of observation methods. Responses to a questionnaire, for example, might be the basis for defining self-reported alcohol use on the previous day. Alternatively, alcohol use by subjects could be measured with Breathalyzer or blood alcohol tests.
Conventionally, in the experimental model, the dependent and independent variables are operationally defined before the experiment begins. However, as we will see in connection with survey research and other methods, it is sometimes appropriate to first make a variety of observations during data collection and then determine the most useful operational definitions of variables during later analyses. Ultimately, however, experimentation requires specific and standardized measurements and observations.

**Pretesting and Posttesting**

In the simplest experimental design, subjects are measured on a dependent variable (pretested), exposed to a stimulus that represents an independent variable, and then remeasured on the dependent variable (posttested). Differences noted between the first and second measurements on the dependent variable are then attributed to the influence of the independent variable.

In our example of alcohol use, we might begin by pretesting the extent of alcohol use among our experimental subjects. Using a questionnaire, we measure the extent of alcohol use reported by each individual and the average level of alcohol use for the whole group. After showing subjects the video on the effects of alcohol, we administer the same questionnaire again. Responses given in this posttest permit us to measure the subsequent extent of alcohol use by each subject and the average level of alcohol use of the group as a whole. If we discover a lower level of alcohol use on the second administration of the questionnaire, the subjects may have been unaware of its purpose. By the time of the second measurement, however, they may have figured out the purpose of the experiment, become sensitized to the questions about drinking, and changed their answers. Thus, the video might seem to have reduced alcohol abuse although, in fact, it did not.

This is an example of a more general problem that plagues many forms of criminal justice research: The very act of studying something may change it. Techniques for dealing with this problem in the context of experimentation are covered throughout the chapter.

**Experimental and Control Groups**

The traditional way to offset the effects of the experiment itself is to use a control group. Social scientific experiments seldom involve only the observation of an experimental group to which a stimulus has been administered. Researchers also observe a control group to which the experimental stimulus has not been administered.

In our example of alcohol abuse, two groups of subjects are examined. To begin, each group is administered a questionnaire designed to measure their alcohol use in general and binge drinking in particular. Then only one of the groups—the experimental group—is shown the video. Later, the researcher administers a posttest of alcohol use to both groups. Figure 7.1 illustrates this basic experimental design.

Using a control group allows the researcher to control for the effects of the experiment itself. If participation in the experiment leads the subjects to report less alcohol use, that should occur in both the experimental and the control groups. If, on the one hand, the overall level of drinking exhibited by the control group decreases between the pretest and posttest as much as for the experimental group, then the apparent reduction in alcohol use must be a function of the experiment or of some external factor, not a function of watching the video specifically. In this situation, we can conclude
that the video did not cause any change in alcohol use.

If, on the other hand, drinking decreases only in the experimental group, then we can be more confident in saying that the reduction is a consequence of exposure to the video (because that’s the only difference between the two groups). Or, alternatively, if drinking decreases more in the experimental group than in the control group, then that, too, is grounds for assuming that watching the video reduced alcohol use.

The need for control groups in social scientific research became clear in a series of employee satisfaction studies conducted by Fritz J. Roethlisberger and William J. Dickson (1939) in the late 1920s and early 1930s. They studied working conditions in the telephone “bank wiring room” of the Western Electric Works in Chicago, attempting to discover what changes in working conditions would improve employee satisfaction and productivity.

To the researchers’ great satisfaction, they discovered that making working conditions better consistently increased satisfaction and productivity. When the workroom was brightened by better lighting, for example, productivity went up again. To substantiate their scientific conclusion, the researchers then dimmed the lights: Productivity again improved!

It became evident then that the wiring room workers were responding more to the attention given them by the researchers than to improved working conditions. As a result of this phenomenon, often called the “Hawthorne effect,” social researchers have become more sensitive to and cautious about the possible effects of experiments themselves. The use of a proper control group—studied intensively, with the working conditions otherwise unchanged—would have revealed this effect in the wiring room study.

The need for control groups in experimentation has been most evident in medical research. Time and again, patients who participated in medical experiments appeared to improve, but it was unclear how much of the improvement came from the experimental treatment and how much from the experiment. Now, in testing the effects of new drugs, medical researchers frequently administer a placebo (for example, sugar pills) to a control group. Thus, the control-group patients believe that they, like members of the experimental group, are re-

---

**Figure 7.1** Basic Experimental Design

![Basic Experimental Design](image-url)
ceiving an experimental drug—and they often improve! If the new drug is effective, however, those who receive that drug will improve more than those who received the placebo.

In criminal justice experiments, control groups are important as a guard against the effects not only of the experiments themselves but also of events that may occur outside the laboratory during the course of experiments. Suppose the alcohol use experiment was being conducted on your campus, and at that time a popular athlete was hospitalized for acute alcohol poisoning after he and a chum drank a bottle of rum. This event might shock the experimental subjects and thereby decrease their reported drinking. Because such an effect should happen about equally for members of the control and experimental groups, lower levels of reported alcohol use in the experimental group than in the control group would again demonstrate the impact of the experimental stimulus: watching the video that describes the health effects of alcohol abuse.

Sometimes, an experimental design requires more than one experimental or control group. In the case of the alcohol video, for example, we might also want to examine the impact of participating in group discussions about why college students drink alcohol, with the intent of demonstrating that peer pressure may promote drinking by people who would otherwise abstain. We might design our experiment around three groups. One group would see the video and participate in the group discussions, another would only see the video, and still another would only participate in group discussions; the control group would do neither. With this kind of design, we could determine the impact of each stimulus separately, as well as their combined effect.

**Double-Blind Experiments**

As we saw in medical experimentation, patients sometimes improve when they think they are receiving a new drug; thus it is often necessary to administer a placebo to a control group.

Sometimes, experimenters have this same tendency to prejudge results. In medical research, the experimenters may be more likely to “observe” improvements among patients who receive the experimental drug than among those receiving the placebo. That would be most likely, perhaps, for the researcher who developed the drug. A double-blind experiment eliminates this possibility because neither the subjects nor the experimenters know which is the experimental group and which is the control. In medical experiments, those researchers who are responsible for administering the drug and for noting improvements are not told which subjects receive the drug and which receive the placebo. Thus, both researchers and subjects are “blind” with respect to who is receiving the experimental drug and who is getting the placebo. Another researcher knows which subjects are in which group, but that person is not responsible for administering the experiment.

An example of a double-blind experiment is the study of the Domestic Violence Reduction Unit in Portland, Oregon, by Annette Jolin and associates (Jolin, Feyerhorn, Fountain, and Friedman, 1998). The researchers evaluated an intervention designed to reduce repeat offenses against victims of domestic violence by instructing them in strategies for “keeping safe” and putting more resources into the investigation of domestic violence incidents.

In the experimental evaluation, after misdemeanor offenses for which male offenders had been arrested and taken to jail, their female victims were randomly assigned to a treatment group that received supplementary services or to a control group that did not receive the intervention. Cases in the treatment group were more intensively investigated, including more thorough efforts to collect evidence. Victims in the treatment group received additional services, including assistance in developing safety plans, information on how to access a variety of services, and transportation to service providers when necessary. Staff randomly assigned
cases to the treatment or control groups, but assignment was coded so the staff were not aware which victims were to receive supplementary services and which were not. Likewise, victims were not informed whether they were in the treatment or the control group. This double-blind experiment was therefore able to test the effects of supplementary domestic violence services without being contaminated by the expectations of victims or researchers.

In social scientific experiments, as in medical experiments, the danger of experimenter bias is further reduced to the extent that the operational definitions of the dependent variables are clear and precise. Thus, in the domestic violence experiment, researchers were less likely to unconsciously bias their interpretation of reports of repeat domestic violence than they were to unconsciously bias their subjective assessment of how satisfied victims were with services provided to them.

**Selecting Subjects**

Before beginning an experiment, we must make two basic decisions about who will participate. First, we must decide on the target population—the group to which the results of our experiment will apply. If our experiment is designed to determine, for example, whether restitution is more effective than probation in reducing recidivism, our target population is some group of persons convicted of crimes. In our hypothetical experiment about the effects of watching a video on the health consequences of alcohol abuse, the target population might be college students.

Second, we must decide how particular members of the target population will be selected for the experiment. Ideally, the control group represents what the experimental group would have been like if it had not been exposed to the experimental stimulus. It is important, therefore, that the experimental and control groups be as similar as possible.

**Randomization**

Having recruited a group of subjects, under the classic experiment we would randomly assign those subjects to either the experimental or the control group. This might be accomplished by numbering all the subjects serially and selecting numbers by means of a random-number table. Or we might assign the odd-numbered subjects to the experimental group and the even-numbered subjects to the control group.

Randomization is a central feature of the classical experiment. The most important characteristic of randomization is that it produces experimental and control groups that are statistically equivalent. Put another way, randomization reduces possible sources of systematic bias in assigning subjects to groups. The basic principle is simple: If subjects are assigned to experimental and control groups through a random process such as flipping a coin, the assignment process is said to be unbiased, and the resultant groups are equivalent.

Although the rationale underlying this principle is a bit complex, understanding how randomization produces equivalent groups is a key point. Farrington and associates (Farrington, Ohlin, and Wilson, 1986:66) compare randomization in criminal justice research to laboratory controls in the natural sciences:

The control of extraneous variables by randomization is similar to the control of extraneous variables in the physical sciences by holding physical conditions (e.g., temperature, pressure) constant. Randomization insures that the average unit in [the] treatment group is approximately equivalent to the average unit in another [group] before the treatment is applied.
You’ve surely heard the expression “All other things being equal.” Randomization makes it possible to assume that all other things are equal.

Experiments and Causal Inference

Experiments potentially control for many threats to the validity of causal inference, but researchers must remain aware of these threats.

The central features of the classical experiment are independent and dependent variables, pretesting and posttesting, and experimental and control groups created through random assignment. Think of these features as building blocks of a research design to demonstrate a cause-and-effect relationship. This point will become clearer by comparing the criteria for causality, discussed in Chapter 4, to the features of the classical experiment, as shown in Figure 7.2.

The experimental design ensures that the cause precedes the effect in time by taking posttest measurements of the dependent variable after introducing the experimental stimulus. The second criterion for causation—an empirical correlation between the cause-and-effect variables—is determined by comparing the pretest (in which the experimental stimulus is not present) to the posttest for the experimental group (after the experimental stimulus is administered). A change from pretest to posttest measures demonstrates correlation.

The final requirement is to show that the observed correlation between cause and effect is not due to the influence of some third variable. The classical experiment makes it possible to satisfy this criterion for cause in two ways. First, the posttest measures for the experimental group (stimulus present) are compared with those for the control group (stimulus not present). If the observed correlation between the stimulus and the dependent variable is due to some other factor, then the two posttest scores will be similar. Second, random assignment produces experimental and control groups that are equivalent and will not differ on some other variable that could account for the empirical correlation between cause and effect.

Experiments and Threats to Validity

The classical experiment is designed to satisfy the three requirements for demonstrating cause-and-effect relationships. But what about threats to the validity of causal inference discussed in Chapter 4? In this section, we consider each of those threats in more detail and describe how the classical experiment reduces many of them. The book by Donald Campbell and Julian Stanley (1966) is the most frequently cited authority on threats to validity. For a fuller discussion, on which we draw heavily, see the book by William Shadish, Thomas Cook, and Donald Campbell (2002). We present these threats in a slightly different order, beginning with threats to internal validity.
Threats to Internal Validity

The problem of threats to internal validity refers to the possibility that conclusions drawn from experimental results may not accurately reflect what went on in the experiment itself. As we stated in Chapter 4, conclusions about cause and effect may be biased in some systematic way. Shadish, Cook, and Campbell (2002:54–60) pointed to several sources of the problem.

As you read about these different threats to internal validity, keep in mind that each is an example of a simple point: possible ways that researchers might be wrong in inferring causation. Although these threats are well known among researchers and often cited by students as things to memorize, they are simply examples of ways we might be wrong in causal inference.

**History**

Historical events may occur during the course of the experiment that confound the experimental results. The hospitalization of a popular athlete for acute alcohol poisoning during an experiment on reducing alcohol use is an example.

**Maturation**

People are continually growing and changing, whether in an experiment or not, and those changes affect the results of the experiment. In a long-term experiment, the fact that the subjects grow older (and wiser?) may have an effect. In shorter experiments, they may become tired, sleepy, bored, or hungry, or change in other ways that affect their behavior in the experiment. A long-term study of alcohol abuse might reveal a decline in binge drinking as the subjects mature.

History and maturation are similar in that they represent a correlation between cause and effect that is due to something other than the independent variable. They’re different in that history represents something that’s outside the experiment altogether, whereas maturation refers to change within the subjects themselves.

**Testing**

Often, the process of testing and retesting influences people’s behavior and thereby confounds the experimental results. Suppose we administer a questionnaire to a group as a way of measuring their alcohol use. Then we administer an experimental stimulus and remeasure their alcohol use. By the time we conduct the posttest, the subjects may have become more sensitive to the issue of alcohol use and so may provide different answers. In fact, they may have figured out that we are trying to determine whether they drink too much. Because excessive drinking is frowned on by university authorities, our subjects will be on their best behavior and give answers that they think we want or that will make them look good.

**Instrumentation**

Thus far, we haven’t said much about the process of measurement in pretesting and posttesting, and it’s appropriate to keep in mind the problems of conceptualization and operationalization discussed in Chapter 5. If we use different measures of the dependent variable (say, different questionnaires about alcohol use), how can we be sure that they are comparable? Perhaps alcohol use seems to have decreased simply because the pretest measure was more sensitive than the posttest measure was. Or if the measurements are being made by the experimenters, their procedures may change over the course of the experiment. That would be a problem of reliability.

Instrumentation is always a potential problem in criminal justice research that uses secondary sources of information such as police records about crime or court records about probation violations. There may be changes in how probation violations are defined or changes in the record-keeping practices of police departments.

The differences between testing and instrumentation threats to internal validity may seem unclear. In general, testing refers to changes in how subjects respond to measurement, whereas instrumentation is concerned with changes in the measurement process itself. If police officers respond differently to pretest and posttest questionnaires about prejudice, for example, that is a testing problem. However, if different
questionnaires about prejudice are used in pre-
test and posttest measurements, instrumenta-
tion is a potential threat.

Statistical Regression Sometimes, it’s ap-
propriate to conduct experiments on subjects
who start out with extreme scores on the de-
pendent variable. Many sentencing policies,
for example, target chronic offenders. Charles
Murray and L. A. Cox (1979) examined a pro-
gram to incarcerate high-rate male juvenile
offenders. They found that rearrest rates for
subjects who had served sentences averaging
11 months were substantially lower than those
for other offenders who were not incarcerated.
These findings were questioned by Michael
Maltz and associates (Maltz, Gordon, McDow-
all, and McCleary, 1980), who pointed out that
the number of crimes committed by any given
offender fluctuates over time. They argued that
subjects in the experimental group—chronic
offenders—were jailed following a period when
their offense rates were abnormally high and
that the decline in posttest arrests simply re-
lected a natural return to less extreme rates
of offending. Even without any experimental
stimulus, then, the group as a whole was likely
to show some improvement over time.

Commonly referred to as “regression to
the mean,” this threat to validity can emerge
whenever researchers are interested in cases
that have extreme scores on some variable. As
a simple example, statisticians often point out
that extremely tall people as a group are likely
to have children shorter than themselves, and
extremely short people as a group are likely to
have children taller than themselves. The dan-
ger, then, is that changes occurring by virtue of
subjects starting out in extreme positions will
be attributed erroneously to the effects of the
experimental stimulus.

Selection Biases Randomization eliminates
the potential for systematic bias in selecting
subjects. But subjects may be chosen in other
ways that threaten validity. Volunteers are of-
ten solicited for experiments conducted on
college campuses. Students who volunteer for
an experiment may not be typical of students
as a whole, however. Volunteers may be more
interested in the subject of the experiment and
more likely to respond to a stimulus. Or, if ex-
perimental subjects are paid some fee, students
in greater financial need may participate, al-
though they may not be representative of other
students.

A common type of selection bias in ap-
plied criminal justice studies results from the
natural caution of public officials. Let’s say
you are a bail commissioner in a large city,
and the mayor wants to try a new program to
increase the number of arrested persons who
are released on bail. The mayor asks you to de-
termine what kinds of defendants should be
eligible for release and informs you that staff
from the city’s criminal justice services agency
will be evaluating the program. In establishing
eligibility criteria, you will probably try to se-
lect defendants who will not be arrested again
while on bail and defendants who will most
likely show up for scheduled court appear-
ances. In other words, you will try to select par-
ticipants who are least likely to fail. This com-
mon and understandable caution is sometimes
referred to as “creaming”—skimming the best
risks off the top. Creaming is a threat to valid-
ity because the low-risk persons selected for re-
lease, although most likely to succeed, do not
represent the jail population as a whole.

Experimental Mortality Experimental sub-
jects often drop out of an experiment before
it is completed, and that can affect statistical
comparisons and conclusions. This is termed
“experimental mortality,” though it is some-
times referred to as attrition. In the classical
experiment involving an experimental and a
control group, each with a pretest and a post-
test, suppose that the heavy drinkers in the
experimental group are so turned off by the
video on the health effects of binge drinking
that they tell the experimenter to forget it and
leave. Those subjects who stick around for the
posttest are less heavy drinkers to start with, and the group results will thus reflect a substantial “decrease” in alcohol use.

In this example, mortality is related to the experimental stimulus itself: Subjects who score highest on the pretest are the most likely to drop out after viewing the video. Mortality may also be a problem in experiments that take place over a long period (people may move away) or in experiments that require a substantial commitment of effort or time by subjects; they may become bored with the study or simply decide it’s not worth the effort.

**Causal Time Order** In criminal justice research, there may be ambiguity about the time order of the experimental stimulus and the dependent variable. Whenever this occurs, the research conclusion that the stimulus caused the dependent variable can be challenged with the explanation that the “dependent” variable actually caused changes in the stimulus. Many early studies of the relationship between different types of punishments and rates of offending exhibited this threat to validity by relying on single interviews with subjects who were asked how they viewed alternative punishments and whether they had committed any crimes.

**Diffusion or Imitation of Treatments** In the event that experimental- and control-group subjects are in communication with each other, it’s possible that experimental subjects will pass on some elements of the experimental stimulus to the control group. David Weisburd and Loraine Green (1995) studied police interventions in high-crime areas by randomly assigning 56 drug “hot spots” to receive either intensive narcotics interventions (experimental condition) or traditional enforcement (control condition). Narcotics squads from the Jersey City, New Jersey, police department were then randomly assigned to experimental or control conditions. Three squads used intensive investigative techniques in the experimental hot spots; three other squads used traditional methods in the control hot spots. The researchers were aware that the integrity of the treatment would be threatened if the officers in the control squads attempted to “mimic” the experimental squad. To guard against this, squads working in experimental and control areas were physically separated.

Critics of highly localized crime prevention argue that reducing crime in one location simply displaces it to another. To test for this, Weisburd and Green (1995) examined measures of crime in two-block areas surrounding the hot spots. They found no evidence of displacement, but they did observe a reduction in some crime measures in areas adjacent to the experimental hot spots. This suggests that some treatment effects “spilled over” into areas that were not subject to experimental conditions.

**Compensatory Treatment** In many criminal justice experiments—such as a special job-training program for incarcerated felons—subjects in the control group may be deprived of something of value to them. In such cases, there may be pressures to offer some form of compensation. Recall the discussion in Chapter 4 of how police in the Kansas City Preventive Patrol Experiment patrolled the perimeter of reactive beats—those with no preventive patrol. As we noted, they more often used lights and sirens when responding to calls for service in the reactive beats. Some police officers compensated for the absence of preventive patrol in a way that reduced the differences among proactive, reactive, and control beats.

It’s helpful to think of treatment diffusion as the accidental spillover of an experimental stimulus, in contrast to the more intentional compensatory treatment by public officials that occurred in the Kansas City case. In applied criminal justice research, such treatment is probably more common, and researchers often take steps to prevent it. For example, in an evaluation of intensive supervision probation (ISP) programs in 11 sites, researchers from the RAND Corporation recognized that probation
officers might provide enhanced supervision to clients in the control group, under the assumption that the experimental program (ISP) was better than traditional probation (Petersilia, 1989). This potential for compensation was reduced in two ways. First, researchers explained to program staff why the experimental and control groups should receive different types of services. Second, a type of double-blind procedure was used whereby researchers tried to disguise the records of subjects in the control group so that probation staff could not distinguish subjects who were participating in the experiment from their regular probation caseload.

Compensatory Rivalry  In real-life experiments, subjects deprived of the experimental stimulus may try to compensate by working harder. Suppose an experimental career development program for corrections officers is the experimental stimulus; the control group may work harder than before in an attempt to keep pace with the “special” experimental subjects.

Demoralization  Feelings of deprivation among the control group may also result in subjects giving up. A career development program for corrections officers may prompt demoralization among control-group subjects who believe their opportunities for advancement will suffer. As a result, it may not be clear whether posttest differences between experimental and control groups in, say, job performance actually reflected program impacts or were due to demoralization among the control group.

Notice that the possibilities of compensatory rivalry and demoralization are based on subjects’ reactions to the experiment; diffusion and compensatory treatment are accidental or intentional extensions of the experimental stimulus to the control group. In studies in which agency staff, not researchers, administer an experimental treatment to subjects, there is a greater potential for intentional compensation. In studies in which subjects in a control group are aware that other subjects are receiving a desirable treatment, compensatory rivalry or demoralization is possible.

Ruling Out Threats to Internal Validity

These, then, are the threats to internal validity cited by Shadish, Cook, and Campbell. The classical experiment, coupled with proper subject selection and assignment, can potentially handle each of the 12 threats to internal validity.

How do researchers determine whether a particular design rules out threats to internal validity? Shadish, Cook, and Campbell provide excellent advice. Ruling out threats to validity requires a “. . . theory of ‘plausibility’ so that we know which of the many possible threats are plausible in [a] particular context” (2002:41). In other words, some threats make more sense in some settings than in others. Sorting this out requires having theory-based expectations to guide figuring out what validity threats might be at work in a particular situation. Let’s look again at the classical experiment, presented graphically in Figure 7.2.

Pursuing the example of the educational video as an attempt to reduce alcohol abuse, if we use the experimental design shown in Figure 7.2, we should expect two findings. For the experimental group, the frequency of drinking measured in their posttest should be less than in their pretest. In addition, when the two posttests are compared, the experimental group should have less drinking than the control group.

This design guards against the problem of history because anything occurring outside the experiment that might affect the experimental group should also affect the control group. There should still be a difference in the two posttest results. The same comparison guards against problems of maturation as long as the subjects have been randomly assigned to the two groups. Testing and instrumentation should not be problems because both the experimental and the control groups are subject to the same tests and experimenter effects.
If the subjects have been assigned to the two groups randomly, statistical regression should affect both equally, even if people with extreme scores on drinking (or whatever the dependent variable is) are being studied. Selection bias is ruled out by the random assignment of subjects.

Experimental mortality can be more complicated to handle because dropout rates may be different between the experimental and control groups. The experimental treatment itself may increase mortality in the group exposed to the video. As a result, the group of experimental subjects that received the posttest will differ from the group that received the pretest. In our example of the alcohol video, it would probably not be possible to handle this problem by administering a placebo, for instance. In general, however, the potential for mortality can be reduced by shortening the time between pretest and posttest, by emphasizing to subjects the importance of completing the posttest, or perhaps by offering cash payments for participating in all phases of the experiment.

The remaining problems of internal invalidity can be avoided through the careful administration of a controlled experimental design. The experimental design we’ve been discussing facilitates the clear specification of independent and dependent variables. Experimental and control subjects can be kept separate to reduce the possibility of diffusion or imitation of treatments. Administrative controls can be applied to avoid compensations given to the control group, and compensatory rivalry can be watched for and taken into account in evaluating the results of the experiment, as can the problem of demoralization.

We emphasize careful administration here. Random assignment, pretest and posttest measures, and use of control and experimental groups do not automatically rule out threats to validity. This caution is especially true in field studies and evaluation research, in which subjects participate in natural settings, and uncontrolled variation in the experimental stimulus may be present. Control over experimental conditions is the hallmark of this approach, but conditions in field settings are usually more difficult to control. In fact, David Weisburd and associates (Weisburd, Petrosino, and Mason, 1993) argue convincingly that field experiments involving a larger number of subjects become more difficult to manage, which offsets the advantages of random assignment and a large number of subjects.

Compare, for example, our hypothetical study of alcohol use among college students with the field experiment on ISP described by Joan Petersilia (1989). In Petersilia’s study, more intensive probation was the independent variable, and recidivism the dependent variable. Subjects were randomly assigned to the experimental group (ISP) or the control group (regular probation).

The alcohol use study could conceivably be completed in about 1 week, using subjects from a class, dormitory, or house on a single campus. ISP programs were evaluated in 11 sites over 4 years, using probation clients as subjects. The video on the health effects of alcohol use is a well-defined treatment that is readily standardized and can easily be controlled by researchers. The experimental treatment in the ISP programs was a reduced caseload for probation workers, together with an increased number of regular contacts with each probation client; program staff, not researchers, administered the experimental treatment. There was a great potential for uncontrolled variation in the delivery of ISP treatments; ISP is not a simple dichotomous treatment as is the video/no-video treatment in the alcohol use experiment.

Finally, the alcohol use questionnaire can easily be administered by researchers, providing reliable measures of alcohol use. Data on recidivism for the ISP study were collected by probation staff in each of the 11 sites. Although recidivism can be readily defined as the number of new arrests after beginning probation, there may have been wide variation in the
ability of staff in the 11 sites to reliably detect new arrests.

These remarks are not intended as criticism of the ISP study. In her description of the evaluation, Petersilia (1989) documents the extensive steps taken by RAND Corporation researchers to control possible validity threats. The important point is that field experiments and evaluations can present many obstacles that are not eliminated simply by adopting a randomized experimental design. Careful administration and control throughout the experiment are necessary to reduce potential threats to internal validity.

Generalizability and Threats to Validity

Potential threats to internal validity are only some of the complications faced by experimenters. They also have the problem of generalizing from experimental findings to the real world. Even if the results of an experiment are an accurate gauge of what happened during that experiment, do they really tell us anything about life in the wilds of society? With our examination of cause and effect in Chapter 4 in mind, we consider two dimensions of generalizability: construct validity and external validity.

Threats to Construct Validity

In the language of experimentation, construct validity is the correspondence between the empirical test of a hypothesis and the underlying causal process that the experiment is intended to represent. Construct validity is thus concerned with generalizing from our observations in an experiment to actual causal processes in the real world. In our hypothetical example, the educational video is how we operationalize the construct of understanding the health effects of alcohol use. Our questionnaire represents the dependent construct of actual alcohol use.

Are these reasonable ways to represent the underlying causal process in which understanding the effects of alcohol use causes people to reduce excessive or abusive drinking? It's a reasonable representation, but also one that is certainly incomplete. People develop an understanding of the health effects of alcohol use in many ways. Watching an educational video is one way; having personal experience, talking to friends and parents, taking other courses, and reading books and articles are others. Our video may do a good job of representing the health effects of alcohol use, but it is an incomplete representation of that construct. Alternatively, the video may be poorly produced, too technical, or incomplete. Then the experimental stimulus may not adequately represent the construct we are interested in—educating students about the health effects of alcohol use. There may also be problems with our measure of the dependent variable: questionnaire items on self-reported alcohol use.

By this time, you should recognize a similarity between construct validity and some of the measurement issues discussed in Chapter 5. Almost any empirical example or measure of a construct is incomplete. Part of construct validity involves how completely an empirical measure can represent a construct or how well we can generalize from a measure to a construct.

A related issue in construct validity is whether a given level of treatment is sufficient. Perhaps showing a single video to a group of subjects would have little effect on alcohol use, but administering a series of videos over several weeks would have a greater impact. We could test this experimentally by having more than one experimental group and varying the number of videos seen by different groups.

Threats to construct validity are problematic in criminal justice experiments, often because researchers do not clearly specify precisely what constructs are to be represented by particular measures or experimental treatments. For example, Anne Schneider (1990) points out that many studies of probation policies do not specify whether an innovative probation program represents a punishment or an
attempt to reform offenders through counseling or other treatments. As a consequence, it is not clear what specific components of an effective probation program might be. Do the punitive components of probation reduce subsequent offending, or does enhanced counseling have a greater impact?

Farrington and associates (Farrington, Ohlin, and Wilson, 1986:92) make a related point: “Most treatments in existing experiments are not based on a well-developed theory but on a vague idea about what might influence offending. The treatments given are often heterogeneous, making it difficult to know which element was responsible for any observed effect.” These authors also note the importance of thinking about levels of constructs. The RAND Corporation evaluation of intensive supervision provides a good example of why this is important. The ISP program explicitly defined enhanced probation as punitive. This leads to the question of how enhanced probation must be. If a typical probation officer carries an average caseload of 100 clients and sees each an average of twice per month, what workload and contact levels are sufficiently intensive to reduce recidivism? Cutting the workload in half and doubling the number of contacts would be more intensive, but would it be intensive enough to produce a decline in recidivism? Again, an experiment could test this question by including more than one experimental group and giving each a different level of probation supervision.

In sum, three elements in enhancing construct validity, therefore, are (1) linking constructs and measures to theory, (2) clearly indicating what constructs are represented by specific measures, and (3) thinking carefully about what levels of treatment may be necessary to produce some level of change in the dependent measure.

**Threats to External Validity**

Will an experimental study, conducted with the kind of control we have emphasized here, produce results that would also be found in more natural settings? Can an intensive probation program shown to be successful in Minneapolis achieve similar results in Miami? External validity represents a slightly different form of generalizability, one in which the question is whether results from experiments in one setting (time and place) will be obtained in other settings, or whether a treatment found to be effective for one population will have similar effects on a different group.

Threats to external validity are greater for experiments conducted under carefully controlled conditions. If the alcohol education experiment reveals that drinking decreased among students in the experimental group, then we can be confident that the video actually reduced alcohol use among our experimental subjects. But will the video have the same effect on high school students or adults if it is broadcast on television? We cannot be certain because the carefully controlled conditions of the experiment might have had something to do with the video’s effectiveness.

In contrast, criminal justice field experiments are conducted in more natural settings. Real probation officers in 11 different local jurisdictions delivered intensive supervision to real probationers in the RAND ISP experiment. Because of the real-world conditions and multiple sites, there were fewer potential threats to external validity. This is not to say that external validity is never a problem in field experiments. The 11 probation agencies that participated in this evaluation may not be typical of probation agencies in other areas. The simple fact that they were willing to participate in the study suggests that staff could be more dedicated or more amenable to trying new approaches to probation. One of the advantages of field experiments in criminal justice is that, because they take place under real-world conditions, results are more likely to be valid in other real-world settings as well.

You may have detected a fundamental conflict between internal and external validity. Threats to internal validity are reduced by
conducting experiments under carefully controlled conditions. But such conditions do not reflect real-world settings, and this restricts our ability to generalize results. Field experiments generally have greater external validity, but their internal validity may suffer because such studies are more difficult to monitor than those taking place in more controlled settings.

Shadish, Cook, and Campbell (2002:101) offered some useful advice for resolving the potential for conflict between internal and external validity. Explanatory studies that test cause-and-effect theories should place greater emphasis on internal validity; applied studies should be more concerned with external validity. This is not a hard-and-fast rule, because internal validity must be established before external validity becomes an issue. That is, applied researchers must have confidence in the internal validity of their cause-and-effect relationships before they ask whether similar relationships would be found in other settings.

**Threats to Statistical Conclusion Validity**

The basic principle of statistical conclusion validity is simple. Virtually all experimental research in criminal justice is based on samples of subjects that represent a target population. Larger samples of subjects, up to a point, are more representative of the target population than are smaller samples. Statistical conclusion validity becomes an issue when findings are based on small samples of cases. Because experiments can be costly and time consuming, they are frequently conducted with relatively small numbers of subjects. In such cases, only large differences between experimental and control groups on posttest measures can be detected with any degree of confidence.

In practice, this means that finding cause-and-effect relationships through experiments depends on two related factors: (1) the number of subjects and (2) the magnitude of posttest differences between the experimental and control groups. Experiments with large numbers of cases may be able to reliably detect small differences, but experiments with smaller numbers can detect only large differences.

Threats to statistical conclusion validity can be magnified by other difficulties in field experiments. If treatment spillover or compensation is a problem, then smaller differences in the experimental stimulus will be delivered to each group. More generally, Weisburd and associates (1993, 2003) concluded, after reviewing a large number of criminal justice experiments, that failure to maintain control over experimental conditions reduces statistical conclusion validity even for studies with large numbers of subjects. Furthermore, these authors found that as sample size increases, so do implementation difficulties, which undermines experimental results in a variety of ways.

**Variations in the Classical Experimental Design**

The basic experimental design is adapted to meet different research applications.

We now turn to a more systematic consideration of variations on the classical experiment that can be produced by manipulating the building blocks of experiments.

Slightly restating our earlier remarks, four basic building blocks are present in experimental designs: (1) the number of experimental and control groups, (2) the number and variation of experimental stimuli, (3) the number of pretest and posttest measurements, and (4) the procedures used to select subjects and assign them to groups. By way of illustrating these building blocks and the ways they are used to produce different designs, we adopt the system of notation used by Campbell and Stanley (1966). Figure 7.3 presents this notation and shows how it is used to represent the classical experiment and examples of variations on this design.

In Figure 7.3, the letter \( O \) represents observations or measurements, and \( X \) represents an experimental stimulus or treatment. Different
time points are displayed as \( t \) with a subscript to represent time order. Thus, for the classical experiment shown in Figure 7.3, \( O \) at \( t_1 \) is the pretest, \( O \) at \( t_3 \) is the posttest, and the experimental stimulus, \( X \), is administered to the experimental group at \( t_2 \), between the pretest and posttest. Measures are taken for the control group at times \( t_1 \) and \( t_3 \), but the experimental stimulus is not administered to the control group.

Now consider the design labeled “Posttest Only.” As implied by its name, no pretest measures are made on either the experimental or the control group. Thinking for a moment about the threats to internal validity, we can imagine situations in which a posttest-only design is appropriate. Testing and retesting might especially influence subjects’ behavior if measurements are made by administering a questionnaire, with subjects’ responses to the posttest potentially affected by their experience in the pretest. A posttest-only design can reduce the possibility of testing as a threat to validity by eliminating the pretest.

Without a pretest, it is obviously not possible to detect change in measures of the dependent variable, but we can still test the effects of the experimental stimulus by comparing posttest measures for the experimental group with posttest measures for the control group. For example, if we are concerned about the possibility of sensitizing subjects in a study of an alcohol education video, we might eliminate the pretest and examine the posttest differences between the experimental and control groups. Randomization is the key to the posttest-only design. If subjects are randomly assigned to experimental and control groups, we expect them to be equivalent. Any posttest differences between the two groups on the dependent variable can then be attributed to the influence of the video.

In general, posttest-only designs are appropriate when researchers suspect that the process of measurement may bias subjects’ responses to a questionnaire or other instrument. This is more likely when only a short time elapses between pretest and posttest measurements. The number of observations made on subjects is a design building block that can be varied as needed. We emphasize here that random assignment is essential in a posttest-only design.

Figure 7.3 also shows a factorial design, which has two experimental groups that receive different treatments, or different levels of a single treatment, and one control group. This design is useful for comparing the effects of different interventions or different amounts of a single treatment. In evaluating an ISP program, we might wish to compare how different levels of contact between probation officers and probation clients affect recidivism. In this case, subjects in one experimental group might receive weekly contact (\( X_1 \)), the other experimental group be seen by probation officers twice each week (\( X_2 \)), and control-group subjects have normal contact (say, monthly) with probation officers. Because more contact is more expensive than less contact is, we would be interested in seeing how much difference in
recidivism was produced by monthly, weekly, and twice-weekly contacts.

Thus, an experimental design may have more than one group receiving different versions or levels of experimental treatment. We can also vary the number of measurements made on dependent variables. No hard-and-fast rules exist for using these building blocks to design a given experiment. A useful rule of thumb, however, is to keep the design as simple as possible to control for potential threats to validity. The specific design for any particular study depends on the research purpose, available resources, and unavoidable constraints in designing and actually carrying out the experiment.

One very common constraint is how subjects or units of analysis are selected and assigned to experimental or control groups. This building block brings us to the subject of quasi-experimental designs.

Quasi-Experimental Designs

When randomization is not possible, researchers can use different types of quasi-experimental designs. By now, the value of random assignment in controlling threats to validity should be apparent. However, it is often impossible to randomly select subjects for experimental and control groups and satisfy other requirements. Most often, there may be practical or administrative obstacles. There may also be legal or ethical reasons why randomization cannot be used in criminal justice experiments.

When randomization is not possible, the next-best choice is often a quasi-experiment. The prefix quasi-, meaning “to a certain degree,” is significant—a quasi-experiment is, to a certain degree, an experiment. In most cases, quasi-experiments do not randomly assign subjects and therefore may suffer from the internal validity threats that are so well controlled in true experiments. Without random assignment, the other building blocks of experimental design must be used creatively to reduce validity threats. Following Shadish, Cook, and Campbell, we will group quasi-experimental designs into two categories: (1) nonequivalent-groups designs and (2) time-series designs. Each can be represented with the same O, X, and t notation used to depict experimental designs.

Nonequivalent-Groups Designs

The name for this family of designs is also meaningful. The main strength of random assignment is that it allows us to assume equivalence in experimental and control groups. When it is not possible to create groups through randomization, we must use some other procedure, one that is not random. If we construct groups through some nonrandom procedure, however, we cannot assume that the groups are equivalent—hence, the label nonequivalent-groups design.

Whenever experimental and control groups are not equivalent, we should select subjects in some way that makes the two groups as comparable as possible. Often, the best way to achieve comparability is through a matching process in which subjects in the experimental group are matched with subjects in a comparison group. The term comparison group is commonly used, rather than control group, to highlight the nonequivalence of groups in quasi-experimental designs. A comparison group does, however, serve the same function as a control group.

Some examples of research that use nonequivalent-group designs illustrate various approaches to matching and the creative use of experimental design building blocks. Examples include studies of child abuse (Widom, 1989), obscene phone calls (Clarke, 1997), and video cameras for crime prevention (Gill and Spriggs, 2005). Figure 7.4 shows a diagram of each design using the X, O, and t notation. The solid line that separates treatment and comparison groups in the figure signifies that subjects have been placed in groups through some nonrandom procedure.

Child Abuse and Later Arrest    Cathy Spatz Widom studied the long-term effects of child abuse—whether abused children are more
likely to be charged with delinquent or adult criminal offenses than are children who were not abused. Child abuse was the experimental stimulus, and the number of subsequent arrests was the dependent variable.

Of course, it is not possible to assign children randomly to groups in which some are abused and others are not. Widom’s design called for selecting a sample of children who, according to court records, had been abused. She then matched each abused subject with a comparison subject—of the same gender, race, age, and approximate socioeconomic status (SES)—who had not been abused. The assumption with these matching criteria was that age at the time of abuse, gender, race, and SES differences might confound any observed relationship between abuse and subsequent arrests.

You may be wondering how a researcher selects important variables to use in matching experimental and comparison subjects. We cannot provide a definitive answer to that question, any more than we can specify what particular variables should be used in a given experiment. The answer ultimately depends on the nature and purpose of the experiment. As a general rule, however, the two groups should be comparable in terms of variables that are likely to be related to the dependent variable under study. Widom matched on gender, race, and SES because these variables are correlated with juvenile and adult arrest rates. Age at the time of reported abuse was also an important variable because children abused at a younger age had a longer “at-risk” period for delinquent arrests.

Widom produced experimental and comparison groups matching individual subjects. It is also possible to construct experimental and comparison groups through aggregate matching, the average characteristics of each group are comparable. This is illustrated in our next example.

### Figure 7.4 Quasi-Experimental Design Examples

<table>
<thead>
<tr>
<th>Widom (1989)</th>
<th>Treatment group</th>
<th>X</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison group</td>
<td>O</td>
<td>t₁</td>
<td>t₂</td>
</tr>
</tbody>
</table>

X = official record of child abuse
O = counts of juvenile or adult arrest

<table>
<thead>
<tr>
<th>Clarke (1990)</th>
<th>Treatment group</th>
<th>O</th>
<th>X</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison group</td>
<td>O</td>
<td>t₁</td>
<td>t₂</td>
<td>t₃</td>
</tr>
</tbody>
</table>

X = caller identification and call tracing
O = customer complaints of obscene calls

<table>
<thead>
<tr>
<th>Gill and Springs (2005)</th>
<th>Target area 1</th>
<th>O</th>
<th>X₁</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison area 1</td>
<td>O</td>
<td>t₁</td>
<td>t₂</td>
<td>t₃</td>
</tr>
</tbody>
</table>

X₁ = CCTV installation in area i
O = police crime data, survey data on fear of crime

### Deterring Obscene Phone Calls

In 1988, the New Jersey Bell telephone company introduced caller identification (ID) and instant call tracing in a small number of telephone exchange areas. Now ubiquitous in mobile phones, caller ID was a new technology in 1988. Instant call tracing allows the recipient of an obscene or threatening call to automatically initiate a procedure to trace the source of the call.

Ronald Clarke (1997) studied the effects of these new technologies in deterring obscene phone calls. Clarke expected that obscene calls would decrease in areas where the new services were available. To test this, he compared records of formal customer complaints about
annoying calls in the New Jersey areas that had the new services to formal complaints in other New Jersey areas where caller ID and call tracing were not available. One year later, the number of formal complaints had dropped sharply in areas serviced by the new technology; no decline was found in other New Jersey Bell areas.

In this study, telephone service areas with new services were the treatment group, and areas without the services were the comparison group. Clarke's matching criterion was a simple one: telephone service by New Jersey Bell, assuming the volume of obscene phone calls was relatively constant within a single phone service area. Of course, matching on telephone service area cannot eliminate the possibility that the volume of obscene phone calls varies from one part of New Jersey to another, but Clarke's choice of a comparison group was straightforward and certainly more plausible than comparing New Jersey to, say, New Mexico.

Clarke's study is a good example of a natural field experiment. The experimental stimulus—caller ID and call tracing—was not specifically introduced by Clarke, but he was able to obtain measures for the dependent variable before and after the experimental stimulus was introduced. This design made it possible for Clarke to infer with reasonable confidence that caller ID and call tracing reduced the number of formal complaints about obscene phone calls.

### Cameras and Crime Prevention

U.S. residents have probably become accustomed to seeing closed-circuit television (CCTV) cameras in stores and at ATMs, but this technology is less used in public spaces such as streets and parking lots. With an estimated 4 million cameras deployed, CCTV is widely used as crime prevention and surveillance tools in the United Kingdom (McCahill and Norris, 2003). CCTV enabled the London Metropolitan Police to quickly identify suspects in the Underground bombing attacks that took place in 2005. Cameras are increasingly used to monitor traffic and even record license plates of cars running traffic lights. But does CCTV have any effect in reducing crime?

Martin Gill and associates (Gill and Spriggs, 2005; Gill, Spriggs, Arhomaniz, et al., 2005) conducted an evaluation of 13 CCTV projects installed in a variety of residential and commercial settings in England. These were a mix of small- and large-scale CCTV projects involving multiple cameras. One area on the outskirts of London included over 500 cameras installed to reduce thefts of and from vehicles in parking facilities. Five projects in London and other urban areas placed 10–15 cameras in low-income housing areas, seeking to reduce burglary and robbery. Researchers examined two types of dependent variables before and after cameras were installed: crimes reported to police, and fear of crime. Fear was measured through surveys of people living in residential areas and samples of people on local streets for commercial areas and parking facilities.

Measuring police data and fear of crime before and after cameras were installed made it possible for Gill and associates to satisfy two criteria for cause—time order and covariation between the independent variable (CCTV) and dependent variables. Random assignment of areas to treatment (CCTV) or control (no CCTV) groups was not possible. This was because the intervention was planned for only a small number of locations of each type (residential, commercial, parking) and because CCTV was carefully tailored to each site. Instead, the researchers created two types of comparison areas. First, comparison areas “. . . were selected by similarity on socio-demographic and geographical characteristics and crime problems” (pp. 123–4). The second type of comparison was “buffer zones,” defined as an area in a 1-mile radius from the edge of the target area where CCTV cameras were installed; buffer zones were defined only for CCTV areas.

The rationale for comparison areas is clear. If CCTV is effective in reducing crime, we should expect declines in target areas, but not in comparison areas. Alternatively, if posttreatment
measures of crime went down in both treatment and comparison areas, we might expect greater declines in the CCTV sites. But what about buffer areas? After defining buffer areas, researchers then subdivided them into concentric rings around a target area, shown as $T$ in Figure 7.5. The stated purpose was to assess any movement of crime around the target area. If CCTV was effective in reducing crime, any reduction should be greatest in the target area; the size of the reduction should decline moving outward from the target area.\footnote{Compare this with the Chicago research by Shaw and McKay described in Chapter 2.}

Short-term results found some reduction of some types of crime in some CCTV areas. In other treatment areas, some crimes increased more than in comparison areas. In particular, Gill and Springs found that public order offenses such as drunkenness tended to increase more in CCTV target areas. Overall, significant drops in crime were found in just 2 of 13 target areas. Fear and related attitudes declined in all target and comparison areas, but authors believed this was largely due to declining crime in all areas.

This example illustrates why nonequivalent comparison groups are important. Because crime declined in most areas and fear declined in all, a simple comparison of pre- and post-intervention measures would have been misleading. That strategy would have suggested that CCTV was responsible for reduced crime and fear. Only by adding the comparison and buffer areas to their research were Gill and Spriggs able to learn that CCTV was probably not the cause of declines because similar patterns were found in many areas where CCTV systems were not installed.

Together, these three studies illustrate different approaches to research design when it is not possible to randomly assign subjects to treatment and control groups. Lacking random assignment, researchers must use creative procedures for selecting subjects, constructing treatment and comparison groups, measuring dependent variables, and exercising other controls to reduce possible threats to validity.

**Cohort Designs**

Chapter 4 mentioned cohort studies as examples of longitudinal designs. We can also view cohort studies as nonequivalent-control-group designs. Recall from Chapter 4 that a cohort may be defined as a group of subjects who enter or leave an institution at the same time. For example, a class of police officers who graduate from a training academy at the same time could be considered a cohort. Or we might view all persons who were sentenced to probation in May as a cohort.

Now think of a cohort that is exposed to some experimental stimulus. The May probation cohort might be required to complete 100 hours of community service in addition to meeting other conditions of probation. If we are interested in whether probationers who receive community service sentences are charged with fewer probation violations, we can compare the performance of the May cohort with that of the April cohort or the June cohort or some other cohort not sentenced to community service. Cohorts that do not receive community service sentences serve as comparison groups. The groups are not equivalent because

![Figure 7.5 Buffer Zones in CCTV Quasi-Experiment](source: Adapted from Gill and Spriggs (2005:40).)


they were not created by random assignment. But if we assume that a comparison cohort does not systematically differ from a treatment cohort on important variables, we can use this design to determine whether community service sentences reduce probation violations.

That last assumption is very important, but it may not be viable. Perhaps a criminal court docket is organized to schedule certain types of cases at the same time, so a May cohort would be systematically different from a June cohort. But if the assumption of comparability can be met, cohorts may be used to construct nonequivalent comparison and experimental groups by taking advantage of the natural flow of cases through some institutional process.

A study by Edmund Risler and associates (Risler, Sweatman, and Nackerud, 1998) on the effectiveness of a judicial waiver law in deterring juvenile crime is a good example of a cohort design. Judicial waiver laws signaled a shift in juvenile justice philosophy, largely because of a perceived increase in violent crime by juveniles and a growing sentiment that juvenile courts were not tough enough on young violent offenders. A judicial waiver law in Georgia required that anyone between the ages of 13 and 17 be tried as an adult if charged with one of several violent crimes: murder, voluntary manslaughter, rape, aggravated sexual battery, aggravated child molestation, aggravated sodomy, or robbery with a firearm. Risler and associates were interested in whether the law achieved its goal of deterrence.

The Georgia law—the experimental stimulus—became effective in 1994. Two samples of juveniles were studied: (1) a preintervention cohort, juveniles processed during the 2-year period prior to the law’s enactment (1992–93); and (2) a postintervention cohort, juveniles processed during the 2-year period after the law took effect (1994–95). Mean arrest rates for juveniles for the seven offenses covered under the judicial waiver law were compared across these two samples. Arrest rates for 1992 and 1993 served as observations for the pretest cohort; those for 1994 and 1995 served as post-test observations.

The two cohorts could not be considered equivalent because they were not produced by random assignment. However, they were assumed to be sufficiently similar for the purpose of the authors’ study.

Risler and associates found that the mean arrests rate for crimes covered under the judicial waiver law actually increased slightly for the postintervention cohort. The authors therefore concluded that the law did not act as a deterrent to reduce violent offending by juveniles.

Time-Series Designs

Time-series designs are common examples of longitudinal studies in criminal justice research. As the name implies, a time-series design involves examining a series of observations on some variable over time. A simple example is examining trends in arrests for drunk driving over time to see whether the number of arrests is increasing, decreasing, or staying constant. A police executive might be interested in keeping track of arrests for drunk driving, or for other offenses, as a way of monitoring the performance of patrol officers. Or state corrections officials might want to study trends in prison admissions as a way of predicting the future need for correctional facilities.

An interrupted time series is a special type of time-series design that can be used in cause-and-effect studies. A series of observations is compared before and after some intervention is introduced. For example, a researcher might want to know whether roadside sobriety checkpoints cause a decrease in fatal automobile accidents. Trends in accidents could be compared before and after the roadside checkpoints are established.

Interrupted time-series designs can be very useful in criminal justice research, especially in applied studies. They do have some limitations, however, just like other ways of structuring research. Shadish, Cook, and Campbell (2002) describe the strengths and limitations...
of different approaches to time-series designs. We will introduce these approaches with a hypothetical example and then describe some specific criminal justice applications.

Continuing with the example of sobriety checkpoints, Figure 7.6 presents four possible patterns of alcohol-related automobile accidents. The vertical line in each pattern shows the time when the roadside checkpoint program is introduced. Which of these patterns indicates that the new program caused a reduction in car accidents?

If the time-series results looked like pattern 1 in Figure 7.6, we might think initially that
the checkpoints caused a reduction in alcohol-related accidents, but there seems to be a general downward trend in accidents that continues after the intervention. It’s safer to conclude that the decline would have continued even without the roadside checkpoints.

Pattern 2 shows that an increasing trend in auto accidents has been reversed after the intervention, but this appears to be due to a regular pattern in which accidents have been bouncing up and down. The intervention was introduced at the peak of an upward trend, and the later decline may be an artifact of the underlying pattern rather than of the new program.

Patterns 1 and 2 exhibit some outside trend, rather than an intervention, that may account
for a pattern observed over time. We may recognize this as an example of history as a validity threat to the inference that the new checkpoint program caused a change in auto accidents. The general decline in pattern 1 may be due to reduced drunk driving that has nothing to do with sobriety checkpoints. Pattern 2 illustrates what is referred to as seasonality in a time series—a regular pattern of change over time. Cook and Campbell (1979) described seasonality as a special case of history. In our example, the data might reflect seasonal variation in alcohol-related accidents that occur around holidays or maybe on football weekends near a college campus.

Patterns 3 and 4 lend more support to the inference that sobriety checkpoints caused a decline in alcohol-related accidents, but the two patterns are different in a subtle way. In pattern 3, accidents decline more sharply from a general downward trend immediately after the checkpoint program was introduced, whereas pattern 4 displays a sharper decline some time after the new program was established. Which pattern provides stronger support for the inference?

In framing your answer, recall what we have said about construct validity. Think about the underlying causal process these two patterns represent, or consider possible mechanisms that might be at work. Pattern 3 suggests that the program was immediately effective and supports what we might call an incapacitation mechanism: Roadside checkpoints enabled police to identify and arrest drunk drivers, thereby getting them off the road and reducing accidents. Pattern 4 suggests a deterrent mechanism: As drivers learned about the checkpoints, they less often drove after drinking, and accidents eventually declined. Either explanation is possible given the evidence presented. This illustrates an important limitation of interrupted time-series designs: They operationalize complex causal constructs in simple ways. Our interpretation depends in large part on how we understand this causal process.

The classic study by Richard Mc Cleary and associates (Mc Cleary, Neinstedt, and Erven, 1982), mentioned in Chapter 5, illustrates the need to think carefully about how well time-series results reflect underlying causal patterns. Recall that Mc Cleary and colleagues reported a sharp decline in burglaries immediately after a special burglary investigation unit was established in a large city. This finding was at odds with their understanding of how police investigations could reasonably be expected to reduce burglary. A special unit might eventually be able to reduce the number of burglaries, after investigating incidents over a period of time and making arrests. But it is highly unlikely that changing investigative procedures would have an immediate impact. This discrepancy prompted Mc Cleary and associates to look more closely at the policy change and led to their conclusion that the apparent decline in burglaries was produced by changes in record-keeping practices. No evidence existed of any decline in the actual number of burglaries.

This example illustrates our discussion of instrumentation earlier in this chapter. Changes in the way police counted burglaries produced what appeared to be a reduction in burglary. Instrumentation can be a particular problem in time-series designs for two reasons. First, observations are usually made over a relatively long time period, which increases the likelihood of changes in measurement instruments. Second, time-series designs often use measures that are produced by an organization such as a police department, criminal court, probation office, or corrections department. There may be changes or irregularities in the way data are collected by these agencies that are not readily apparent to researchers and that are, in any case, not subject to their control.

As another example, David Finkelhor and Lisa Jones (2004) systematically examine data showing a nationwide 40 percent decline in child sexual abuse cases from 1990 to 2000. Because child abuse had become a prominent
concern in the 1990s, the researchers were especially interested in determining whether any changes in recording practices in one or more states might account for the decline. They systematically considered such explanations as changing definitions or data collection procedures, more stringent screening of cases by investigators, less reporting of sexual abuse cases, and other possible sources of instrumentation. After comparing different sources of data for a variety of states, Finkelhor and Jones conclude that at least some portion of the decline in reported cases is due to a real decline in the number of child sexual abuse cases. Their careful report is an excellent example of thinking through plausible threats to instrumentation in a time series.

Variations in Time-Series Designs

If we view the basic interrupted time-series design as an adaptation of basic design building blocks, we can consider how modifications can help control for many validity problems. The simplest time-series design studies one group—the treatment group—over time. Rather than making one pretest and one posttest observation, the interrupted time-series design makes a longer series of observations before and after introducing an experimental treatment.

What if we considered the other building blocks of experimental design? Figure 7.7 presents the basic design and some variations using the familiar \( O, X, \) and \( t \) notation. In the basic design, shown at the top of Figure 7.7, many pretest and posttest observations are made on a single group that receives some treatment. We could strengthen this design by adding a comparison series of observations on some group that does not receive the treatment. If, for example, roadside sobriety checkpoints were introduced all over the state of Ohio but were not used at all in Michigan, then we could compare auto accidents in Ohio (the treatment series) with auto accidents in Michigan (the comparison series). If checkpoints caused a reduction in alcohol-related accidents, we would expect to see a decline in Ohio following the intervention, but there should be no change or a lesser decline in Michigan over the same time period. The second part of Figure 7.7 shows this design—an interrupted time series with a nonequivalent comparison group. The two series are not equivalent because we did not randomly assign drivers to Ohio or Michigan. So Young Kim and Wesley Skogan (2003) present a good example of this design in their analysis of police problem solving in Chicago. They examined changes in crime for police beats where specific problems were targeted, to comparison beats where no crime-specific interventions were developed.

A single-series design may be modified by introducing and then removing the intervention, as shown in the third part of Figure 7.7. We might test sobriety checkpoints by setting them up every weekend for a month and then not setting them up for the next few months. If the checkpoints caused a reduction in alcohol-related accidents, we might expect an increase

---

**Figure 7.7** Interrupted Time-Series Designs

---
after they were removed. Or the effects of weekend checkpoints might persist even after we removed them.

Because different states or cities sometimes introduce new drunk-driving programs at different times, we might be able to use what Shadish, Cook, and Campbell (2002:192) called a time-series design with switching replications. The bottom of Figure 7.7 illustrates this design. For example, assume that Ohio begins using checkpoints in May 1998 and Michigan introduces them in July of the same year. A switching replications design could strengthen our conclusion that checkpoints reduce accidents if we saw that a decline in Ohio began in June and a similar pattern was found in Michigan beginning in August. The fact that similar changes occurred in the dependent variable in different states at different times, corresponding to when the program was introduced, would add to our confidence in stating that sobriety checkpoints actually reduced auto accidents.

**Variable-Oriented Research and Scientific Realism**

Another way to think about a time-series design is as a study of one or a few cases with many observations. If we design a time-series study of roadside checkpoints in Ohio, we will be examining one case (Ohio) with many observations of auto accidents. Or a design that compares Ohio and Michigan will examine many observations for two cases. Thinking once again about design building blocks, notice how we have slightly restated one of those building blocks. Instead of considering the number of experimental and control groups, our attention centers on the number of subjects or cases in our study. In Figure 7.7, the first and third time-series designs have one case each; the second and fourth designs examine two cases each.

Classical experiments and quasi-experiments with large numbers of subjects are examples of what Charles Ragin (2000) terms **case-oriented research**, in which many cases are examined to understand a small number of variables. Time-series designs and case studies are examples of **variable-oriented research**, in which a large number of variables are studied for a small number of cases or subjects. Suppose we wish to study inmate-on-inmate assaults in correctional facilities. With case-oriented approach, we might send a questionnaire to a sample of 500 correctional facilities, asking facility staff to provide information about assaults, facility design, inmate characteristics, and housing conditions. Here, we are gathering information on a few variables from a large number of correctional facilities. Using a variable-oriented approach, we might visit one or a few facilities to conduct in-depth interviews with staff, observe the condition of facilities, and gather information from institutional records. Here, we are collecting information on a wide range of variables from a small number of institutions.

The **case-study** design is an example of variable-oriented research. Here, the researcher’s attention centers on an in-depth examination of one or a few cases on many dimensions. Robert Yin (2003) points out that the terms case and case study are used broadly. Cases can be individual people, neighborhoods, correctional facilities, courtrooms, or other aggregations. Our running example of race profiling has many elements of a series of case studies. Researchers gathered information from different sources to measure several types of variables for one state (New Jersey, North Carolina, Pennsylvania, etc.) and a varying number of other cases (state police stations, or counties) in the state.

Robert Yin cautions that the case study design is often misunderstood as representing “qualitative” research or participant observation study. Instead, Yin advises that the case study is a design strategy and that the labels qualitative and quantitative are not useful ways to distinguish design strategies. Case studies might appear qualitative because they focus on one or a small number of units. But many case
studies employ sophisticated statistical techniques to examine many variables for those units. An example illustrates how misleading it can be to associate case studies with qualitative research.

In what has come to be known as the “Boston Gun Project,” Anthony Braga and associates (Braga, Kennedy, Waring, and Piehl, 2001) studied violence by youth gangs in Boston neighborhoods. Theirs was an applied explanatory study. They worked with local officials to better understand gang violence, develop ways to reduce it, and eventually assess the effects of their interventions. Neither a classical experiment nor a nonequivalent-groups design was possible. Researchers sought to understand and reduce violence by all gangs in the city. Their research centered on gangs, not individuals, though some interventions targeted particular gang members.

Researchers collected a large amount of information about gangs and gang violence from several sources. Earlier reports (Kennedy, Piehl, and Braga, 1996) described something called “network analysis,” in which researchers examined relationships between gangs in different neighborhoods and conflicts over turf within neighborhoods. Police records of homicides, assaults, and shootings were studied. Based on extensive data on a small number of gangs, researchers collaborated with public officials, neighborhood organizations, and a coalition of religious leaders—the “faith community.” A variety of interventions were devised, but most were crafted from a detailed understanding of the specific nature of gangs and gang violence as they existed in Boston neighborhoods. David Kennedy (1998) summarizes these using the label “pulling levers,” signifying that key gang members were vulnerable to intensive monitoring via probation or parole. The package of strategies was markedly successful: Youth homicides were reduced from about 35–40 each year in the 20 years preceding the program to about 15 per year in the first 5 postintervention years (Braga, 2002:70).

The Boston research is also a good example of the scientific realist approach of Ray Pawson and Nick Tilley (1997). Researchers examined a small number of subjects—gangs and gang members—in a single city and in the context of specific neighborhoods where gangs were active. Extensive data were gathered on the mechanisms of gang violence. Interventions were tailored to those mechanisms in their context. Braga and associates (2001) emphasize that the success of the Boston efforts was due to the process by which researchers, public officials, and community members collaboratively studied gang violence and then developed appropriate policy actions based on their analyses. Other jurisdictions mistakenly tried to reproduce Boston’s interventions, with limited or no success, failing to recognize that the interventions were developed specifically for Boston. In case-study language, researchers examined many variables for one site and based policy decisions on that analysis. In the words of scientific realism, researchers studied the gang violence mechanism in the Boston context. In other contexts (Baltimore or Minneapolis, for example), gang violence operated as a different mechanism; the “levers” pulled in Boston did not work elsewhere. Braga and associates emphasize that the problem-solving process is exportable to other settings but that the interventions used in Boston are not (2001:220).

How do case studies address threats to validity? In the most general sense, case studies attempt to isolate causal mechanisms from possible confounding influences by studying very precisely defined subjects. Donald Campbell (2003:ix–x) likened this to laboratory experiments in the natural sciences, in which researchers try to isolate causal variables from outside influences. Case-study research takes place in natural field settings, not in laboratories. But the logic of trying to isolate causal mechanisms by focusing on one or a few cases is a direct descendant of the rationale for experimental isolation in laboratories.
Figure 7.8 summarizes advice from Yin (2003:33–39) on how to judge the quality of case-study designs in language that should now be familiar. Construct validity is established through multiple sources of evidence, the establishment of chains of causation that connect independent and dependent variables, and what are termed “member checks”—asking key informants to review tentative conclusions about causation. Examples of techniques for strengthening internal validity are theory-based pattern matching and time-series analysis. The first criterion follows Shadish, Cook, and Campbell, calling on researchers to...
percent of the resident population of a jurisdiction, this suggests a racial disparity. The problem with this comparison is that drivers are mobile by definition. The majority of cars on the southern third of the New Jersey Turnpike have out-of-state license plates. This approach might work on local streets in a large jurisdiction, but even then it ignores the fundamental fact that people are mobile. Researchers in England recognized this in their study of four municipalities (Miller, 2000).

- Driving population. This is better because it compares people stopped by police to people who are actually on the road. In settings like the New Jersey Turnpike, the race of drivers can be observed at tollbooths, as we described in Chapter 5. However, this method produces information about drivers, not traffic violators. Police claim to stop cars because they observe some violation, not because of race.

- Night-time traffic stops. This comparison strategy assumes that police are not able to detect the race of drivers at nighttime. So if the proportion of traffic stops of nonwhite drivers at night is lower than the proportion stopped during the day, some discriminatory targeting might be at work in daylight hours. Even though there is merit in such logic, New Jersey state police tend to make fewer stops at night. There have also been reports of “spot-lighting” by state troopers, where they use spotlights, or park their vehicle at right angles to the roadway and shine lights on the face of passing drivers (Verneiro and Zoubek, 1999).

- Radar-timed stops. Similar logic is at work here. If police make stops based on radar readings, it’s assumed they will be less likely to stop people based on the appearance of race. In this respect, police are presumed to exercise less discretion in relying solely on radar to decide whom to stop. Again, it’s not quite that simple. From their research in New Jersey, Maxfield and Andresen sat in state police cars while troopers used radar. In many cases the racial appearance of drivers was easily visible. Most vehicles were speeding, though some more than others. Together these observations suggest that state police could use race as a factor in deciding which speeding cars to stop and which to ignore.

- Independent observations of speeding. As we saw in the last chapter, researchers in three states collected their own data on speeding. In principle this offers the best comparison against police stop data: comparing the race distribution of police stop data to the race distribution of speeders from data collected independently. If the percentages are similar, it supports police claims that they stop only traffic violators. If they are different, this would be stronger evidence of unexplained disparity.

We say “unexplained disparity” intentionally. Such a disparity would support discrimination as an explanation, but it would not prove that differences were due to discrimination. That’s because we cannot be as certain as we would like that all other plausible explanations have been eliminated. Slightly restated, we cannot rule out threats to internal validity of a claim that discrimination causes disparities.

make specific theory-based predictions about what pattern of results will support hypothesized causal relationships. Alternative explanations, also termed “rival hypotheses,” are less persuasive when specific predictions of results are actually obtained. For example, Braga and associates (2001) predicted that gun killings among male Boston residents under age 25 would decline following implementation of the package of interventions in the Boston gun strategy. Although other explanations are possible for the sharp observed declines, the specific focus of the researchers’ interventions and the concomitant results undermine the
credibility of rival hypotheses. Having many measures of variables over time strengthens internal validity if observations support our predicted expectations about cause. We saw earlier how nonequivalent time-series comparisons and switching replications can enhance findings. This is also consistent with pattern matching—we make specific statements about what patterns of results we expect in our observations over time.

Finally, a single case study is vulnerable to external validity threats because it is rooted in the context of a specific site. Conducting multiple case studies in different sites illustrates the principle of replication. By replicating research findings, we accumulate evidence. We may also find that causal relationships are different in different settings, as did researchers who tried to transplant specific interventions from the Boston Gun Project. Although such findings can undermine the generalizability of causality, they also help us understand how causal mechanisms can operate differently in different settings.

Time-series designs and case studies are examples of variable-oriented research. A case study with many observations over time can be an example of a time-series design. Adding one or more other cases offers opportunities to create nonequivalent comparisons. Time-series designs, case studies, and nonequivalent comparisons are quasi-experimental designs—they are conducted in the manner of experiments, using design building blocks in different ways.

### Case Study Approach

<table>
<thead>
<tr>
<th>Construct Validity</th>
<th>Internal Validity</th>
<th>External Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple sources of evidence</td>
<td>Pattern-matching</td>
<td>Replicate through multiple case studies</td>
</tr>
<tr>
<td>Establish chain of causation</td>
<td>Time-series analysis</td>
<td></td>
</tr>
<tr>
<td>Member checks</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 7.8** Case Studies and Validity

*Source: Adapted from Yin (2003:34).*

By now, it should be clear that there are no simple formulas or recipes for designing an experimental or quasi-experimental study. Researchers have an almost infinite variety of ways of varying the number and composition of groups of subjects, selecting subjects, determining how many observations to make, and deciding what types of experimental stimuli to introduce or study.

Variations on the classical experimental designs are especially useful for explanatory research and in evaluation studies, but exploratory and descriptive studies usually use other methods. Surveys conducted at one point in time, for example, may be used to explore or describe such phenomena as fear of crime or public attitudes toward punishment. Longitudinal studies of age cohorts are often the best ways to examine criminal careers or developmental causes of delinquency.

Even when experimental designs might be the best choice, it is not always possible to construct treatment and control groups, to use random assignment, or even to analyze a series of observations over time. Research on CCTV is an example.

As we stated early in this chapter, experiments are best suited to topics that involve well-defined concepts and propositions. Experiments and quasi-experiments also require that researchers be able to exercise, or at least approximate, some degree of control over an experimental stimulus. Finally, these designs depend on the ability to unambiguously establish
the time order of experimental treatments and observations on the dependent variable. Often, it is not possible to achieve the necessary degree of control.

In designing research projects, researchers should be alert to opportunities for using experimental designs. Researchers should also be aware of how quasi-experimental designs can be developed when randomization is not possible. Experiments and quasi-experiments lend themselves to a logical rigor that is often much more difficult to achieve in other modes of observation. The building blocks of research design can be used in creative ways to address a variety of criminal justice research questions. Careful attention to design issues, and to how design elements can reduce validity threats, is essential to the research process.

**Main Points**

- Experiments are an excellent vehicle for the controlled testing of causal processes. Experiments may also be appropriate for evaluation studies.
- The classical experiment tests the effect of an experimental stimulus on some dependent variable through the pretesting and posttesting of experimental and control groups.
- It is less important that a group of experimental subjects be representative of some larger population than that experimental and control groups be similar to each other.
- Randomization is the best way to achieve comparability in the experimental and control groups.
- The classical experiment with random assignment of subjects guards against most of the threats to internal invalidity.
- Because experiments often take place under controlled conditions, results may not be generalizable to real-world constructs. Or findings from an experiment in one setting may not apply to other settings.
- The classical experiment may be modified to suit specific research purposes by changing the number of experimental and control groups, the number and types of experimental stimuli, and the number of pretest or posttest measurements.
- Quasi-experiments may be conducted when it is not possible or desirable to use an experimental design.
- Nonequivalent-groups and time-series designs are two general types of quasi-experiments.
- Time-series designs and case studies are examples of variable-oriented research, in which a large number of variables are examined for one or a few cases.
- Both experiments and quasi-experiments may be customized by using design building blocks to suit particular research purposes.
- Not all research purposes and questions are amenable to experimental or quasi-experimental designs because researchers may not be able to exercise the required degree of control.

**Key Terms**

- Case-oriented research
- Generalizability
- Case study
- Independent variable
- Classical experiment
- Quasi-experiment
- Control group
- Randomization
- Dependent variable
- Variable-oriented research
- Experimental group

**Review Questions and Exercises**

1. If you do not remember participating in D.A.R.E.—Drug Abuse Resistance Education—you have probably heard or read something about it. Describe an experimental design to test the causal hypothesis that D.A.R.E. reduces drug use. Is your experimental design feasible? Why or why not?

2. Experiments are often conducted in public health research, where a distinction is made between an efficacy experiment and an effectiveness experiment. Efficacy experiments focus on whether some new health program works under ideal conditions; effectiveness experiments test the program under typical conditions that health professionals encounter in their day-to-day work. Discuss how efficacy experiments and effectiveness experiments reflect concerns about internal validity threats on the one hand and generalizability on the other.

3. Crime hot spots are areas where crime reports, calls for police service, or other measures of crime are especially common. Police in departments with a good analytic capability routinely
identify hot spots and launch special tactics to reduce crime in these areas. What kinds of validity threats should researchers be especially attentive to in studying the effects of police interventions on hot spots?

4. David Weisburd and associates (Weisburd, Lum, and Petrosino, 2001) studied relationships between the types of research designs and the findings of applied research projects. The authors report that experimental designs are more likely to find no relationship between independent and dependent variables; quasi-experimental designs are more likely to support hypothesized relationships between independent and dependent variables. Given your understanding of threats to validity, present at least two alternative explanations for the findings by Weisburd and associates.

Online Sessions for Chapter 7

Online Sessions present further examples of selected material covered in each chapter. Your instructor may choose to use these for in-class discussion (as I do with my students) or you can access them at home. To ensure that you always have access to live, correct links for the websites described here, we provide the necessary links on the companion website for Research Methods for Criminal Justice and Criminology at http://www.thomsonedu.com/criminaljustice/maxfield. Once at the companion website, select this specific chapter, click on “Chapter Resources,” then click on “Web Links.”

Experimentation in Social Science

Randomization is one of the hallmarks of a true experimental design. Visit William Trochim’s Knowledge Base website to help you to understand and summarize the difference between random assignment and random sampling. Here Trochim also describes building blocks as elements of design, which “… provides the glue that holds the research project together.” http://www.socialearchresearchmethods.net/kb/design.htm

If you are ready to randomly assign subjects to conditions, visit the Research Randomizer website. Here you will recognize the important role of experiments in psychology, while learning more about the mechanics of random assignment.

http://www.randomizer.org/

Sites maintained by U.S. government agencies describe the role of randomized trials in health and education research. Visit one or more of these sites and compare typical designs to those used in criminal justice research.

http://www.cancer.gov/clinicaltrials/learning/what-is-randomization

Experimental Criminology and Crime Science

Over the last 10 years, many criminal justice researchers have promoted increased use of experimental designs. This can be largely traced to the publication of Preventing Crime: What Works, What Doesn’t, What’s Promising (Sherman et al., 1997). One of the most interesting features of that report is a measure of the strength of research methods used by various crime prevention studies. Go to the report’s website and review Chapter 2, “Thinking About Crime Prevention,” by Lawrence Sherman. Also read the appendix on research methods by Sherman and Denise Gottfredson. Pay particular attention to the “scale of evidentiary strength for cause and effect.” Then consider how these two sections of the report use the language of research design and validity threats from this chapter of your text. http://www.ncjrs.gov/works/

In the years since that report was published, two related organizations have been established to compile experimental studies in criminology, and publish research articles using experimental and quasi-experimental designs. The Academy of Experimental Criminology publishes the Journal of Experimental Criminology. Information about the Academy and a link to its journal are on the Academy web page. http://www.crim.upenn.edu/aec/index.html

Researchers have also established the Campbell Collaboration’s Crime & Justice Coordinating Group to compile and classify research on crime prevention and related topics. Named after Donald Campbell, the Collaboration is loosely modeled after efforts to compile public health research. Visit the Collaboration’s website and try to find examples of studies in some specific area of interest to you. http://www.campbellcollaboration.org/CCJG/index.asp

Researchers at the Jill Dando Institute of Crime Science advocate approaches to criminal justice research that rely less on the kinds of formal designs advocated by experimental criminologists. Instead, most projects are based on two principles: scientific realism and replication. We
discussed some of the basics of scientific realism in this chapter. Replication is emphasized as a way to accumulate knowledge about “what works” in crime prevention by repeatedly testing interventions in different settings. In fact, Gloria Laycock, director of the Dando Institute, claims replication, not randomized designs, should be viewed as the gold standard of criminological research.

General descriptions of crime science, together with many examples of applied research in crime problems, can be found on the Dando Institute’s website. See especially the sections on “Research Reports,” “Crime Science Series,” and “Crime Science: Short Reports.” Incidentally, the Dando Institute was named after Jill Dando, a famous television broadcaster in England, who was murdered in 2000. Her on-air colleague, Nick Ross, led a fund-raising effort to create an institute where crime science could be advanced. http://www.jdi.ucl.ac.uk/publications/

Additional Readings

Kim, So Young, and Skogan, Wesley G., “Statistical Analysis of Time Series Data on Problem Solving,” Community Policing Working Paper #27 (Center for Policy Research, Northwestern University, 2003) (http://www.northwestern.edu/ipr/publications/policing.html). Kim and Skogan present a number of time-series studies to examine the effects of problem solving by Chicago police. This is a good example of switching replications time-series designs by researchers at the university where Campbell and Cook did their pioneering work on quasi-experimental designs.

Pawson, Ray, and Tilley, Nick, Realistic Evaluation (Thousand Oaks, CA: Sage, 1997). We mentioned this book in Chapter 4. Pawson and Tilley argue that experiments and quasi-experiments focus too narrowly on threats to internal validity. Instead, they propose a different view of causation and different approaches to assessing cause.

Shadish, William R., Cook, Thomas D., and Campbell, Donald T., Experimental and Quasi-experimental Designs for Generalized Causal Inference (Boston: Houghton Mifflin, 2002). An update of the definitive guide to quasi-experimentation, this book focuses on basic principles of research design. In addition to numerous pointers on designing research, the authors stress that designing out validity threats is much preferred to trying to control them through later statistical analysis.

Weisburd, David, Lum, Cynthia M., and Petrosino, Anthony, “Does Research Design Affect Study Outcomes in Criminal Justice?” The Annals, Vol. 578 (2001), pp. 50–70. The authors make the intriguing claim that stronger experimental designs are more likely to find no causal relationships, whereas quasi-experimental designs more often find relationships. Read this article carefully (whether or not you complete the exercise described above) and decide whether you agree with the authors’ conclusions.

This page intentionally left blank
Having covered the basics of structuring research, from general issues to research design, let’s dive into the various observational techniques available for criminal justice research.

Chapter 8 examines how social scientists go about selecting people or things for observation. Our discussion of sampling addresses the fundamental scientific issue of generalizability. As we’ll see, it is possible for us to select a few people or things for observation and then apply what we observe to a much larger group of people or things than we actually observed. It is possible, for example, to ask 1000 people how they feel about using video cameras to enforce traffic laws and then accurately predict how tens of millions of people feel about it.

Chapter 9 describes survey research and other techniques for collecting data by asking people questions. We’ll cover different ways of asking questions and discuss the various uses of surveys and related techniques in criminal justice research.

Chapter 10, on field research, examines what is perhaps the most natural form of data collection: the direct observation of phenomena in natural settings. As we will see, observations can be highly structured and systematic (such as counting pedestrians who walk by some specified point) or less structured and more flexible.

Chapter 11 discusses ways to take advantage of some of the data available all around us. Researchers often examine data collected by criminal justice and other public agencies. Content analysis is a method of collecting data through carefully specifying and counting communications such as news stories, court opinions, or even recorded visual images. Criminal justice researchers may also conduct secondary analysis of data collected by others.
Chapter 8

Sampling

Sampling makes it possible to select a few hundred or thousand people for study and discover things that apply to many more people who are not studied.

Introduction  213
The Logic of Probability Sampling  213
Conscious and Unconscious Sampling Bias  215
Representativeness and Probability of Selection  216
Probability Theory and Sampling Distribution  217
The Sampling Distribution of 10 Cases  218
From Sampling Distribution to Parameter Estimate  221
Estimating Sampling Error  222
Confidence Levels and Confidence Intervals  223
Probability Theory and Sampling Distribution Summed Up  224
Probability Sampling  225
Populations and Sampling Frames  225
Simple Random Sampling  226
Systematic Sampling  227
Stratified Sampling  227
Disproportionate Stratified Sampling  228
Multistage Cluster Sampling  229
Multistage Cluster Sampling with Stratification  230
Illustration: Two National Crime Surveys  232
The National Crime Victimization Survey  232
The British Crime Survey  233
Probability Sampling in Review  234
Nonprobability Sampling  234
(continued)
Introduction

*How we collect representative data is fundamental to criminal justice research.*

Much of the value of research depends on how data are collected. A critical part of criminal justice research is deciding what will be observed and what won’t. If you want to study drug users, for example, which drug users should you study? This chapter discusses the logic and fundamental principles of sampling and then describes different general approaches for selecting subjects or other units.

Sampling is the process of selecting observations. Sampling is ordinarily used to select observations for one of two related reasons. First, it is often not possible to collect information from all persons or other units we wish to study. We may wish to know what proportion of all persons arrested in U.S. cities have recently used drugs, but collecting all that data would be virtually impossible. Thus, we have to look at a sample of observations.

The second reason for sampling is that it is often not necessary to collect data from all persons or other units. Probability sampling techniques enable us to make relatively few observations and then generalize from those observations to a much wider population. For example, if we are interested in what proportion of high school students have used marijuana, collecting data from a probability sample of a few thousand students will serve just as well as trying to study every high school student in the country.

Although probability sampling is central to criminal justice research, it cannot be used in many situations of interest. A variety of non-probability sampling techniques are available in such cases. Nonprobability sampling has its own logic and can provide useful samples for criminal justice inquiry. In this chapter, we examine both the advantages and the shortcomings of such methods, and we discuss where they fit in the larger picture of sampling and collecting data. Keep in mind one important goal of all sampling: to reduce, or at least to understand, potential biases that may be at work in selecting subjects.

The Logic of Probability Sampling

*Probability sampling helps researchers generalize from observed cases to unobserved ones.*

In selecting a group of subjects for study, social science researchers often use some type of sampling. Sampling in general refers to selecting part of a population. In selecting samples, we want to do two related things. First, we select samples to represent some larger population of people or other things. If we are interested in attitudes about a community correctional facility, for example, we might draw a sample of neighborhood residents, ask them some...
questions, and use their responses to represent the attitudes of all neighborhood residents. Or, in studying cases in a criminal court, we may not be able to examine all cases, so we select a sample to represent that population of all cases processed through some court.

Second, we may want to generalize from a sample to an unobserved population the sample is intended to represent. If we interview a sample of community residents, we may want to generalize our findings to all community residents—those we interviewed and those we did not. We might similarly expect that our sample of criminal court cases can be generalized to the population of all criminal court cases.

A special type of sampling that enables us to generalize to a larger population is known as **probability sampling**, a method of selection in which each member of a population has a known chance or probability of being selected. Knowing the probability that any individual member of a population could be selected makes it possible for us to make predictions that our sample accurately represents the larger population.

If all members of a population are identical in all respects—demographic characteristics, attitudes, experiences, behaviors, and so on—there is no need for careful sampling procedures. Any sample will be sufficient. In this extreme case of homogeneity, in fact, a single case will be sufficient as a sample to study characteristics of the whole population.

In reality, of course, the human beings who make up any real population are heterogeneous, varying in many ways. Figure 8.1 offers a simplified illustration of a heterogeneous population: The 100 members of this small population differ by gender and race. We'll use this hypothetical micropopulation to illustrate various aspects of sampling.

A sample of individuals from a population, if it is to provide useful descriptions of the to-
The sample population, must contain essentially the same variations that exist in the population. This is not as simple as it might seem. Let’s look at some of the possible biases in selection or ways researchers might go astray. Then we will see how probability sampling provides an efficient method for selecting a sample that should adequately reflect variations that exist in the population.

Conscious and Unconscious Sampling Bias

At first glance, it may seem as if sampling is a rather straightforward matter. To select a sample of 100 lawyers, a researcher might simply go to a courthouse and interview the first 100 lawyers who walk through the door. This kind of sampling method is often used by untrained researchers, but it is subject to serious biases. In connection with sampling, bias simply means that those selected are not “typical” or “representative” of the larger populations they have been chosen from. This kind of bias is virtually inevitable when a researcher picks subjects casually.

Figure 8.2 illustrates what can happen when we simply select people who are convenient for study. Although women make up only 50 percent of our micropopulation, those closest to the researcher (people in the upper right-hand corner of Figure 8.2) happen to be 70 percent women. Although the population is 12 percent African American, none were selected into this sample of people who happened to be conveniently situated near the researcher.

Moving beyond the risks inherent in simply studying people who are convenient, we need to consider other potential problems as well. To begin, our own personal leanings or biases may affect the sample selected in this manner; hence, the sample will not truly represent the population of lawyers. Suppose a researcher is a little intimidated by lawyers who look particularly prosperous, believing that they might ridicule his research effort. He might consciously
or unconsciously avoid interviewing them. Or he might believe that the attitudes of “establishment” lawyers are irrelevant to his research purposes and avoid interviewing them.

Even if the researcher seeks to interview a “balanced” group of lawyers, he won’t know the exact proportions of different types of lawyers who make up such a balance and won’t always be able to identify the different types merely by watching them walk by.

The researcher might make a conscious effort to interview, say, every 10th lawyer who enters the courthouse, but he still cannot be sure of a representative sample because different types of lawyers visit the courthouse with different frequencies, and some never go to the courthouse at all. Thus, the resulting sample will overrepresent lawyers who visit the courthouse more often.

Similarly, “call-in polls”—in which radio stations ask people to call specified telephone numbers to register their opinions—cannot be trusted to represent the general population. At the very least, not everyone in the population is even aware of the poll. Those who are aware of it have some things in common simply because they listen to the same radio station. As market researchers understand very well, a classical music station has a different audience than a hard rock station. Adding even more bias to the sample, those who are motivated to take part in the poll are probably different from others who are not so motivated.

A similar problem affects polls linked to web blogs or mass e-mail. Blogs tend to be selective; people regularly visit blogs that present views on personal and political issues they endorse (Hewitt, 2005). As a result, the population of people who respond to web blog polls can only represent the population of people who regularly visit individual blogs. As a general principle, the more self-selection is involved the more bias will be introduced into the sample.

The possibilities for inadvertent sampling bias are endless and not always obvious. Fortunately, some techniques can help us avoid bias.

Representativeness and Probability of Selection

Although the term representativeness has no precise, scientific meaning, it carries a commonsense meaning that makes it useful in the discussion of sampling. As we’ll use the term here, a sample is representative of the population from which it is selected if the aggregate characteristics of the sample closely approximate those same aggregate characteristics in the population. If the population, for example, contains 50 percent women, a representative sample will also contain “close to” 50 percent women. Later in this chapter, we’ll discuss “how close” in detail. Notice that samples need not be representative in all respects; representativeness is limited to those characteristics that are relevant to the substantive interests of the study. This is consistent with the more extensive discussion of what representative means in Shadish, Cook, and Campbell (2002:354–56).

A basic principle of probability sampling is that a sample will be representative of the population from which it is selected if all members of the population have an equal chance of being selected in the sample. Samples that have this quality are often labeled equal probability of selection method (EPSEM) samples. This principle forms the basis of probability sampling.

Even carefully selected EPSEM samples are seldom, if ever, perfectly representative of the populations from which they are drawn. Nevertheless, probability sampling offers two special advantages. First, probability samples, though never perfectly representative, are typically more representative than other types of samples because they avoid the biases discussed in the preceding section. In practice, there is a greater likelihood that a probability sample will be representative of the population from which it is drawn than a nonprobability sample will be.

Second, and more important, probability sampling permits us to estimate the accuracy or representativeness of the sample. Conceivably, a researcher might wholly by chance select
a sample that closely represents the larger population. The odds are against doing so, however, and we cannot estimate the likelihood that a haphazard sample will achieve representativeness. The probability sample can provide an accurate estimate of success or failure, because probability samples enable us to draw on probability theory.

**Probability Theory and Sampling Distribution**

Probability theory permits inferences about how sampled data are distributed around the value found in a larger population.

With a basic understanding of the logic of probability sampling in hand, we can examine how probability sampling works in practice. We will then be able to devise specific sampling techniques and assess the results of those techniques. To do so, we first need to understand four important concepts.

A **sample element** is that unit about which information is collected and that provides the basis of analysis. Typically, in survey research, elements are people or certain types of people. However, other kinds of units can be the elements for criminal justice research—correctional facilities, police beats, or court cases, for example. Elements and units of analysis are often the same in a given study, although the former refers to sample selection and the latter to data analysis.

A **population** is the theoretically specified grouping of study elements. Whereas the vague “delinquents” might be the target for a study, a more precise description of the population includes the definition of the element “delinquents” (for example, a person charged with a delinquent offense) and the time referent for the study (charged with a delinquent offense in the previous 6 months). Translating the abstract “adult drug addicts” into a workable population requires specifying the age that defines “adult” and the level of drug use that constitutes an “addict.” Specifying “college student” includes a consideration of full- and part-time students, degree and nondegree candidates, undergraduate and graduate students, and so on.

A **population parameter** is the value for a given variable in a population. The average income of all families in a city and the age distribution of the city’s population are parameters. An important portion of criminal justice research involves estimating population parameters on the basis of sample observations.

The summary description of a given variable in the sample is called a **sample statistic**. Sample statistics are used to make estimates of population parameters. Thus, the average income computed from a sample and the age distribution of that sample are statistics, and those statistics are used to estimate income and age parameters in a population.

The ultimate purpose of sampling is to select a set of elements from a population in such a way that descriptions of those elements (sample statistics) accurately portray the parameters of the total population from which the elements are selected. Probability sampling enhances the likelihood of accomplishing this aim and also provides methods for estimating the degree of probable success.

The key to this process is random selection. In random selection, each element has an equal chance of being selected independent of any other event in the selection process. Flipping a coin is the most frequently cited example: The “selection” of a head or a tail is independent of previous selections of heads or tails.

There are two reasons for using random selection methods. First, this procedure serves as a check on conscious or unconscious bias on the part of the researcher. The researcher who selects cases on an intuitive basis might choose cases that will support his or her research expectations or hypotheses. Random selection erases this danger. Second, and more important, with random selection we can draw on probability theory, which allows us to estimate population parameters and to estimate how accurate our statistics are likely to be.
The Sampling Distribution of 10 Cases

Suppose there are 10 people in a group, and each has a certain amount of money in his or her pocket. To simplify, let’s assume that one person has no money, another has $1, another has $2, and so forth up to the person who has $9. Figure 8.3 illustrates this population of 10 people.¹

Our task is to determine the average amount of money one person has—specifically, the mean number of dollars. If you simply add up the money shown in Figure 8.3, the total is $45, so the mean is $4.50 (45/10). Our purpose in the rest of this example is to estimate that mean without actually observing all 10 individuals. We’ll do that by selecting random samples from the population and using the means of those samples to estimate the mean for the whole population.

To start, suppose we select—at random—a sample of only 1 person from the 10. Depending on which person we select, we will estimate the group’s mean as anywhere from $0 to $9. Figure 8.4 shows a display of those 10 possible samples. The 10 dots shown on the graph represent the 10 “sample” means we will get as estimates of the population. The range of the dots on the graph is the sampling distribution, defined as the range of sample statistics we will obtain if we select many samples. Figure 8.4 shows how all of our possible samples of 1 are distributed. Obviously, it is not a good idea to select a sample of only 1 because we stand a good chance of missing the true mean of $4.50 by quite a bit.

What if we take samples of 2 each? As you can see from Figure 8.5, increasing the sample size improves our estimations. Each dot represents a possible sample. There are 45 possible samples of two elements: $0/$1, $0/$2, . . . , $7/$8, $8/$9. Moreover, some of these samples produce the same means. For example, $0/$6, $1/$5, and $2/$4 all produce means of $3. In Figure 8.5, the three dots shown above the $3 mean represent those 3 samples.

Notice that the means we get from the 45 samples are not evenly distributed. Rather, ¹We thank Hanan Selvin for suggesting this way of introducing probability sampling.
Figure 8.4  The Sampling Distribution of Samples of 1

Figure 8.5  The Sampling Distribution of Samples of 2
Figure 8.6  The Sampling Distribution of Samples of 3, 4, 5, and 6
they are somewhat clustered around the true value of $4.50. Only 2 samples deviate by as much as $4 from the true value ($0/$1 and $8/$9), whereas 5 of the samples give the true estimate of $4.50, and another 8 samples miss the mark by only $.50 (plus or minus).

Now suppose we select even larger samples. What will that do to our estimates of the mean? Figure 8.6 presents the sampling distributions of samples of 3, 4, 5, and 6. The progression of the sampling distributions is clear. Every increase in sample size improves the distribution of estimates of the mean in two related ways. First, in the distribution for samples of 5, for example, no sample means are at the extreme ends of the distribution. Why? Because it is not possible to select five elements from our population and obtain an average of less than $2 or greater than $7. The second way sampling distributions improve with larger samples is that sample means cluster more and more around the true population mean of $4.50. Figure 8.6 clearly shows this tendency.

**From Sampling Distribution to Parameter Estimate**

Let's turn now to a more realistic sampling situation and see how the notion of sampling distribution applies. Assume that we wish to study the population of Placid Coast, California, to assess the levels of approval or disapproval of a proposed law to ban possession of handguns within the city limits.

Our target population is all adult residents. In order to draw an actual sample, we need some sort of list of elements in our population; such a list is called a *sampling frame*. Assume our sampling frame is a voter registration list of, say, 20,000 registered voters in Placid Coast. The elements are the individual registered voters.

The variable under consideration is attitudes toward the proposed law: approve and disapprove. Measured in this way, attitude toward the law is a *binomial variable*; it can have only two values. We’ll select a random sample of, say, 100 persons for the purpose of estimating the population parameter for approval of the proposed law.

Figure 8.7 presents all the possible values of this parameter in the population—from 0 percent approval to 100 percent approval. The midpoint of the line—50 percent—represents half the voters approving of the handgun ban and the other half disapproving.

To choose our sample, we assign each person on the voter registration list a number and use a computer program to generate 100 random numbers. Then we interview the 100 people whose numbers have been selected and ask for their attitudes toward the handgun ban: whether they approve or disapprove. Suppose this operation gives us 48 people who approve of the law and 52 who disapprove. We present this statistic by placing a dot at the point representing 48 percent, as shown in Figure 8.8.

Now suppose we select another sample of 100 people in exactly the same fashion and measure their approval or disapproval of the proposed law. Perhaps 51 people in the second sample approve of the law. We place another dot in the appropriate place on the line in Figure 8.8. Repeating this process once more, we may discover that 52 people in the third sample approve of the handgun ban; we add a third dot to Figure 8.8.

Figure 8.8 now presents the three different sample statistics that represent the percentages
of people in each of the three random samples who approved of the proposed law. Each of the random samples, then, gives us an estimate of the percentage of people in the total population of registered voters who approve of the handgun law. Unfortunately, we now have three separate estimates.

To rescue ourselves from this dilemma, let’s draw more and more samples of 100 registered voters each, question each of the samples concerning their approval or disapproval, and plot the new sample statistics on our summary graph. In drawing many such samples, we discover that some of the new samples provide duplicate estimates, as in our earlier illustration with 10 cases. Figure 8.9 shows the sampling distribution of hundreds of samples. This is often referred to as a normal or bell-shaped curve.

Notice that by increasing the number of samples selected and interviewed we have also increased the range of estimates provided by the sampling operation. In one sense, we have increased our dilemma in attempting to find the parameter in the population. Fortunately, probability theory provides certain important rules about the sampling distribution shown in Figure 8.9.

**Estimating Sampling Error**

Probability theory can help resolve our dilemma with some basic statistical concepts. First, if many independent random samples are selected from a population, then the sample statistics provided by those samples will be distributed around the population parameter in a known way. Thus, although Figure 8.9 shows a wide range of estimates, more of them are in the vicinity of 50 percent than elsewhere in the graph. Probability theory tells us, then, that the true value is in the vicinity of 50 percent.

Second, probability theory gives us a formula for estimating how closely the sample statistics are clustered around the true value:

\[
 s = \sqrt{\frac{P \times Q}{n}}
\]
where $s$ is the **standard error**—defined as a measure of sampling error—as is the number of cases in each sample, and $P$ and $Q$ are the population parameters for the binomial. If 60 percent of registered voters approve of the ban on handguns and 40 percent disapprove, then $P$ and $Q$ are 60 percent and 40 percent, or .6 and .4, respectively.

To see how probability theory makes it possible for us to estimate sampling error, suppose that in reality 50 percent of the people approve of the proposed law and 50 percent disapprove. These are the population parameters we are trying to estimate with our samples. Recall that we have been selecting samples of 100 cases each. When these numbers are plugged into the formula, we get:

\[ s = \sqrt{\frac{5 \times .5}{100}} = .05 \]

The standard error equals .05, or 5 percent.

In probability theory, the standard error is a valuable piece of information because it indicates how closely the sample estimates will be distributed around the population parameter. The standard error tells us how sample statistics will be dispersed or clustered around a population parameter. Probability theory tells us that approximately 34 percent (.3413) of the sample estimates will fall within one standard error increment above the population parameter, and another 34 percent will fall within one standard error increment below the parameter. In our example, the standard error increment is 5 percent, so we know that 34 percent of our samples will give estimates of approval between 50 percent (the parameter) and 55 percent (one standard error above); another 34 percent of the samples will give estimates between 50 percent and 45 percent (one standard error below the parameter). Taken together, then, we know that roughly two-thirds (68 percent) of the samples will give estimates between 45 and 55 percent, which is within 5 percent of the parameter.

The standard error is also a function of the sample size—an inverse function. This means that as the sample size increases the standard error decreases. And as the sample size increases, the several samples will be clustered nearer to the true value. Figure 8.6 illustrates this clustering. Another rule of thumb is evident in the formula for the standard error: Because of the square root operation, the standard error is reduced by half if the sample size is quadrupled. In our example, samples of 100 produce a standard error of 5 percent; to reduce the standard error to 2.5 percent, we would have to increase the sample size to 400.

All of this information is provided by established probability theory in reference to the selection of large numbers of random samples. If the population parameter is known and many random samples are selected, probability theory allows us to predict how many of the samples will fall within specified intervals from the parameter.

Of course, this discussion illustrates only the logic of probability sampling. It does not describe the way research is actually conducted. Usually, we do not know the parameter; we conduct a sample survey precisely because we want to estimate that value. Moreover, we don’t actually select large numbers of samples; we select only one sample. What probability theory does is provide the basis for making inferences about the typical research situation. Knowing what it would be like to select thousands of samples allows us to make assumptions about the one sample we do select and study.

**Confidence Levels and Confidence Intervals**

Probability theory specifies that 68 percent of that fictitious large number of samples will produce estimates that fall within one standard error of the parameter. As researchers, we can turn the logic around and infer that any single random sample has a 68 percent chance of falling within that range. In this regard, we speak of **confidence levels**: We are 68 percent confident that our sample estimate is within
one standard error of the parameter. Or we may say that we are 95 percent confident that the sample statistic is within two standard errors of the parameter, and so forth. Quite reasonably, our confidence level increases as the margin for error is extended. We are virtually positive (99.9 percent) that our statistic is within three standard errors of the true value.

Although we may be confident (at some level) of being within a certain range of the parameter, we seldom know what the parameter is. To resolve this dilemma, we substitute our sample estimate for the parameter in the formula; lacking the true value, we substitute the best available guess.

The result of these inferences and estimations is that we are able to estimate a population parameter and also the expected degree of error on the basis of one sample drawn from a population. We begin with this question: What percentage of the registered voters in Placid Coast approve of the proposed handgun ban? We select a random sample of 100 registered voters and interview them. We might then report that our best estimate is that 50 percent of registered voters approve of the gun ban and that we are 95 percent confident that between 40 and 60 percent (plus or minus two standard errors) approve. The range from 40 to 60 percent is called the confidence interval. At the 68 percent confidence level, the confidence interval is 45–55 percent.

The logic of confidence levels and confidence intervals also provides the basis for determining the appropriate sample size for a study. Once we decide on the sampling error we can tolerate, we can calculate the number of cases needed in our sample.

**Probability Theory and Sampling Distribution Summed Up**

This, then, is the basic logic of probability sampling. Random selection permits the researcher to link findings from a sample to the body of probability theory so as to estimate the accuracy of those findings. All statements of accuracy in sampling must specify both a confidence level and a confidence interval. The researcher must report that he or she is x percent confident that the population parameter is between two specific values.

In this example, we have demonstrated the logic of sampling error using a binomial variable—a variable analyzed in percentages. A different statistical procedure would be required to calculate the standard error for a mean, for example, but the overall logic is the same.

Notice that nowhere in this discussion did we consider the size of the population being studied. This is because the population size is almost always irrelevant. A sample of 2000 respondents drawn properly to represent residents of Vermont will be no more accurate than a sample of 2000 drawn properly to represent residents in the United States, even though the Vermont sample would be a substantially larger proportion of that small state's residents than would the same number chosen to represent the nation's residents. The reason for this counterintuitive fact is that the equations for calculating sampling error assume that the populations being sampled are infinitely large, so all sample would equal zero percent of the whole.

Although this is not literally true in practice, a sample of 2000 represents only 0.32 percent of people living in Vermont in 2004, and a sample of 2000 U.S. residents represents a mere 0.0007 percent of the national population in 2004. Unless a sample represents, say, 5 percent or more of the population it is drawn from, that proportion is irrelevant. In those rare cases of large proportions being selected, a “finite population correction” can be calculated to adjust the confidence intervals. Simply subtract the proportion from 1.0 and multiply the result times the sampling error. As you can see, with proportions close to zero, this will make no difference. If, on the other hand, your sample were half of the population, the sampling error would be cut in half by this procedure. In the extreme, if you included the whole
population in your sample, the sample-to-
population proportion would be 1.0, and you
would multiple the calculated standard error
by 0.0—suggesting there was no sampling er-
ror, which would, of course, be the case. (How
cool is that?)

Two cautions are in order before we con-
clude this discussion of the basic logic of prob-
ability sampling. First, the survey uses of prob-
ability theory as discussed here are technically
not wholly justified. The theory of sampling
distribution makes assumptions that almost
never apply in survey conditions. The exact
proportion of samples contained within speci-
fied increments of standard errors, for example,
mathematically assumes an infinitely large
population, an infinite number of samples,
and sampling with replacement—that is, every
sampling unit selected is “thrown back into the
pot” and could be selected again. Second, our
discussion has greatly oversimplified the in-
ferential jump from the distribution of several
samples to the probable characteristics of one
sample.

We offer these cautions to provide perspec-
tive on the uses of probability theory in sam-
ping. Researchers in criminal justice and other
social sciences often appear to overestimate
the precision of estimates produced by the use
of probability theory. Variations in sampling
techniques and nonsampling factors may fur-
ther reduce the legitimacy of such estimates.
For example, those selected in a sample who fail
or refuse to participate further detract from the
representativeness of the sample.

Nevertheless, the calculations discussed in
this section can be extremely valuable to you
in understanding and evaluating your data. Al-
though the calculations do not provide as
precise estimates as some researchers might as-
sume, they can be quite valid for practical pur-
poses. They are unquestionably more valid than
less rigorously derived estimates based on less
rigorous sampling methods. Most important,
being familiar with the basic logic underlying
the calculations can help you react sensibly
both to your own data and to those reported
by others.

**Probability Sampling**

*Different types of probability sampling designs can be used alone or in combination for different re-
search purposes.*

As researchers and as consumers of research we
need to understand the theoretical foundations
of sampling. It is no less important to appreci-
ate the less-than-perfect conditions that exist
in the field. Most of we have considered so far
assumes we are using simple random sampling.
In reality researchers have a number of options
in choosing their sampling method, each with
its own advantages and disadvantages.

**Populations and Sampling Frames**

Our discussion begins with a more practical
consideration of one key feature of all prob-
ability sampling designs: the relationship be-
tween populations and sampling frames.

A sampling frame is the list or quasi-list of
elements from which a probability sample is
selected. Put another way, a sampling frame is
a list or quasi-list of our target population. We
say “quasi-list” because, even though an actual
list might not exist, we can draw samples as if
there were a list. Properly drawn samples pro-
vide information appropriate for describing
the population of elements that compose the
sampling frame—nothing more. This point is
important in view of the common tendency for
researchers to select samples from a particu-
lar sampling frame and then make assertions
about a population that is similar, but not
identical, to the study population defined by
the sampling frame.

For example, if we want to study the atti-
dudes of corrections administrators toward de-
terminant sentencing policies, we might select
a sample by consulting the membership roster
of the American Correctional Association. In
this case, the membership roster is our sam-
ping frame, and corrections administrators
are the population we wish to describe. However, unless all corrections administrators are members of the American Correctional Association and all members are listed in the roster, it would be incorrect to generalize results to all corrections administrators.

Studies of organizations are often the simplest from a sampling standpoint because organizations typically have membership lists. In such cases, the list of members may be an acceptable sampling frame. If a random sample is selected from a membership list, then the data collected from that sample may be taken as representative of all members—if all members are included in the list. It is, however, imperative that researchers learn how complete or incomplete such lists might be and limit their generalizations to listed sample elements rather than to an entire population.

Other lists of individuals may be especially relevant to the research needs of a particular study. Lists of licensed drivers, automobile owners, welfare recipients, taxpayers, holders of weapons permits, and licensed professionals are just a few examples. Although it may be difficult to gain access to some of these lists, they provide excellent sampling frames for specialized research purposes.

Telephone directories are frequently used for “quick and dirty” public opinion polls. Undeniably, they are easy and inexpensive to use, and that is no doubt the reason for their popularity. Still, they have several limitations. A given directory will not include new subscribers or those who have requested unlisted numbers. Sampling is further complicated by the inclusion of nonresidential listings in directories. Moreover, telephone directories are sometimes taken to be a listing of a city’s population, which is simply not the case. Lower-income people are less likely to have telephones, and higher-income people may have more than one line. A growing number of households are served only by wireless phone service and so are not listed in directories. A recent national study reported that 7 percent of households had only wireless phones, and 2 percent had no telephone service (Blumberg, Luke, and Cynamon, 2006). Telephone companies may not publish listings for temporary residents such as students. And persons who live in institutions or group quarters—dormitories, nursing homes, rooming houses, and the like—are not listed in phone directories.

Street directories and tax maps may be used for as a sampling frame for households, but they may also suffer from incompleteness and possible bias. For example, illegal housing units such as basement apartments are unlikely to appear on official records. As a result, such units have no chance for selection, and sample findings will not be representative of those units, which are often substandard and overcrowded.

In a more general sense, it’s worth viewing sampling frames as operational definitions of a study population. Just as operational definitions of variables describe how abstract concepts will be measured, sampling frames serve as a real-world version of an abstract study population. For example, we may want to study how criminologists deal with ethical issues in their research. We don’t know how many criminologists exist out there, but we can develop a general idea about the population of criminologists. We could also operationalize the concept by using the membership directory for the American Society of Criminology—that list is our operational definition of the population of criminologists.

**Simple Random Sampling**

*Simple random sampling* forms the basis of probability theory and the statistical tools we use to estimate population parameters, standard error, and confidence intervals. More accurately, such statistics assume unbiased sampling, and simple random sampling is the foundation of unbiased sampling.

Once a sampling frame has been established, a simple random sample can be produced by assigning a single number to each element in the
frame, not skipping any number in the process. A table of random numbers, or a computer program for generating them, is then used to select elements for the sample.

If the sampling frame is a computerized database or some other form of machine-readable data, a simple random sample can be selected automatically by computer. In effect, the computer program numbers the elements in the sampling frame, generates its own series of random numbers, and prints out the list of elements selected.

**Systematic Sampling**

Simple random sampling is seldom used in practice, primarily because it is not usually the most efficient method, and it can be tedious if done manually. It typically requires a list of elements. And when such a list is available, researchers usually use systematic sampling rather than simple random sampling.

In systematic sampling, the researcher chooses all elements in the list for inclusion in the sample. If a list contains 10,000 elements and we want a sample of 1000, we select every 10th element for our sample. To ensure against any possible human bias, we should select the first element at random. Thus, to systematically select 1000 from a list of 10,000 elements, we begin by selecting a random number between 1 and 10. The element having that number, plus every 10th element following it, is included in the sample. This method technically is referred to as a systematic sample with a random start.

In practice, systematic sampling is virtually identical to simple random sampling. If the list of elements is indeed randomized before sampling, one might argue that a systematic sample drawn from that list is, in fact, a simple random sample.

Systematic sampling has one danger. A periodic arrangement of elements in the list can make systematic sampling unwise; this arrangement is usually called “periodicity.” If the list of elements is arranged in a cyclical pattern that coincides with the sampling interval, a biased sample may be drawn. Suppose we select a sample of apartments in an apartment building. If the sample is drawn from a list of apartments arranged in numerical order (for example, 101, 102, 103, 104, 201, 202, and so on), there is a danger of the sampling interval coinciding with the number of apartments on a floor or some multiple of it. Then the samples might include only northwest-corner apartments or only apartments near the elevator. If these types of apartments have some other particular characteristic in common (for example, higher rent), the sample will be biased. The same potential danger would apply in a systematic sample of houses in a subdivision arranged with the same number of houses on a block.

In considering a systematic sample from a list, then, we need to carefully examine the nature of that list. If the elements are arranged in any particular order, we have to figure out whether that order will bias the sample to be selected and take steps to counteract any possible bias.

In summary, systematic sampling is usually superior to simple random sampling, in terms of convenience if nothing else. Problems in the ordering of elements in the sampling frame can usually be remedied quite easily.

**Stratified Sampling**

We have discussed two methods of selecting a sample from a list: random and systematic. Stratification is not an alternative to these methods, but it represents a possible modification in their use. Simple random sampling and systematic sampling both ensure a degree of representativeness and permit an estimate of the sampling error present. Stratified sampling is a method for obtaining a greater degree of representativeness—decreasing the probable sampling error. To understand why that is the case, we must return briefly to the basic theory of sampling distribution.

Recall that sampling error is reduced by two factors in the sample design: (1) A large sample produces a smaller sampling error than a small
sample does, and (2) a homogeneous population produces samples with smaller sampling errors than a heterogeneous population does. If 99 percent of the population agrees with a certain statement, it is extremely unlikely that any probability sample will greatly misrepresent the extent of agreement. If the population is split 50-50 on the statement, then the sampling error will be much greater.

Stratified sampling is based on this second factor in sampling theory. Rather than selecting our sample from the total population at large, we select appropriate numbers of elements from homogeneous subsets of that population. To get a stratified sample of university students, for example, we first organize our population by college class and then draw appropriate numbers of freshmen, sophomores, juniors, and seniors. In a nonstratified sample, representation by class is subject to the same sampling error as other variables. In a sample stratified by class, the sampling error on this variable is reduced to zero.

Even more complex stratification methods are possible. In addition to stratifying by class, we might also stratify by gender, grade-point average, and so forth. In this fashion, we could ensure that our sample contains the proper numbers of freshman men with a 4.0 average, freshman women with a 4.0 average, and so forth.

The ultimate function of stratification, then, is to organize the population into homogeneous subsets (with heterogeneity between subsets) and to select the appropriate number of elements from each. To the extent that the subsets are homogeneous on the stratification variables, they may be homogeneous on other variables as well. Because age is usually related to college class, a sample stratified by class will be more representative in terms of age as well.

The choice of stratification variables typically depends on what variables are available. Gender can often be determined in a list of names. Many local government sources of information on housing units are arranged geographically. Age, race, education, occupation, and other variables are often included on lists of persons who have had contact with criminal justice officials.

In selecting stratification variables, however, we should be concerned primarily with those that are presumably related to the variables we want to represent accurately. Because gender is related to many variables and is often available for stratification, it is frequently used. Age and race are related to many variables of interest in criminal justice research. Income is also related to many variables, but it is often not available for stratification. Geographic location within a city, state, or nation is related to many things. Within a city, stratification by geographic location usually increases representativeness in social class and ethnicity.

Stratified sampling ensures the proper representation of the stratification variables to enhance representation of other variables related to them. Taken as a whole, then, a stratified sample is likely to be more representative on a number of variables than a simple random sample is.

Disproportionate Stratified Sampling

Another use of stratification is to purposively produce samples that are not representative of a population on some variable, referred to as disproportionate stratified sampling. Because the purpose of sampling, as we have been discussing, is to represent a larger population, you may wonder why anyone would want to intentionally produce a sample that was not representative.

To understand the logic of disproportionate stratification, consider again the role of population homogeneity in determining sample size. If members of a population vary widely on some variable of interest, then larger samples must be drawn to adequately represent that population. Similarly, if only a small number of people in a population exhibit some attribute
or characteristic of interest, then a large sample must be drawn to produce adequate numbers of elements that exhibit the uncommon condition. Disproportionate stratification is a way of obtaining sufficient numbers of these “rare” cases by selecting a number disproportionate to their representation in the population.

The best example of disproportionate sampling in criminal justice is a national crime survey in which one goal is to obtain some minimum number of crime victims in a sample. Because crime victimization for certain offenses—such as robbery or aggravated assault—is relatively rare on a national scale, persons who live in large urban areas, where serious crime is more common, are disproportionately sampled.

The British Crime Survey (BCS) is a nationwide survey of people ages 16 and over in England and Wales. Over its first 20 years (since 1982) the BCS selectively oversampled people or areas to yield larger numbers of designated subjects than would result from proportionate random samples of the population of England and Wales. The BCS conducted in 2000 included special questions to better understand contacts between ethnic minorities and police, and ethnic minorities were disproportionately oversampled to produce a large enough number of ethnic minority respondents for later analysis (Kershaw, Budd, Kinshott, et al., 2000).

**Multistage Cluster Sampling**

The preceding sections have described reasonably simple procedures for sampling from lists of elements. Unfortunately, however, many interesting research problems require the selection of samples from populations that cannot easily be listed for sampling purposes; that is, sampling frames are not readily available. Examples are the population of a city, state, or nation, and all police officers in the United States. In such cases, the sample design must be much more complex. Such a design typically involves the initial sampling of groups of elements—clusters—followed by the selection of elements within each of the selected clusters. This procedure is called multistage cluster sampling.

Cluster sampling may be used when it is either impossible or impractical to compile an exhaustive list of the elements that compose the target population, such as all law enforcement officers in the United States. Often, however, population elements are already grouped into subpopulations, and a list of those subpopulations either exists or can be created.

Population elements, or aggregations of those elements, are referred to as sampling units. In the simplest forms of sampling, elements and units are the same thing—usually, people. But in cases in which a listing of elements is not available, we can often use some other unit that includes a grouping of elements.

Because U.S. law enforcement officers are employed by individual cities, counties, or states, it is possible to create lists of those political units. For cluster sampling, then, we could sample the list of cities, counties, and states in some manner as discussed previously (for example, a systematic sample stratified by population). Next, we could obtain lists of law enforcement officers from agencies in each of the selected jurisdictions. We could then sample each of the lists to provide samples of police officers for study.

Another typical situation concerns sampling among population areas such as a city. Although there is no single list of a city’s population, citizens reside on discrete city blocks or census blocks. It is possible, therefore, to select a sample of blocks initially, create a list of persons who live on each of the selected blocks, and then sample persons from that list. In this case, blocks are treated as the primary sampling unit.

In a more complex design, we might sample blocks, list the households on each selected block, sample the households, list the persons who reside in each household, and, finally, sample persons within each selected household. This multistage sample design will lead to the
ultimate selection of a sample of individuals without requiring the initial listing of all individuals in the city's population.

Multistage cluster sampling, then, involves the repetition of two basic steps: listing and sampling. The list of primary sampling units (city blocks) is compiled and perhaps stratified for sampling. Next, a sample of those units is selected. The list of secondary sampling units is then sampled, and the process continues.

Cluster sampling is highly recommended for its efficiency, but the price of that efficiency is a less accurate sample. A simple random sample drawn from a population list is subject to a single sampling error, but a two-stage cluster sample is subject to two sampling errors. First, the initial sample of clusters represents the population of clusters only within a range of sampling error. Second, the sample of elements selected within a given cluster represents all the elements in that cluster only within a range of sampling error. Thus, for example, we run a certain risk of selecting a sample of disproportionately wealthy city blocks, plus a sample of disproportionately wealthy households within those blocks. The best solution to this problem involves the number of clusters selected initially and the number of elements selected within each cluster.

Recall that sampling error is reduced by two factors: (1) an increase in the sample size and (2) increased homogeneity of the elements being sampled. These factors operate at each level of a multistage sample design. A sample of clusters will best represent all clusters if a large number are selected and if all clusters are very much alike. A sample of elements will best represent all elements in a given cluster if a large number are selected from the cluster and if all the elements in the cluster are very much alike.

A good general guideline for cluster design is to maximize the number of clusters selected while decreasing the number of elements within each cluster. But this scientific guideline must be balanced against an administrative constraint. The efficiency of cluster sampling is based on the ability to minimize the list of population elements. By initially selecting clusters, we need only list the elements that make up the selected clusters, not all elements in the entire population. Increasing the number of clusters, however, reduces this efficiency in cluster sampling. A small number of clusters may be listed more quickly and more cheaply than a large number. Remember that all the elements in a selected cluster must be listed even if only a few are to be chosen in the sample.

The final sample design will reflect these two constraints. In effect, we will probably select as many clusters as we can afford. So as not to leave this issue too open-ended, here is a rule of thumb: Population researchers conventionally aim for the selection of 5 households per census block. If a total of 2000 households are to be interviewed, researchers select 400 blocks with 5 household interviews on each. Figure 8.10 presents a graphic overview of this process.

As we turn to more detailed procedures in cluster sampling, keep in mind that this method almost inevitably involves a loss of accuracy. First, as noted earlier, a multistage sample design is subject to a sampling error at each stage. Because the sample size is necessarily smaller at each stage than the total sample size, the sampling error at each stage will be greater than would be the case for a single-stage random sample of elements. Second, sampling error is estimated on the basis of observed variance among the sample elements. When those elements are drawn from relatively homogeneous clusters, the estimated sampling error will be too optimistic and so must be corrected in light of the cluster sample design.

Multistage Cluster Sampling with Stratification

Thus far, we have looked at cluster sampling as though a simple random sample were selected at each stage of the design. In fact, we can use stratification techniques to refine and improve the sample being selected. The basic options
Stage One: Identify blocks and select a sample. (Selected blocks are shaded.)

Stage Two: Go to each selected block and list all households in order. (Example of one listed block.)

1. 491 Rosemary Ave.
2. 487 Rosemary Ave.
3. 473 Rosemary Ave.
4. 455 Rosemary Ave.
5. 437 Rosemary Ave. **
6. 423 Rosemary Ave.
7. 411 Rosemary Ave.
8. 403 Rosemary Ave.
9. 1101 4th St.
10. 1123 4th St.
11. 1137 4th St. **
12. 1157 4th St.
13. 1169 4th St.
14. 1187 4th St.
15. 402 Thyme Ave.
16. 408 Thyme Ave.
17. 424 Thyme Ave. **
18. 446 Thyme Ave.
19. 458 Thyme Ave.
20. 480 Thyme Ave.
21. 498 Thyme Ave.
22. 1186 5th St.
23. 1174 5th St. **
24. 1160 5th St.
25. 1140 5th St.
26. 1122 5th St.
27. 1118 5th St.
28. 1116 5th St.
29. 1104 5th St. **
30. 1102 5th St.

Stage Three: For each list, select a sample of households. (In this example, every sixth household has been selected starting with #5, which was selected at random.)

Figure 8.10 Multistage Cluster Sampling
available are essentially the same as those possible in single-stage sampling from a list. In selecting a national sample of law enforcement officers, for example, we might initially stratify our list of agencies by type (state, county, municipal), geographic region, size, and rural or urban location.

Once the primary sampling units (law enforcement agencies) have been grouped according to the relevant, available stratification variables, either simple random or systematic sampling techniques can be used to select the sample. We might select a specified number of units from each group or stratum, or we might arrange the stratified clusters in a continuous list and systematically sample that list.

To the extent that clusters are combined into homogeneous strata, the sampling error at this stage will be reduced. The primary goal of stratification, as before, is homogeneity.

In principle, stratification can take place at each level of sampling. The elements listed within a selected cluster might be stratified before the next stage of sampling. Typically, however, that is not done because we strive for relative homogeneity within clusters. If clusters are sufficiently similar, it is not necessary to stratify again.

Illustration: Two National Crime Surveys

Two national crime surveys show different ways of designing samples to achieve desired results.

Our discussion of sampling designs suggests that researchers can combine many different techniques of sampling and their various components in different ways to suit various needs. In fact, the different components of sampling can be tailored to specific purposes in much the same way that research design principles can be modified to suit various needs. Because sample frames suitable for simple random sampling are often unavailable, researchers use multistage cluster sampling to move from aggregate sample units to actual sample elements. We can add stratification to ensure that samples are representative of important variables. And we can design samples to produce elements that are proportionate or disproportionate to the population.

Two national crime surveys illustrate how these various building blocks may be combined in complex ways: (1) the National Crime Victimization Survey (NCVS), conducted by the U.S. Census Bureau; and (2) the British Crime Survey (BCS). Each is a multistage cluster sample, but the two surveys use different strategies for sampling to produce sufficient numbers of respondents in different categories. Our summary description is adapted from BJS (2006) for the NCVS and from Keith Bolling and associates (2004) for the BCS.

The National Crime Victimization Survey

Although various parts of the NCVS have been modified since the surveys were begun in 1972, the basic sampling strategies have remained relatively constant. The most significant changes have been fluctuations in sample size and a shift to telephone interviewing, with samples of telephone number listings eventually leading to households.

The NCVS seeks to represent the nationwide population of persons ages 12 and over who are living in households. We noted in Chapter 6 that the phrase “living in households” is significant; this is especially true in our current discussion of sampling. NCVS procedures are not designed to sample homeless persons or people who live in institutional settings such as military group housing, temporary housing, or correctional facilities. Also, because the sample targets persons who live in households, it cannot provide estimates of crimes in which a commercial establishment or business is the victim.

Because there is no national list of households in the United States, multistage cluster sampling must be used to proceed from larger
units to households and their residents. The national sampling frame used in the first stage defines primary sampling units (PSUs) as large metropolitan areas, nonmetropolitan counties, or groups of contiguous counties (to represent rural areas).

The largest 93 PSUs are specified as “self-representing” and are automatically included in the first stage of sampling. The remaining PSUs are stratified by size, population density, reported crimes, and other variables. An additional 110 non-self-representing PSUs are then selected with a probability proportionate to the population of the PSU. Thus, for example, if one stratum includes Bugtussle, Texas (population 7000), Punkinseed, Indiana (5000), and Rancid, Missouri (3000), the probability that each PSU will be selected is: 7 in 15 for Bugtussle, 5 in 15 for Punkinseed, and 3 in 15 for Rancid.

The second stage of sampling involves designating four different sampling frames within each PSU. Each of these frames is used to select different types of subsequent units. First, the housing unit frame lists addresses of housing units from census records. Second, a group quarters frame lists group quarters such as dormitories and rooming houses from census records. Third, a building permit frame lists newly constructed housing units from local government sources. Finally, an area frame lists census blocks (physical geographic units), from which independent address lists are generated and sampled. Notice that these four frames are necessary because comprehensive, up-to-date lists of residential addresses are not available in this country.

For the 2004 NCVS, these procedures yielded a sample of approximately 42,000 housing units. Completed interviews were obtained from about 74,500 individuals living in households. The sample design for the NCVS is an excellent illustration of the relationship between sample size and variation in the target population. Because serious crime is a relatively rare event when averaged across the entire U.S. population, very large samples must be drawn. And because no single list of the target population exists, samples are drawn in several stages.

For further information, consult NCVS documentation maintained by the Bureau of Justice Statistics (http://www.ojp.usdoj.gov/bjs/cvictgen.htm). Also see the “National Crime Victimization Survey Resource Guide,” maintained at the National Archive of Criminal Justice Data (http://www.icpsr.umich.edu/nacjd/ncvs).

The British Crime Survey

We have seen that NCVS sampling procedures begin with demographic units and work down to selection of housing units. BCS sampling is simplified by the existence of a national list of something close to addresses. The Postcode Address File (PAF) lists postal delivery points nationwide and is further subdivided to distinguish “small users,” those addresses receiving less than 50 items per day. Even though 50 pieces of mail might still seem like quite a bit, this classification makes it easier to distinguish household addresses from commercial ones.

Postcode sectors, roughly corresponding to U.S. five-digit zip codes, are easily defined clusters of addresses from the PAF. Samples of addresses are then selected from within these sectors. In most cases, 32 addresses are selected from within the postcode.

In addition, BCS researchers devised “booster samples” to increase the number of respondents who were ethnic minorities or ages 16–24. Victimization experiences of ethnic minorities were of special interest to police and other public officials. Young people were oversampled to complete a special questionnaire of self-report behavior items.

The ethnic minority booster was accomplished by first selecting respondents using formal sampling procedures. Interviewers then sought information about four housing units adjacent to the selected unit in an effort to determine if any residents were nonwhite. If adjacent units housed minority families, one
was selected to be interviewed for the ethnic minority booster sample. This is an example of what Steven Thompson (1997) calls “adaptive sampling.” Probability samples are first selected, then those respondents are used to identify other individuals who meet some criterion. Increasing the number of respondents ages 16–24 was simpler—interviewers sought additional respondents in that age group within sampled households.

One final sampling dimension reflects the regional organization of police in England and Wales into 43 police areas. The BCS was further stratified to produce 600–700 interviews in each police area to support analysis within those areas.

Apart from the young-person booster, once individual households are selected, one person age 16 or over was randomly chosen to provide information for all household members. Sampling procedures initially produced about 54,700 addresses for the year 2004 BCS. About 8 percent of these were eliminated because they were vacant, had been demolished, or contained a business, not a private household. Of the remaining 50,000 addresses, interviews were completed with 37,213 individuals for a response rate of about 74 percent.

Although sampling designs for both the BCS and the NCVS are more complex than we have represented in this discussion, the important point is how multistage cluster sampling is used in each. Notice two principal differences between the samples. First, the NCVS uses proportionate sampling to select a large number of respondents who may then represent the relatively rare attribute of victimization. The BCS samples a disproportionate number of minority and young residents, who are more likely to be victims of crime. Second, sampling procedures for the BCS are somewhat simpler than those for the NCVS, largely because a suitable sampling frame exists at the national level. Stratification and later-stage sampling is conducted to more efficiently represent each police area and to oversample minority respondents.

**Probability Sampling in Review**

Depending on the field situation, probability sampling can be very simple or extremely complex, time consuming, and expensive. Whatever the situation, however, it is usually the preferred method for selecting study elements. It’s worth restating the two main reasons for this.

First, probability sampling avoids conscious or unconscious biases in element selection on the part of the researcher. If all elements in the population have an equal (or unequal and subsequently weighted) chance of selection, there is an excellent chance that the sample so selected will closely represent the population of all elements.

Second, probability sampling permits estimates of sampling error. Although no probability sample will be perfectly representative in all respects, controlled selection methods permit the researcher to estimate the degree of expected error.

Despite these advantages, it is sometimes impossible to use standard probability sampling methods. Sometimes, it isn’t even appropriate to do so. In those cases, researchers turn to nonprobability sampling.

**Nonprobability Sampling**

*In many research applications, nonprobability samples are necessary or advantageous.*

You can no doubt envision situations in which it would be either impossible or unfeasible to select the kinds of probability samples we have described. Suppose we want to study auto thieves. There is no list of all auto thieves, nor are we likely to be able to create anything other than a partial and highly selective list. Moreover, probability sampling is sometimes inappropriate even if it is possible. In many such situations, **nonprobability sampling** procedures are called for. Recall that probability samples are defined as those in which the probability that any given sampling element will be selected is known. Conversely, in nonprobability
sampling, the likelihood that any given element will be selected is not known.

We’ll examine four types of nonprobability samples in this section: (1) purposive or judgmental sampling, (2) quota sampling, (3) the reliance on available subjects, and (4) snowball sampling.

**Purposive or Judgmental Sampling**

Occasionally, it may be appropriate to select a sample on the basis of our own knowledge of the population, its elements, and the nature of our research aims—in short, based on our judgment and the purpose of the study. Such a sample is called a *purposive sample*.

We may wish to study a small subset of a larger population in which many members of the subset are easily identified but the enumeration of all of them would be nearly impossible. For example, we might want to study members of community crime prevention groups; many members are easily visible, but it is not feasible to define and sample all members of community crime prevention organizations. In studying a sample of the most visible members, however, we may collect data sufficient for our purposes.

Criminal justice research often compares practices in different jurisdictions—cities or states, for example. In such cases, study elements may be selected because they exhibit some particular attribute. For instance, Cassia Spohn and Julie Horney (1991) were interested in how differences among states in rape shield laws affected the use of evidence in sexual assault cases. Strong rape shield laws restricted the use of evidence or testimony about a rape victim’s sexual behavior, whereas weak laws routinely permitted such testimony. Spohn and Horney selected a purposive sample of six states for analysis based on the strength of their rape shield laws. Similarly, Michael Leiber and Jayne Stairs (1999) were interested in how economic inequality combined with race to affect sentencing practices in Iowa juvenile courts. After controlling for economic status, they found that African American defendants received more restrictive sentences than white defendants did. Leiber and Stairs purposively selected three jurisdictions to obtain sample elements with adequate racial diversity in the state of Iowa. The researchers then selected over 5000 juvenile cases processed in those three courts.

Researchers may also use purposive or judgmental sampling to represent patterns of complex variation. In their study of closed-circuit television (CCTV) systems, Martin Gill and Angela Spriggs (2005) describe how sites were sampled to reflect variation in type of area (residential, commercial, city center, large parking facilities). Some individual CCTV projects were selected because of certain specific features—they were installed in a high-crime area, or the CCTV setup was notably expensive. One element of this study involved interviews to assess changes in fear of crime following CCTV installation. Spriggs and associates (2005) sampled passersby on city center streets. They first selected purposive samples of areas, and spread their interviews across four day/time periods. This was done to reflect variation in the types of people encountered on different streets at different times. Sampling strategies were thus adapted because of expected heterogeneity that would have been difficult to capture with random selection.

One of the best-known social science applications of judgmental sampling is the selection of voting precincts for exit polls on election days. On the basis of previous voting results in a given area (city, state, nation), TV networks select voting precincts that, in combination, produce results similar to those of the entire area. The theory is that the selected precincts represent a cross section of the entire electorate. Each time an election is held, analysts evaluate the adequacy of selected precincts and make revisions, additions, or deletions. The goal is to update the group of precincts to ensure that it provides a good representation of all precincts.

Pretesting a questionnaire is another situation in which purposive sampling is common.
If, for example, we plan to study people’s attitudes about court-ordered restitution for crime victims, we might want to test the questionnaire on a sample of crime victims. Instead of selecting a probability sample of the general population, we might select some number of known crime victims, perhaps from court records.

**Quota Sampling**

Like probability sampling, *quota sampling* addresses the issue of representativeness, although the two methods approach the issue quite differently. Quota sampling begins with a matrix or table describing the characteristics of the target population we wish to represent. To do this, we need to know, for example, what proportion of the population is male or female and what proportions fall into various age categories, education levels, ethnic groups, and so forth. In establishing a national quota sample, we need to know what proportion of the national population is, say, urban, eastern, male, under 25, white, working-class, and all the combinations of these attributes.

Once we have created such a matrix and assigned a relative proportion to each cell in the matrix, we can collect data from people who have all the characteristics of a given cell. We then assign all the persons in a given cell a weight appropriate to their portion of the total population. When all the sample elements are weighted in this way, the overall data should provide a reasonable representation of the total population.

Although quota sampling may resemble probability sampling, it has two inherent problems. First, the quota frame (the proportions that different cells represent) must be accurate, and it is often difficult to get up-to-date information for this purpose. A quota sample of auto thieves or teenage vandals would obviously suffer from this difficulty. Second, biases may exist in the selection of sample elements within a given cell—even though its proportion of the population is accurately estimated. An interviewer, instructed to interview five persons who meet a given complex set of characteristics, may still avoid people who live at the top of seven-story walk-ups, have particularly run-down homes, or own vicious dogs.

Quota and purposive sampling may be combined to produce samples that are intuitively, if not statistically, representative. For example, Kate Painter and David Farrington (1998) designed a survey to study marital and partner violence. They wished to represent several variables: marital status, age, an occupational measure of social status, and each of 10 standard regions in the United Kingdom. A probability sample was rejected because the authors wished to get adequate numbers of respondents in each of several categories, and some of the categories were thought to be relatively uncommon. Instead, the authors selected quota samples of 100 women in each of 10 regions and sought equal numbers of respondents in each of five occupational status categories.

**Reliance on Available Subjects**

Relying on available subjects—that is, stopping people at a street corner or some other location—is sometimes misleadingly called “convenience sampling.” University researchers frequently conduct surveys among the students enrolled in large lecture classes. The ease and economy of such a method explain its popularity; however, it seldom produces data of any general value. It may be useful to pretest a questionnaire, but it should not be used for a study purportedly describing students as a whole.

Reliance on available subjects can be an appropriate sampling method in some situations. It is generally best justified if the researcher wants to study the characteristics of people who are passing the sampling point at some specified time. For example, in her study of street lighting as a crime prevention strategy, Painter (1996) interviewed samples of pedestrians as they walked through specified areas of London just before and 6 weeks after improvements were made in lighting conditions. Painter clearly understood the scope and limits
of this sampling technique. Her findings are described as applying to people who actually use area streets after dark, while recognizing that this population may be quite different from the population of area residents. Interviewing a sample of available evening pedestrians is an appropriate sampling technique for generalizing to the population of evening pedestrians, and the population of pedestrians will not be the same as the population of residents.

In a more general sense, samples like Painter’s select elements of a process—the process that generates evening pedestrians—rather than elements of a population. If we can safely assume that no systematic pattern generates elements of a process, then a sample of available elements as they happen to pass by can be considered to be representative. So, for example, if you are interested in studying crimes reported to police, then a sample of, say, every seventh crime report over a 2-month period will be representative of the general population of crime reports over that 2-month period.

Sometimes nonprobability and probability sampling techniques can be combined. For example, most attempts to sample homeless or street people rely on available subjects found in shelters, parks, or other locations. Semaan and associates (2002) suggest that once areas are located where homeless people congregate, individuals found there can be enumerated then sampled. Here’s a semihypothetical example.

In recent years Maxfield has observed that many people who appear to be homeless congregate at the corner of 9th Avenue and 41st Street in Manhattan. An efficient strategy for interviewing samples of homeless people would be a time–space sample where, for example, each hour individuals would be counted and some fraction sampled. Let’s say we wished to interview 30 people, and spread those interviews over a 6-hour period; we would try to interview five people per hour. So each hour we would count the number of people within some specific area (say 20 at 1:00 P.M.), then divide that number by five to obtain a sampling fraction (4 in this case). Recalling our earlier discussion of systematic probability sampling, we would then select a random starting point to identify the first person to interview. Then select the fourth person after that, and so on. This approach would yield an unbiased sample that represented the population of street people on one Manhattan corner over a 6-hour period.

As it happens, 41st Street and 9th Avenue in Manhattan is the rear entrance to the Port Authority bus terminal. Marcus Felson and associates (1996) described efforts to reduce crime and disorder in the Port Authority terminal, a place they claim is the world’s busiest bus station. Among the most important objectives were to reduce perceptions of crime problems and to improve how travelers felt about the Port Authority terminal. These are research questions appropriate to some sort of survey. Because more than 170,000 passengers pass through the bus station on an average spring day, obtaining a sufficiently large sample of users presents no difficulty. The problem was how to select a sample. Felson and associates point out that stopping passengers on their way to or from a bus was out of the question. Most passengers are commuters whose journey to and from work is timed to the minute, with none to spare for interviewers’ questions. Here’s how Felson and associates describe the solution and the sampling strategy it embodied (1996:90–91):

Response rates would have been low if the Port Authority had tried to interview rushing customers or to hand out questionnaires to be returned later. Their solution was ingenious. The Port Authority drew a sample of outgoing buses . . . and placed representatives aboard. After the bus had departed, he or she would hand out a questionnaire to be completed during the trip . . . [and] collect these questionnaires as each customer arrived at the destination. This procedure produced a very high
response rate and high completion rate for each item.

Snowball Sampling

Another type of nonprobability sampling that closely resembles the available-subjects approach is called snowball sampling. Commonly used in field observation studies or specialized interviewing, snowball sampling begins by identifying a single subject or small number of subjects and then ask the subject(s) to identify others like him or her who might be willing to participate in a study.

Criminal justice research on active criminals or deviants frequently uses snowball sampling techniques. The researcher often makes an initial contact by consulting criminal justice agency records to identify, say, someone

PUTTING IT ALL TOGETHER

SAMPLING DRIVERS, PLACES, AND “KNOWN SPEEDERS”

Different studies of traffic enforcement and racial profiling employed a variety of probability and nonprobability sampling strategies. Target populations included drivers, places, and known traffic offenders. New Jersey researchers sampled vehicles entering the state’s turnpike. Engel and colleagues selected Pennsylvania counties from a range of purposive criteria that reflected roadway use. Many different samples were involved in North Carolina research. We’ll consider a stratified sample of “known North Carolina speeders” (2003:340).

Drivers

You can probably imagine the population of drivers on a roadway, but how would you define it? Place and time would be important dimensions, for example: “All drivers between milepost 1 and milepost 50 on the New Jersey Turnpike at 4:00 P.M. on Tuesday.” Because the New Jersey Turnpike is a toll road with access restricted to a relatively small number of points, that would be a relatively manageable population. Even so, that population would continually change as cars entered and exited the turnpike over the course of an hour or so.

Lange et al. took advantage of the limited access feature, as well as the fact that vehicles exiting the turnpike had to stop at tollbooths. This slowed vehicles, making it possible for researchers to observe features of cars and drivers that were of interest to their research. More important for sampling purposes, the cars leaving the turnpike were examples of a process or flow of units. The researchers sampled vehicles from that process. To accomplish that in an unbiased way, research staff stationed in tollbooths used hand-held computers to time the flow of vehicles. About every 2.5 minutes, an alarm sounded on the computer, signaling that the third vehicle approaching the tollbooth after the alarm was to be selected (2005:199). With this procedure, vehicles were selected for inclusion in the sample before drivers could be observed by research staff, thus eliminating any bias that might emerge if researchers chose which vehicles to select. Sampling also included procedures for selecting tollbooth exits to represent different sections of the highway, as well as day of week and time of day.

Counties and Road Segments

As a relatively uniform highway with few access points, the New Jersey Turnpike produced few sampling challenges. In contrast, Pennsylvania researchers sought locations that represented the considerable variety of roadways and conditions covered by troopers throughout the state. In selecting their sample, Engel and associates considered several characteristics of Pennsylvania counties that might affect roadway use and attract long-distance travelers. They specified three general constructs to represent, and seven
variables to measure those constructs at the county level:

- General roadway use
  - County population
  - Total roadway miles within each county
  - Interstate highway miles within each county
- Probable roadway use by minorities
  - County Hispanic population
  - County black population
- Likelihood that county road use did not reflect the residential population
  - Presence of tourist attractions, colleges and universities, or historical sites
  - Presence of seasonal attractions such as amusement parks or ski resorts

These seven variables were combined in a statistical procedure known as factor analysis, to produce summary scores for all Pennsylvania counties. Counties were then sorted into four groups ranging from high to low on the summary scores of roadway use, minority use, and nonlocal driver use. A total of 20 counties were selected from these four groups. An additional seven counties were selected because the racial distribution of drivers stopped was much different than the racial distribution of county residents.

Together, these procedures represented several sampling concepts described in this chapter. Counties were stratified by things that affected road use, and likely road use by minorities. Some counties were purposively selected based on high proportions of minority traffic stops. Eleven of 20 sampled counties were selected from the high-score group, producing a disproportionate number of sample units that had high levels of road use by minorities and nonlocal drivers. Finally, Engel et al. selected specific road segments within counties where they later conducted observations, resulting in a multistage sample design.

**Known Speeders**

In connection with their research in North Carolina, William Smith and associates used three general techniques to measure speeding: police records, independent observations, and self-report surveys. Like the Bureau of Justice Statistics survey on contacts with police, one goal of a survey is to sample people who had been stopped, but might not appear in police records. But Smith and associates were also interested in the validity of self-reported speeding and police contacts. How truthful were respondents asked about speeding and traffic stops?

To address this question, the researchers sampled court records of North Carolina residents who had received a speeding ticket between June 1999 and May 2000. Because they wished to include a large number of black drivers ticketed for speeding, the researchers used disproportionate stratified sampling so about half of selected names were for drivers whose race was classified as black on driving records. In this case the target population was persons known to have been cited for speeding. The sampling frame was a database containing records of all persons ticket for speeding, together with their address and phone number. Because race was recorded on driver’s licenses, it was easy to sort the records then use a random sampling computer procedure to select a sample stratified by race.

convicted of auto theft and placed on probation. That person is interviewed and asked to suggest other auto thieves whom researchers could contact. Stephen Baron and Timothy Hartnagel (1998) studied violence among homeless youths in Edmonton, Canada, identifying their sample through snowball techniques. Similarly, snowball sampling is often used to study drug users and dealers. Leon Pettiway (1995) describes crack cocaine markets in Philadelphia through the eyes of his snowball sample. Bruce Jacobs and Jody Miller (1998) accumulated a sample of 25 female crack dealers in St. Louis to study specific techniques to avoid arrest.

Contacting an initial subject or informant who will then refer the researcher to other subjects can be especially difficult in studies of active offenders. As in most aspects of
criminal justice research, the various approaches to initiating contacts for snowball sampling have advantages and disadvantages. Beginning with subjects who have a previous arrest or conviction is usually the easiest method for researchers, but it suffers from potential bias by depending on offenders who are known to police or other officials (McCall, 1978).

Because snowball samples are used most commonly in field research, we’ll return to this method of selecting subjects in Chapter 10 on field methods and observation. In the meantime, recent studies by researchers at the University of Missouri–St. Louis offer good examples of snowball samples of offenders that are not dependent on contacts with criminal justice officials. Beginning with a street-savvy ex-offender, these researchers identified samples of burglars (Wright and Decker, 1994), members of youth gangs (Decker and Van Winkle, 1996), and armed robbers (Wright and Decker, 1997). It’s especially difficult to identify active offenders as research subjects, but these examples illustrate notably clever uses of snowball sampling techniques.

Nonprobability Sampling in Review
Snowball samples are essentially variations on purposive samples (we want to sample juvenile gang members) and on samples of available subjects (sample elements identify other sample elements that are available to us). Each of these is a nonprobability sampling technique. And, like other types of nonprobability samples, snowball samples are most appropriate when it is impossible to determine the probability that any given element will be selected in a sample. Furthermore, snowball sampling and related techniques may be necessary when the target population is difficult to locate or even identify. Approaching pedestrians who happen to pass by, for example, is not an efficient way to select a sample of prostitutes or juvenile gang members. In contrast, approaching pedestrians is an appropriate sampling method for studying pedestrians, whereas drawing a probability sample of urban residents to identify people who walk in specific areas of the city would be costly and inefficient.

Like other elements of criminal justice research, sampling plans must be adapted to specific research applications. When it’s important to make estimates of the accuracy of our samples, and when suitable sampling frames are possible, we use probability sampling techniques. When no reasonable sampling frame is available, and we cannot draw a probability sample, we cannot make estimates about sample accuracy. Fortunately, in such situations, we can make use of a variety of approaches for drawing nonprobability samples.

❖ Main Points

- The logic of probability sampling forms the foundation for representing large populations with small subsets of those populations.
- The chief criterion of a sample’s quality is the degree to which it is representative—the extent to which the characteristics of the sample are the same as those of the population from which it was selected.
- The most carefully selected sample is almost never a perfect representation of the population from which it was selected. Some degree of sampling error always exists.
- Probability sampling methods provide one excellent way of selecting samples that will be quite representative. They make it possible to estimate the amount of sampling error that should be expected in a given sample.
- The chief principle of probability sampling is that every member of the total population must have some known nonzero probability of being selected into the sample.
- Our ability to estimate population parameters with sample statistics is rooted in the sampling distribution and probability theory. If we draw a large number of samples of a given size, sample statistics will cluster around the true population parameter. As sample size increases, the cluster becomes tighter.
- A variety of sampling designs can be used and combined to suit different populations and research purposes. Each type of sampling has its own advantages and disadvantages.
• Simple random sampling is logically the most fundamental technique in probability sampling although it is seldom used in practice.
• Systematic sampling involves using a sampling frame to select units that appear at some specified interval—for example, every 8th, or 15th, or 1023rd unit. This method is functionally equivalent to simple random sampling.
• Stratification improves the representativeness of a sample by reducing the sampling error.
• Disproportionate stratified sampling is especially useful when we want to select adequate numbers of certain types of subjects who are relatively rare in the population we are studying.
• Multistage cluster sampling is frequently used when there is no list of all the members of a population.
• The NCVS and the BCS are national crime surveys based on multistage cluster samples. Sampling methods for each survey illustrate different approaches to representing relatively rare events.
• Nonprobability sampling methods are less statistically representative and less reliable than probability sampling methods. However, they are often easier to use and can be appropriate for many research applications.
• Purposive sampling is used when researchers wish to select specific elements of a population. This may be because the elements are believed to be representative, extreme cases or because they represent the range of variation expected in a population.
• In quota sampling, researchers begin with a detailed description of the characteristics of the total population and then select sample members in a way that includes the different composite profiles that exist in the population.
• In cases in which it’s not possible to draw nonprobability samples through other means, researchers often rely on available subjects. Professors sometimes do this—students in their classes are available subjects.
• Snowball samples accumulate subjects through chains of referrals and are most commonly used in field research.

**Key Terms**

- Binomial variable
- Confidence interval
- Cluster sample
- Confidence level
- Disproportionate stratified sampling
- Equal probability of selection method (EPSEM)
- Nonprobability sample
- Population
- Population parameter
- Probability sample
- Purposive sample
- Quota sample
- Sample element
- Sample statistic
- Simple random sample
- Snowball sampling
- Standard error
- Stratification
- Systematic sampling

**Review Questions and Exercises**

1. Review the summary of the Drug Abuse Warning Network (DAWN) program in Chapter 6. Specify the target population, study population, sampling frame, and elements used in the DAWN program. Describe what type of sample DAWN uses, and discuss the advantages and disadvantages of sampling procedures.

2. Discuss possible study populations, elements, sampling units, and sampling frames for drawing a sample to represent the populations listed here. You may wish to limit your discussion to populations in a specific state or other jurisdiction.
   a. Municipal police officers
   b. Felony court judges
   c. Auto thieves
   d. Licensed automobile drivers
   e. State police superintendents
   f. Persons incarcerated in county jails

**Online Sessions for Chapter 8**

Online Sessions present further examples of selected material covered in each chapter. Your instructor may choose to use these for in-class discussion (as I do with my students) or you can access them at home. To ensure that you always have access to live, correct links for the websites described here, we provide the necessary links on the companion website for *Research Methods for Criminal Justice and Criminology* at http://www.thomsonedu.com/criminaljustice/maxfield. Once at the companion website, select this specific chapter, click on “Chapter Resources,” then click on “Web Links.”

**Confidence Intervals in Opinion Polls**

Polls reporting presidential approval ratings are among the more well-known uses of probability sampling.
and survey research to represent public opinion. The PollKatz website summarizes responses from 15 polls regularly reported in U.S. mass media. Each uses a telephone interview from a random sample of telephone numbers. But each poll produces slightly different results. The PollKatz website arrays these results on a single graph so you can see patterns of differences. This home page also presents a nice discussion of standard errors and confidence intervals. See also links to web pages for individual polls.

Be aware that the PollKatz site presents a distinct anti–George Bush point of view. Nonetheless it offers a nice insight into several dimensions of sampling and public opinion polls. [http://www.pollkatz.homestead.com/](http://www.pollkatz.homestead.com/)

**Resources for Probability Sampling** William Trochim’s Knowledge Base site includes some good information on different types of sampling. The site is better in its discussion of probability sampling than nonprobability sampling. It also presents information about various sampling terms and concepts. See especially the distinction between random selection and random assignment, something we discussed in Chapter 7. [http://www.socialresearchmethods.net/kb/sampprob.htm](http://www.socialresearchmethods.net/kb/sampprob.htm)

**Tips on Snowball Sampling** Snowball sampling is a type of nonprobability sampling strategy commonly used in field research. The University of Surrey Social Research Update website presents some very useful guidelines on using snowball approaches to find hard-to-reach populations. See Update #33. [http://www.soc.surrey.ac.uk/sru/SRU33.html](http://www.soc.surrey.ac.uk/sru/SRU33.html)

**Sample Size and Other Basic Statistical Tools** Steve Simon maintains a marvelous website on statistics—STeve’s Attempts to Teach Statistics (STATS)—at Children’s Mercy Hospital in St. Louis. For our chapter on sampling, see links to various tips on how to compute sample size, and general guidelines on sample sizes. He offers a great analogy to explain why the size of a target population has little effect on the size of a sample. “No matter how large the pot, every cook knows that it only takes a single sip from a well-stirred soup to determine the taste.” Material here is not technical, but it does assume some familiarity with basic statistics. Nevertheless, you will find many topics of interest. From the STATS home page, look for a link to “selecting an appropriate sample size” or something like that; the site changes periodically. [http://www.childrensmercy.org/stats/](http://www.childrensmercy.org/stats/)

**Descriptions of Samples from Hundreds of Studies** The National Archive of Criminal Justice Data (NACJD) contains criminal justice data files from numerous sources. Each data set has extensive documentation listing the source of the data, the sampling frame, the sample, and so on. You can learn more about samples, target populations, and sampling strategy by reading descriptions of data archived at NACJD.

From the website’s home page, select the “Search/Browse Holding” link. Choices on the next page are inexplicably presented as “Browse BJS-funded studies” or “Browse NIJ-funded studies.” You’ll find greater variety in the NIJ lists, but more uniform descriptions for BJS studies. Other browse and search options are available, so start with one that seems sensible.

Scroll down the list of studies, and click on “description” for at least two. After reading the description, list the following: the population (universe), the sampling frame, the study population, the observation unit, and the sampling method that was used to gather the sample. [http://www.icpsr.umich.edu/NACJD/](http://www.icpsr.umich.edu/NACJD/)

**Additional Readings**

General Accounting Office, *Using Statistical Sampling* (Washington, DC: U.S. General Accounting Office, 1992). One of a series of handbooks on evaluation methods, this is a rich source of practical information on how to draw probability samples. You will also find suggestions on how to work around various difficulties in actually selecting samples, as well as illustrations of sampling people and other types of elements.

Kish, Leslie, *Survey Sampling* (New York: Wiley, 1965). Unquestionably the definitive work on sampling in social research. Kish’s coverage ranges from the simplest matters to the most complex and mathematical. He is both highly theoretical and downright practical. Easily readable and difficult passages intermingle as Kish dissects everything you could want or need to know about each aspect of sampling.

Patton, Michael Quinn, *Qualitative Research and Evaluation Methods*, 3rd ed. (Thousand Oaks, CA: Sage, 2001). Though its focus is evaluation,
this book presents one of the best discussions of nonprobability sampling available. Patton covers a wide range of variations on purposive sampling.


We’ll examine how mail, interview, and telephone surveys can be used in criminal justice research. We’ll also consider other ways of collecting data by asking people questions.

<table>
<thead>
<tr>
<th>Topics</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>245</td>
</tr>
<tr>
<td>Topics Appropriate to Survey Research</td>
<td>246</td>
</tr>
<tr>
<td>Counting Crime</td>
<td>246</td>
</tr>
<tr>
<td>Self-Reports</td>
<td>246</td>
</tr>
<tr>
<td>Perceptions and Attitudes</td>
<td>247</td>
</tr>
<tr>
<td>Targeted Victim Surveys</td>
<td>247</td>
</tr>
<tr>
<td>Other Evaluation Uses</td>
<td>248</td>
</tr>
<tr>
<td>General-Purpose Crime Surveys</td>
<td>248</td>
</tr>
<tr>
<td>Guidelines for Asking Questions</td>
<td>248</td>
</tr>
<tr>
<td>Open-Ended and Closed-Ended Questions</td>
<td>249</td>
</tr>
<tr>
<td>Questions and Statements</td>
<td>249</td>
</tr>
<tr>
<td>Make Items Clear</td>
<td>249</td>
</tr>
<tr>
<td>Short Items Are Best</td>
<td>250</td>
</tr>
<tr>
<td>Avoid Negative Items</td>
<td>250</td>
</tr>
<tr>
<td>Biased Items and Terms</td>
<td>250</td>
</tr>
<tr>
<td>Designing Self-Report Items</td>
<td>251</td>
</tr>
<tr>
<td>Questionnaire Construction</td>
<td>253</td>
</tr>
<tr>
<td>General Questionnaire Format</td>
<td>253</td>
</tr>
<tr>
<td>Contingency Questions</td>
<td>253</td>
</tr>
<tr>
<td>Matrix Questions</td>
<td>254</td>
</tr>
<tr>
<td>Ordering Questions in a Questionnaire</td>
<td>256</td>
</tr>
</tbody>
</table>

DON’T START FROM SCRATCH! 257

Self-Administered Questionnaires 257

(continued)
Introduction

Asking people questions is the most common data collection method in social science.

A little-known survey was attempted among French workers in 1880. A German political sociologist mailed some 25,000 questionnaires to workers to determine the extent of their exploitation by employers. The rather lengthy questionnaire included items such as these:

Does your employer or his representative resort to trickery in order to defraud you of a part of your earnings? If you are paid piece rates, is the quality of the article made a pretext for fraudulent deductions from your wages?

The survey researcher in this case was not George Gallup but Karl Marx (1880:208). Although 25,000 questionnaires were mailed out, there is no record of any being returned. And you need not know much about survey methods to recognize the loaded questions posed by Marx.

Survey research is perhaps the most frequently used mode of observation in sociology and political science, and surveys are often used in criminal justice research as well. You have no doubt been a respondent in some sort of survey, and you may have conducted surveys yourself.

We begin this chapter by discussing the criminal justice topics that are most appropriate for survey methods. Next, we cover the basic principles of how to ask people questions for research purposes, including some of the details of questionnaire construction. We describe the three basic ways of administering questionnaires—self-administration, face-to-face interviews, and telephone interviews—and summarize the strengths and weaknesses of each method. After discussing more specialized
interviewing techniques, such as focus groups, we conclude the chapter with some advice on the benefits and pitfalls of conducting your own surveys.

**Topics Appropriate to Survey Research**

*Surveys have a wide variety of uses in basic and applied criminal justice research.*

Surveys may be used for descriptive, explanatory, exploratory, and applied research. They are best suited for studies that have individual people as the units of analysis. They are often used for other units of analysis as well, such as households or organizations. Even in these cases, however, one or more individual people act as respondents or informants.

For example, researchers sometimes use victimization incidents as units of analysis in examining data from crime surveys. The fact that some people may be victimized more than once and others not at all means that victimization incidents are not the same units as individuals. However, a survey questionnaire must still be administered to people who provide information about victimization incidents. In a similar fashion, the National Jail Census, conducted every 5 or so years for the Bureau of Justice Statistics, collects information about local detention facilities. Jails are the units of analysis, but information about each jail is provided by individuals. Quite a lot of research on police practices was conducted following passage of the 1994 Crime Bill, and in most cases, law enforcement agencies were units of analysis; individual people, however, provided information for the surveys of police departments.

We now consider some broad categories of research applications in which survey methods are especially appropriate.

**Counting Crime**

Chapter 6 covered this use of surveys in detail. Asking people about victimizations is a measure of crime that adjusts for some of the problems found in data collected by police. Of course, survey measures have their own shortcomings. Most of these difficulties, such as recall error and reluctance to discuss victimization with interviewers, are inherent in survey methods. Nevertheless, victim surveys have become important sources of data about the volume of crime in the United States and in other countries.

**Self-Reports**

Surveys that ask people about crimes they may have committed were also discussed in Chapter 6. Thornberry and Krohn describe self-report surveys as “... an integral part of the way crime and delinquency is studied” (2000:35). For research that seeks to explore or explain why people commit criminal, delinquent, or deviant acts, asking questions is the best method available.

Within the general category of self-report surveys, two different applications are distinguished by their target population and sampling methods. Studies of offenders select samples of respondents known to have committed crimes, often prisoners. Typically, the focus is on the frequency of offending—how many crimes of various types are committed by active offenders over a period of time. A study of incarcerated felons by Jan and Marcia Chaiken (1982) is among the best-known self-report surveys of offenders. John Ball and associates (Ball, Rosen, Flueck, and Nurco, 1982) studied a sample of heroin addicts, documenting the crimes they committed over several years.

The other type of self-report survey focuses on the prevalence of offending—how many people commit crimes, in contrast to how many crimes are committed by a target population of offenders. Such surveys typically use samples that represent a broader population, such as U.S. households, adult males, or high school seniors. The National Youth Survey (NYS), sponsored by the National Institute for Mental Health and the Office of Juvenile Justice and Delinquency Prevention, began in 1976. The NYS is a panel study (described in Chapter 4) that began with a national probability sample
of children who were then ages 11–17. The most recent follow-up interview with these subjects was conducted in 2004. Since the survey’s first wave, data from the NYS have been used to make estimates of the prevalence of criminal and delinquent acts. Delbert Elliott and associates (Elliott, Huizinga, and Menard, 1989) present comprehensive results from the NYS, and Finn-Age Esbensen and Delbert Elliott (1994) describe patterns of drug use over eight waves of the survey. Other examples of self-report surveys were described in Chapter 6.

General-population surveys and surveys of offenders tend to present different types of difficulties in connection with the validity and reliability of self-reports. Recall error and the reporting of fabricated offenses may be problems in a survey of high-rate offenders (Horney and Marshall, 1992); respondents in general-population self-report surveys may be reluctant to disclose illegal behavior. Jennifer Roberts and associates (2005) describe how weekly interviews were required to assess self-reported violence in a sample of psychiatric patients. When we discuss questionnaire construction later in this chapter, we will present examples and suggestions for creating self-report items.

Perceptions and Attitudes
Another application of surveys in criminal justice is to learn how people feel about crime and criminal justice policy. Public views about sentencing policies, gun control, police performance, and drug abuse are often solicited in opinion polls. Begun in 1972, the General Social Survey is an ongoing survey of social indicators in the United States. Questions about fear of crime and other perceptions of crime problems are regularly included. Since the mid-1970s, a growing number of explanatory studies have been conducted on public perceptions of crime and crime problems. A large body of research on fear of crime has grown, in part, from the realization that fear and its behavioral consequences are much more widespread among the population than is actual criminal victimization (Hale, 1996).

Targeted Victim Surveys
Victim surveys that target individual cities or neighborhoods are important tools for evaluating policy innovations. Many criminal justice programs seek to prevent or reduce crime in some specific area, but crimes reported to police cannot be used to evaluate many types of programs.

To see why this is so, consider a hypothetical community policing program that encourages neighborhood residents to report all suspected crimes to the police. We saw in Chapter 6 that many minor incidents are not reported because victims believe that police will not want to be bothered. But if a new program stresses that police actually want to be bothered, the proportion of crimes reported may increase, resulting in what appears to be an increase in crime.

The solution is to conduct targeted victim surveys before and after introducing some policy change. Such victim surveys are especially appropriate for evaluating any policy that may increase crime reporting as a side effect.

Consider also that large-scale surveys such as the National Crime Victimization Survey (NCVS) cannot be used to evaluate local crime prevention programs. This is because, as we have seen, the NCVS is designed to represent the national population of persons who live in households. NCVS data for the 11 largest states can be analyzed separately (Maltz, 1998), but the NCVS is not representative of any particular local jurisdiction. However, a recent report by Janet Lauritsen and Robin Schaum (2005) examined NCVS data on burglary, robbery, and assault victimization for Chicago, Los Angeles, and New York. This required special access to location information on NCVS records. It is not possible to identify the specific location of victimizations from NCVS data available for routine analysis.

The community victim surveys designed by the BJS and the COPS office help with each of these needs. Local surveys can be launched specifically to evaluate local crime prevention efforts. Or innovative programs can be timed
to correspond to regular cycles of local surveys. In each case, the BJS/COPS (Weisel, 1999) guide presents advice on drawing samples to represent local jurisdictions.

Another type of targeted victim survey is one that focuses on particular types of incidents that might target more narrowly defined population segments. A good example is the National Violence Against Women Survey, a joint effort of the National Institute of Justice and a violence prevention bureau in the National Institutes of Health (Tjaden and Thoennes, 2000). Screening questions present explicit descriptions of sexual and other violence with the specific purpose of providing better information about these incidents that have proved difficult to measure through general-purpose crime surveys.

Other Evaluation Uses
Other types of surveys may be appropriate for applied studies. A good illustration of this is a continuing series of neighborhood surveys to evaluate community policing in Chicago. Here’s an example of how the researchers link their information needs to surveys (Chicago Community Policing Evaluation Consortium, 2004:2):

Because it is a participatory program, CAPS depends on the effectiveness of campaigns to bring it to the public’s attention and on the success of efforts to get the public involved in beat meetings and other district projects. The surveys enable us to track the public’s awareness and involvement in community policing in Chicago.

In general, surveys can be used to evaluate policy that seeks to change attitudes, beliefs, or perceptions. For example, consider a program designed to promote victim and witness cooperation in criminal court by reducing case processing time. At first, we might consider direct measures of case processing time as indicators of program success. If the program goal is to increase cooperation, however, a survey that asks how victims and witnesses perceive case processing time will be more appropriate.

General-Purpose Crime Surveys
As the name implies, these surveys are designed for more than one of the purposes we have mentioned. Since its inception in 1982, the BCS has included special batteries of questions about various topics: fear of crime (1984), contacts with the police (1988, 1996, 2000), self-reported drug use (1994), security measures used to protect households and vehicles (1996), and attitudes toward criminal sentencing (2000).

The NCVS has traditionally focused on counting crime. In recent years, however, efforts have been devoted to incrementally redesigning the survey to broaden its scope (Bachman and Taylor, 1994; Cantor and Lynch, 2000). Beginning in 1996, the NCVS added a series of questions on police–public contacts other than those in connection with victimization. The most interesting type of such contact is a traffic stop—survey respondents describe their experiences with police in connection with traffic violations. Using the NCVS as a research platform in this way has produced a new source of data with which to assess the scope of race-based disparities in police–public contacts. See DuRose (2005) for examples of tabulations from this feature of the NCVS.

Guidelines for Asking Questions
How questions are asked is the single most important feature of survey research.

A defining feature of survey methods is that research concepts are operationalized by asking people questions. Several general guidelines can assist in framing and asking questions that serve as excellent operationalizations of variables. It is important also to be aware of pitfalls that can result in useless and even misleading information. We’ll begin with some of the options available for creating questionnaires.
Open-Ended and Closed-Ended Questions

In asking questions, researchers have two basic options, and each can accommodate certain variations. The first is open-ended questions, in which the respondent is asked to provide his or her own answers. For example, the respondent may be asked, “What do you feel is the most important crime problem facing the police in your city today?” and be provided with a space to write in the answer (or be asked to report it orally to an interviewer). The other option is closed-ended questions, in which the respondent is asked to select an answer from among a list provided by the researcher.

Closed-ended questions are especially useful because they provide more uniform responses and are more easily processed. They often can be transferred directly into a data file. Open-ended responses, in contrast, must be coded before they can be processed for analysis. This coding process often requires that the researcher interpret the meaning of responses, which opens up the possibility of misunderstanding and researcher bias. Also, some respondents may give answers that are essentially irrelevant to the researcher’s intent.

The chief shortcoming of closed-ended questions lies in the researcher’s structuring of responses. When the relevant answers to a given question are relatively clear, there should be no problem. In some cases, however, the researcher’s list of responses may fail to include some important answers. When we ask about “the most important crime problem facing the police in your city today,” for example, our checklist might omit certain crime problems that respondents consider important.

In constructing closed-ended questions, we are best guided by two of the requirements for operationalizing variables stated in Chapter 5. First, the response categories provided should be exhaustive: They should include all the possible responses that might be expected. Often, researchers ensure this by adding a category labeled something like “Other (Please specify: ____).” Second, the answer categories must be mutually exclusive: The respondent should not feel compelled to select more than one. In some cases, researchers solicit multiple answers, but doing so can create difficulties in subsequent data processing and analysis. To ensure that categories are mutually exclusive, we should carefully consider each combination of categories, asking whether a person could reasonably choose more than one answer. In addition, it is useful to add an instruction that respondents should select the one best answer. However, this is still not a satisfactory substitute for a carefully constructed set of responses.

Questions and Statements

The term questionnaire suggests a collection of questions, but a typical questionnaire probably has as many statements as questions. This is because researchers often are interested in determining the extent to which respondents hold a particular attitude or perspective. Researchers try to summarize the attitude in a fairly brief statement; then they present that statement and ask respondents whether they agree or disagree with it. Rensis Likert formalized this procedure through the creation of the Likert scale, a format in which respondents are asked whether they strongly agree, agree, disagree, or strongly disagree, or perhaps strongly approve, approve, and so forth.

Both questions and statements may be used profitably. Using both in a questionnaire adds flexibility in the design of items and can make the questionnaire more interesting as well.

Make Items Clear

It should go without saying that questionnaire items must be clear and unambiguous, but the broad proliferation of unclear and ambiguous questions in surveys makes the point worth restressing here. Researchers commonly become so deeply involved in the topic that opinions and perspectives that are clear to them will not be at all clear to respondents, many of whom
have given little or no thought to the topic. Or researchers may have only a superficial understanding of the topic and so may fail to specify the intent of a question sufficiently. The question “What do you think about the governor’s decision concerning prison furloughs?” may evoke in the respondent a counterquestion or two: “Which governor’s decision?” “What are prison furloughs?” Questionnaire items should be precise so that the respondent knows exactly what the researcher wants an answer to.

Frequently, researchers ask respondents for a single answer to a combination question. Such “double-barreled” questions seem to occur most often when the researcher has personally identified with a complex question. For example, the researcher might ask respondents to agree or disagree with the statement “The Department of Corrections should stop releasing inmates for weekend furloughs and concentrate on rehabilitating criminals.” Although many people will unequivocally agree with the statement and others will unequivocally disagree, still others will be unable to answer. Some might want to terminate the furlough program and punish—not rehabilitate—prisoners. Others might want to expand rehabilitation efforts while maintaining weekend furloughs; they can neither agree nor disagree without misleading the researcher.

Short Items Are Best

In the interest of being unambiguous and precise and pointing to the relevance of an issue, researchers often create long, complicated items. That should be avoided. In the case of self-administered questionnaires, respondents are often unwilling to study an item in order to understand it. The respondent should be able to read an item quickly, understand its intent, and select or provide an answer without difficulty. In general, it’s safe to assume that respondents will read items quickly and give quick answers; therefore, short, clear items that will not be misinterpreted under those conditions are best. Questions read to respondents in person or over the phone should be similarly brief.

Avoid Negative Items

A negation in a questionnaire item paves the way for easy misinterpretation. Asked to agree or disagree with the statement “Drugs such as marijuana should not be legalized,” many respondents will overlook the word not and answer on that basis. Thus, some will agree with the statement when they are in favor of legalizing marijuana, and others will agree when they oppose it. And we may never know which is which.

Biased Items and Terms

Recall from the earlier discussion of conceptualization and operationalization that there are no ultimately true meanings for any of the concepts we typically study in social science. This same general principle applies to the responses we get from persons in a survey. The meaning of a given response to a question depends in large part on the wording of the question. That is true of every question and answer. Some questions seem to encourage particular responses more than do other questions. Questions that encourage respondents to answer in a particular way are biased. Most researchers recognize the likely effect of a question such as “Do you support the president’s use of intensive interrogation to promote the safety and security of all Americans?” and no reputable researcher would use such an item. The biasing effect of items and terms is far subtler than this example suggests, however.

The mere identification of an attitude or position with a prestigious (or unpopular) person or agency can bias responses. For example, an item that starts with “Do you agree or disagree with the recent Supreme Court decision that . . . ” might have this effect. We are not suggesting that such wording will necessarily produce consensus or even a majority in support of the position identified with the prestigious person or agency. Rather, support will likely be
greater than what would have been obtained without such identification.

Sometimes, the impact of different forms of question wording is relatively subtle. For example, Kenneth Rasinski (1989) analyzed the results of several General Social Survey studies of attitudes toward government spending. He found that the way programs were identified had an impact on the amount of public support they received. Here are some comparisons:

<table>
<thead>
<tr>
<th>More Support</th>
<th>Less Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Assistance to the poor”</td>
<td>“Welfare”</td>
</tr>
<tr>
<td>“Halting the rising crime rate”</td>
<td>“Law enforcement”</td>
</tr>
<tr>
<td>“Dealing with drug addiction”</td>
<td>“Drug rehabilitation”</td>
</tr>
</tbody>
</table>

In 1986, for example, 63 percent of respondents said too little money was being spent on “assistance to the poor,” whereas in a matched survey that year, only 23 percent said we were spending too little on “welfare.” The main guidance we offer for avoiding bias is that researchers imagine how they would feel giving each of the answers they offer to respondents.

If they would feel embarrassed, perverted, inhumane, stupid, irresponsible, or anything like that, then they should give some serious thought to whether others will be willing to give those answers. Researchers must carefully examine the purpose of their inquiry and construct items that will be most useful to it.

We also need to be generally wary of what researchers call the “social desirability” of questions and answers. Whenever we ask people for information, they answer through a filter of what will make them look good. That is especially true if they are being interviewed in a face-to-face situation.

**Designing Self-Report Items**

Social desirability is one of the problems that plague self-report crime questions in general-population surveys. Adhering to the ethical principles of confidentiality and anonymity, as well as convincing respondents that we are doing so, is one way of getting more truthful responses to self-report items. Other techniques can help us avoid or reduce problems with self-report items.

One method, used in earlier versions of the BCS, is to introduce a group of self-report items with a disclaimer and to sanitize the presentation of offenses. The self-report section of the 1984 BCS began with this introduction:

There are lots of things which are actually crimes, but which are done by lots of people, and which many people do not think of as crimes. On this card [printed card handed to respondents] are a list of eight of them. For each one can you tell me how many people you think do it—most people, a lot of people, or no one.

Respondents then read a card, shown in Figure 9.1, that presented descriptions of various offenses. Interviewers first asked respondents how many people they thought ever did X, where X corresponded to the letter for an offense shown in Figure 9.1. Next, respondents were asked whether they had ever done X. Interviewers then moved on down the list of letters for each offense on the card.

This procedure incorporates three techniques to guard against the socially desirable response of not admitting to having committed a crime. First, the disclaimer seeks to reassure respondents that “many people” do not really think of various acts as crimes. Second, respondents are asked how many people they think commit each offense before being asked whether they have done so themselves. This takes advantage of a common human justification for engaging in certain kinds of behavior—other people do it. Third, asking whether they “have ever done X” is less confrontational than asking whether they “have ever cheated on an expense account.” Again, the foibles of human behavior are at work here, in much the same way that people use euphemisms such as “rest room” and “sleep together” for “toilet” and
“have sexual intercourse.” It is, of course, not realistic to expect that such ploys will reassure all respondents. Furthermore, disclaimers about serious offenses such as rape or bank robbery would be ludicrous. But such techniques illustrate how thoughtful wording and introductions can be incorporated into sensitive questions.

Self-report surveys of known offenders encounter different problems. Incarcerated persons may be reluctant to admit committing crimes because of the legal consequences. High-rate offenders may have difficulty distinguishing among a number of different crimes or remembering even approximate dates. As discussed in Chapter 3, promises of immunity from prosecution for offenses revealed in certain forms of sponsored research can allay the fears of many active offenders. Sorting out dates and details of individual crimes among high-rate offenders requires different strategies.

One technique that is useful in surveys of active offenders is to interview subjects several times at regular intervals. For example, Lisa Maher (1997) interviewed her sample of heroin- or cocaine-addicted women repeatedly (sometimes daily) over the course of 3 years. Each subject was asked about her background, intimate relationships with men, income-generating activities, and drug use habits. Having regular interviews helped respondents recall offending.

A similar approach was used by Julie Horney and Ineke Haen Marshall (1992), who questioned methods used by RAND Corporation researchers to obtain self-report data from prisoners. In the RAND study (Chaiken and Chaiken, 1982), high-rate offenders were asked to estimate their usual level of criminal activity, a procedure that Horney and Marshall believed might produce inflated estimates of crimes committed. Their solution was to present inmates with a “crime calendar” on which they recorded month-by-month lists of offenses. In an experimental comparison of the crime calendar approach with the RAND method, Horney and Marshall found no differences in the numbers of crimes recalled by subjects. It is likely, however, that the monthly diary format helped subjects more accurately recall the numbers and characteristics of crimes they had committed. More recent research on the use of crime calendars (Roberts, Mulvey, Horney, et al., 2005) finds that more frequent interviews are necessary for recall by high-rate offenders, and that crime calendars are best suited for tracking more serious offenses.

Obtaining valid and reliable results from self-report items is challenging, but self-report survey techniques are important tools for addressing certain types of criminal justice research questions. Because of this, researchers are constantly striving to improve self-report items. See the collection of essays by Joseph Gfroerer and associates (Gfroerer, Eyerman, and Chromy, 2002) for a detailed discussion of issues involved in measuring self-reported drug use through the National Household Survey on Drug Abuse. A National Research Council

---

| A. | Taken office supplies from work (such as stationery, envelopes, and pens) when not supposed to. |
| B. | Taken things other than office supplies from work (such as tools, money, or other goods) when not supposed to. |
| C. | Fiddled expenses [fiddled is the Queen’s English equivalent of fudged]. |
| D. | Deliberately traveled [on a train] without a ticket or paid too low a fare. |
| E. | Failed to declare something at customs on which duty was payable. |
| F. | Cheated on tax. |
| G. | Used cannabis (hashish, marijuana, ganga, grass). |
| H. | Regularly driven a car when they know they have drunk enough to be well above the legal limit. |

**Figure 9.1** Showcard for Self-Report Items, 1984 British Crime Survey
*Source: Adapted from the 1984 British Crime Survey (NOP Market Research Limited, 1985).*
Computer technology has made it possible to significantly improve self-reported items. We present examples later in the chapter when we describe different modes of survey administration.

**Questionnaire Construction**

*After settling on question content, researchers must consider the format and organization of all items in a questionnaire.*

Because questionnaires are the fundamental instruments of survey research, we now turn our attention to some of the established techniques for constructing them. The following sections are best considered as a continuation of our theoretical discussions in Chapter 5 of conceptualization and measurement.

Of course, how a questionnaire is constructed depends on how the questionnaire will be administered to respondents. Later in this chapter, we will consider the three modes of administration: (1) self-administered, (2) in-person interview, and (3) telephone interview. For now, we simply point out that a rough continuum of questionnaire complexity is associated with the three modes of administration. In general, questionnaires administered through in-person interviews are the most complex. Self-administered questionnaires tend to be the least complex and must be more carefully formatted than questionnaires administered by trained interviewers, either in person or by telephone.

**General Questionnaire Format**

The format of a questionnaire is just as important as the nature and wording of the questions. An improperly laid-out questionnaire can cause respondents to miss questions, confuse them about the nature of the data desired, and, in the extreme, lead them to throw the questionnaire away.

As a general rule, the questionnaire should be uncluttered. Inexperienced researchers tend to fear that their questionnaire will look too long, so they squeeze several questions onto a single line, abbreviate questions, and try to use as few pages as possible. Such efforts are ill-advised and even counterproductive. Putting more than one question on a line will cause some respondents to miss the second question altogether. Some respondents will misinterpret abbreviated questions. And, more generally, respondents who have spent considerable time on the first page of what seemed a short questionnaire will be more demoralized than will respondents who quickly completed the first several pages of what initially seemed a long form. Moreover, the latter will have made fewer errors and will not have been forced to reread confusing, abbreviated questions. Nor will they have been forced to write a long answer in a tiny space.

**Contingency Questions**

Quite often in questionnaires, certain questions are clearly relevant to only some of the respondents and irrelevant to others. A victim survey, for example, presents batteries of questions about victimization incidents that are meaningful only to crime victims.

Frequently, this situation—realizing that the topic is relevant only to some respondents—arises when we wish to ask a series of questions about a certain topic. We may want to ask whether respondents belong to a particular organization and, if so, how often they attend meetings, whether they have held office in the organization, and so forth. Or we might want to ask whether respondents have heard anything about a certain policy proposal, such as opening a youth shelter in the neighborhood, and then investigate the attitudes of those who have heard of it.

The subsequent questions in series such as these are called “contingency questions”; whether they are to be asked and answered is contingent on the response to the first question.
in the series. The proper use of contingency questions can make it easier for respondents to complete the questionnaire because they do not have to answer questions that are irrelevant to them.

Contingency questions can be presented in several formats. The one shown in Figure 9.2 is probably the clearest and most effective. Note that the questions shown in the figure could have been dealt with in a single question: “How many times, if any, have you smoked marijuana?” The response categories then would be: “Never,” “Once,” “2 to 5 times,” and so forth. This single question would apply to all respondents, and each would find an appropriate answer category. Such a question, however, might put pressure on some respondents to report having smoked marijuana, because the main question asks how many times they have smoked it. The contingency question format illustrated in Figure 9.2 reduces the subtle pressure on respondents to report having smoked marijuana. This discussion shows how seemingly theoretical issues of validity and reliability are involved in so mundane a matter as how to format questions on a piece of paper.

Used properly, even complex sets of contingency questions can be constructed without confusing respondents. Sometimes, a set of contingency questions is long enough to extend over several pages.

Victim surveys typically include many contingency questions. Figure 9.3 presents a few questions from the NCVS questionnaire used in 2004. All respondents are asked a series of screening questions to reveal possible victimizations. Persons who answer yes to any of the screening questions then complete a crime incident report that presents a large number of items designed to measure details of the victimization incident.

As Figure 9.3 shows, the crime incident report itself also contains contingency questions. You might notice that even this brief adaptation from the NCVS screening and crime incident report questionnaires is rather complex. NCVS questionnaires are administered primarily through computer-assisted telephone interviews in which the flow of contingency questions is more or less automated. It would be difficult to construct a self-administered victimization questionnaire with such complicated contingency questions.

Matrix Questions

Often, researchers want to ask several questions that have the same set of answer categories. This happens whenever the Likert response categories are used. Then it is often possible to construct a matrix of items and answers, as illustrated in Figure 9.4.

This format has three advantages. First, it uses space efficiently. Second, respondents probably find it easier to complete a set of questions presented in this fashion. In addition, this format may increase the comparability of
Screening Question:

36a. I’m going to read you some examples that will give you an idea of the kinds of crimes this study covers. As I go through them, tell me if any of these happened to you in the last 6-months, that is, since [date].

Was something belonging to YOU stolen, such as-

(a) Things that you carry, like luggage, a wallet, purse, briefcase, book-
(b) Clothing, jewelry or calculator-
(c) Bicycle or sports equipment-
(d) Things in your home like a TV, stereo, or tools-
(e) Things outside your home, such as a garden hose or lawn furniture-
(f) Things belonging to children in the household-
(g) Things from a vehicle, such as a package, groceries, camera, or tapes-
(h) Did anyone ATTEMPT to steal anything belonging to you?

Crime Incident Report:

20a. Were you or any other member of this household present when this incident occurred?

___ Yes [ask item 20b]
___ No [skip to 56, page 8]

20b. Which household members were present

___ Respondent only [ask item 21]
___ Respondent and other household member(s) [ask item 21]
___ Only other household member(s) [skip to 59, page 8]

21. Did you personally see an offender?

___ Yes
___ No

..........

56. Do you know or have you learned anything about the offender(s)—for instance, whether there was one or more than one offender involved, whether it was someone young or old, or male or female?

___ Yes [ask 57]
___ No [skip to 88, page 11]
responses given to different questions for the respondent as well as for the researcher. Because respondents can quickly review their answers to earlier items in the set, they might choose between, say, “strongly agree” and “agree” on a given statement by comparing their strength of agreement with their earlier responses in the set.

Some dangers are inherent in using this format as well. Its advantages may promote structuring an item so that the responses fit into the matrix format when a different, more idiosyncratic, set of responses might be more appropriate. Also, the matrix question format can generate a response set among some respondents. This means that respondents may develop a pattern of, say, agreeing with all the statements, without really thinking about what the statements mean. That is especially likely if the set of statements begins with several that indicate a particular orientation (for example, a conservative political perspective) and then offers only a few subsequent ones that represent the opposite orientation. Respondents might assume that all the statements represent the same orientation and, reading quickly, misread some of them, thereby giving the wrong answers. This problem can be reduced somewhat by alternating statements that represent different orientations and by making all statements short and clear.

A more difficult problem is when responses are generated through respondent boredom or fatigue. This can be avoided by keeping matrix questions and the entire questionnaire as short as possible. Later in this chapter, in the section on comparing different methods of questionnaire administration, we will describe a useful technique for avoiding response sets generated by respondent fatigue.

Ordering Questions in a Questionnaire
The order in which questions are asked can also affect the answers given. The content of one question can affect the answers given to later ones. For example, if several questions ask about the dangers of illegal drug use and then a question asks respondents to volunteer (open ended) what they believe to be the most serious crime problems in U.S. cities, drug use will receive more mentions than would otherwise be the case. In this situation, it is preferable to ask the open-ended question first.

If respondents are asked to rate the overall effectiveness of corrections policy, they will answer subsequent questions about specific aspects of correctional institutions in a way that is consistent with their initial assessment. The converse is true as well: If respondents are first asked specific questions about prisons and other correctional facilities, their subsequent overall assessment will be influenced by the earlier question.

The best solution is sensitivity to the problem. Although we cannot avoid the effect of question order, we should attempt to estimate what that effect will be. Then we will be able to interpret results in a meaningful fashion. If the order of questions seems an especially important issue in a given study, we could construct several versions of the questionnaire that contain the different possible orderings of questions. We could then determine the effects of ordering. At the very least, different versions of the questionnaire should be pretested.

The desired ordering of questions differs somewhat between self-administered questionnaires and interviews. In the former, it is usually best to begin the questionnaire with the most interesting questions. Potential respondents who glance casually at the first few questions should want to answer them. At the same time, however, the initial questions should be neither threatening nor sensitive. It might be a bad idea to begin with questions about sexual behavior or drug use. Requests for duller demographic data (age, gender, and the like) should generally be placed at the end of a self-administered questionnaire. Placing these questions at the beginning, as many inexperienced researchers are tempted to do,
gives the questionnaire the initial appearance of a routine form, and the person who receives it may not be motivated to complete it.

Just the opposite is generally true for in-person interview and telephone surveys. When the potential respondent’s door first opens, the interviewer must begin to establish rapport quickly. After a short introduction to the study, the interviewer can best begin by enumerating the members of the household, obtaining demographic data about each. Such questions are easily answered and generally nonthreatening. Once the initial rapport has been established, the interviewer can move into more sensitive areas. An interview that begins with the question “Do you ever worry about strangers appearing at your doorstep?” will probably end rather quickly.

Finally, it’s common for less experienced researchers to assume that questionnaires must be newly constructed for each application. In contrast, it’s almost always possible—and usually preferable—to use an existing questionnaire as a point of departure. See the box “Don’t Start from Scratch!” for more on this.

Self-Administered Questionnaires

Self-administered questionnaires are generally the least expensive and easiest to complete.

Although the mail survey is the traditional method used in self-administered studies, several other methods are also possible. In some cases, it may be appropriate to administer the

DON’T START FROM SCRATCH!

It’s always easier to modify an existing questionnaire for a particular research application than it is to start from scratch. It’s also difficult to imagine asking questions that nobody has asked before. Here are examples of websites that present complete questionnaires or batteries of questionnaire items.

- Bureau of Justice Statistics (BJS). In addition to administering the NCVS, the BJS collects information from a variety of justice organizations. Copies of recent questionnaires for all BJS-sponsored surveys are available at http://www.ojp.usdoj.gov/bjs/quest.htm.
- California Healthy Kids Survey. This set of questionnaires is useful for assessing behavior routines. Most include items on alcohol, tobacco, and other drug use; fighting; and other behaviors of potential interest for school-based interventions. English and Spanish versions are available for elementary, middle, and high school: http://www.wested.org/hks/.
- Centers for Disease Control (CDC). Various centers within the CDC regularly collect a variety of health-related data through questionnaires and other data collection systems. Copies of instruments are available at http://www.cdc.gov/nchs/express.htm.
- The Measurement Group. This provides links to questionnaires designed for use in public health studies, but many of these include items of potential interest to treatment-related initiatives: http://www.themeasurementgroup.com/evalbttn.htm.
- University of Surrey Question Bank. Maintained by a university in England, the Question Bank includes links to complete questionnaires for a wide variety of surveys conducted in the United Kingdom and other countries. You can find a master list of surveys, or browse questionnaires by topic. An excellent resource: http://qb.soc.surrey.ac.uk/docs/home.htm.

Source: Adapted from Maxfield (2001).
questionnaire to a group of respondents gathered at the same place at the same time, such as police officers at roll call or prison inmates at some specially arranged assembly. Or probationers might complete a questionnaire when they report for a meeting with their probation supervisor. The Monitoring the Future survey (see Chapter 6) has high school seniors complete self-administered questionnaires in class.

Some experimentation has been conducted on the home delivery of questionnaires. A research worker delivers the questionnaire to the home of sample respondents and explains the study. Then the questionnaire is left for the respondent to complete, and the researcher picks it up later.

Home delivery and the mail can be used in combination as well. Questionnaires can be mailed to families, and then research workers may visit the homes to pick up the questionnaires and check them for completeness. In the opposite approach, survey packets are hand delivered by research workers with a request that the respondents mail the completed questionnaires to the research office. In general, when a research worker delivers the questionnaire, picks it up, or both, the completion rate is higher than for straightforward mail surveys.

More recently, the explosive growth of the Internet has made it possible to have respondents complete self-administered questionnaires online. Before discussing web-based questionnaires, let us turn our attention to the fundamentals of mail surveys, still the most typical form of self-administration used by researchers.

**Mail Distribution and Return**

The basic method for collecting data through the mail is transmittal of a questionnaire accompanied by a letter of explanation and a self-addressed, stamped envelope for returning the questionnaire. You have probably received a few. As a respondent, you are expected to complete the questionnaire, put it in the envelope, and mail it back. If, by any chance, you have received such a questionnaire and failed to return it, it would be a valuable exercise for you to recall your reasons for not returning it—and keep those in mind any time you plan to send questionnaires to others.

One big reason people do not return questionnaires is that it seems like too much trouble. To overcome this problem, researchers have developed ways to make the return of questionnaires easier. One method involves a self-mailing questionnaire that requires no return envelope. The questionnaire is designed so that when it is folded in a particular fashion, the return address appears on the outside. That way, the respondent doesn’t have to worry about losing the envelope.

Timing is important to consider in the actual mailing of questionnaires. In most cases, the holiday months of November, December, and January should be avoided. Overall mail volume is greatest during those periods, which can substantially slow down the process of both distribution and return of questionnaires. And because a greater volume of mail is flowing through post offices, people receive more mail of all types. If a questionnaire arrives in the company of glossy gift catalogs, holiday greetings, bills, and assorted junk mail, respondents will be more likely to discard the survey packet.

**Warning Mailings and Cover Letters**

The U.S. population, especially that proportion residing in urban areas, is becoming increasingly mobile. This mobility, together with the fact that sampling frames used to obtain addresses may be dated, has prompted researchers to use warning mailings for the purpose of verifying, or “cleaning,” addresses. Certain types of warning mailings can also be effective in increasing response rates.

Warning mailings work like this: After researchers generate a sample, they send a postcard to each selected respondent, with the no-
tation “Address correction requested” printed on the postcard. If the addressee has moved and left a forwarding address, the actual questionnaire is sent to the new address. In cases in which someone has moved and not left a forwarding address, or more than a year has elapsed and the post office no longer has information about a new address, the postcard is returned marked something like “Addressee unknown.” Selected persons who still reside at the original listed address are “warned” in suitable language to expect a questionnaire in the mail. In such cases, postcards should briefly describe the purpose of the survey for which the respondent has been selected.

Warning letters can be more effective than postcards in increasing response rates, and they can also serve the purpose of cleaning addresses. Letters printed on letterhead stationery can present a longer description of the survey’s purpose and a more reasoned explanation of why it is important for everyone to respond. Cover letters accompanying the actual questionnaire offer a similar opportunity to increase response rates. Two features of cover letters warrant some attention. First, the content of the letter is obviously important. The message should communicate why a survey is being conducted, how and why the respondent was selected, and why it is important for the respondent to complete the questionnaire. In line with our discussion of the protection of human subjects in Chapter 3, the cover letter should also assure respondents that their answers will be confidential.

Second, the cover letter should identify the institutional affiliation or sponsorship of the survey. The two alternatives are (1) some institution that the respondent respects or can identify with or (2) a neutral but impressive-sounding affiliation. For example, if we are conducting a mail survey of police chiefs, printing our cover letter on International Association of Chiefs of Police (IACP) stationery and having the letter signed by an official in the IACP might increase the response rate. Of course, we cannot adopt such a procedure unless the survey actually is endorsed by the IACP.

By the same token, it is important to avoid controversial affiliations or those inappropriate for the target population. The National Organization for Reform of Marijuana Laws, for instance, is not suitable for most target populations. A university affiliation is appropriate in many cases, unless the university is on bad terms with the target population.

Follow-Up Mailings

Follow-up mailings may be administered in a number of ways. In the simplest, nonrespondents are sent a letter of additional encouragement to participate. A better method, however, is to send a new copy of the survey questionnaire with the follow-up letter. If potential respondents have not returned their questionnaires after 2 or 3 weeks, the questionnaires probably have been lost or misplaced.

The methodological literature on follow-up mailings strongly suggests that they are effective in increasing return rates in mail surveys. In general, the longer a potential respondent delays replying, the less likely he or she is to do so at all. Properly timed follow-up mailings provide additional stimuli to respond.

The effects of follow-up mailings may be seen by monitoring the number of questionnaires received over time. Initial mailings will be followed by a rise in and subsequent subsiding of returns, and follow-up mailings will spur a resurgence of returns. In practice, three mailings (an original and two follow-ups) are the most effective.

Acceptable Response Rates

A question frequently asked about mail surveys concerns the percentage return rate that should be achieved. Note that the body of inferential statistics used in connection with survey analysis assumes that all members of the initial sample complete and return their questionnaires. Because this almost never happens, response bias becomes a concern. Researchers
must test (and hope for) the possibility that respondents look essentially like a random sample of the initial sample and thus a somewhat smaller random sample of the total population. For example, if the gender of all people in the sample is known, a researcher can compare the percentages of males and females indicated on returned questionnaires with the percentages for the entire sample.

Nevertheless, overall response rate is one guide to the representativeness of the sample respondents. If the response rate is high, then there is less chance of significant response bias than if the rate is low. But what is a high response rate?

A quick review of the survey literature uncovers a wide range of response rates. Each may be accompanied by a statement like “This is regarded as a relatively high response rate for a survey of this type.” Even so, it’s possible to state some rules of thumb about return rates. A response rate of at least 50 percent is adequate for analysis and reporting. A response of at least 60 percent is good. And a response rate of 70 percent is very good.

Bear in mind, however, that these are only rough guides; they have no statistical basis, and a demonstrated lack of response bias is far more important than a high response rate. Response rates tend to be higher for surveys that target a narrowly defined population; general-population surveys yield lower response rates.

As you can imagine, one of the more persistent discussions among survey researchers concerns ways to increase response rates. Ingenious techniques have been developed to address this problem. Some researchers have experimented with novel formats. Others have tried paying respondents to participate. The problem with paying, of course, is that it’s expensive to make meaningful payments to hundreds or thousands of respondents. A compromise approach is to include all respondents in a lottery where they have a chance to win one prize of some value. You may recognize this as a common strategy for customer-satisfaction and other polls—if you participate, you have a chance to win a vacation to Las Vegas or some such prize of equal value.

Don Dillman (1999) has undertaken an extensive review of the various techniques survey researchers use to increase return rates on mail surveys, and he evaluates the impact of each. More importantly, Dillman stresses the necessity of paying attention to all aspects of the study—what he calls the “total design method”—rather than one or two special gimmicks.

**Computer-Based Self-Administration**

Advances in computer and telecommunication technology over the past several decades have produced additional options for distributing and collecting self-administered questionnaires. Jeffrey Walker (1994) describes variations on conducting surveys by fax machine. Questionnaires are faxed to respondents, who are asked to fax their answers back. As the Internet and web have permeated work and leisure activities, different types of computer-assisted self-administered surveys have become more common.

David Shannon and associates (Shannon, Johnson, Scarcy, and Lott, 2002) describe three general types of electronic surveys. The first is a disk-based survey. Respondents load a questionnaire from a disk or CD into their own computer, key in responses to survey items, and then either mail the disk back to researchers or transmit the information electronically. As the earliest form of electronic survey, the disk-based survey is a relic of stand-alone personal computers. Disk-based surveys are now virtually obsolete as personal computers are now routinely connected to the web in one way or another.

The second type, e-mail surveys, has a few variations. Researchers can include a few simple questions in an e-mail message and ask respondents to reply by e-mail. More elaborate versions can embed complex formatted question-
naires in e-mail messages. Respondents might be asked to open an attached file that contains a questionnaire, or they might be directed to another web page that contains a formatted questionnaire. That brings us to the third type of electronic survey described by Shannon and associates—a questionnaire posted on a web page.

The advantages of this method are obvious. Responses are automatically recorded in computer files, saving time and money. Web page design tools make it possible to create attractive questionnaires that include contingency questions, matrices, and other complex tools for presenting items to respondents. Dillman, long recognized for his total design approach to conducting mail surveys, has written a comprehensive guide to conducting mail, web-based, and other self-administered surveys.

All electronic versions of self-administered questionnaires face a couple of problems. The first concerns representativeness: Will the people who can be surveyed online be representative of meaningful populations, such as all U.S. adults, all registered voters, or all residents of particular urban neighborhoods? This criticism has also been raised with regard to surveys via fax and, in the mid-twentieth century, with regard to telephone surveys. Put in terms that should be familiar from the previous chapter, how closely do available sampling frames for electronic surveys match possible target populations? If, for example, our target population is university students, how can we obtain a list of e-mail addresses or other identifiers that will enable us to survey a representative sample? It’s easy to think of other target populations of interest to criminal justice researchers that might be difficult to reach via e-mail or web-based questionnaires.

The second problem is an unfortunate consequence of the rapid growth of e-mail and related technologies. Just as junk mail clutters our physical mailboxes with all sorts of advertising, “spam” and other kinds of unwanted messages pop up all too often in our virtual mailboxes. The proliferation of junk e-mail has led to the development of antispam filters that screen out unwanted correspondence. Unfortunately, such programs also can screen out unfamiliar but well-meaning mail such as e-mail questionnaires. Similar problems with telemarketing have made it increasingly difficult to conduct surveys by telephone.

In yet another example of technology advances being accompanied by new threats, the spread of computer viruses has made people even more cautious about opening e-mail or attachments from unfamiliar sources. This problem, and the electronic version of junk mail, can be addressed in a manner similar to warning mailings for printed questionnaires. Before sending an e-mail with an embedded or linked questionnaire, researchers can distribute e-mail messages from trusted sources that warn recipients to expect to receive a questionnaire, and urge them to complete it.

We should keep one basic principle in mind when considering whether a self-administered questionnaire can be distributed electronically: Web-based surveys depend on access to the web, which, of course, implies having a computer. The use of computers and the web continues to increase rapidly. Although access to this technology still is unequally distributed across socioeconomic classes, web-based surveys can be readily conducted for many target populations of interest to criminal justice researchers.

Recalling our discussion in Chapter 8, the correspondence between a sampling frame and target population is a crucial feature of sampling. Most justice professionals and criminal justice organizations routinely use the web and e-mail. Lower-cost but generalizable victim surveys can use web-based samples of university students to distribute questionnaires. Or printed warning letters can be mailed, inviting respondents to complete either traditional or e-mail self-administered questionnaires. Just as e-mail, electronic bill-paying, and other transactions have replaced letters, check-writing, and other correspondence, self-administered
surveys will increasingly be conducted on web-based computers. At the end of this chapter we describe a small sample of resources for conducting web-based surveys.

**In-Person Interview Surveys**

*Face-to-face interviews are best for complex questionnaires and other specialized needs.*

The in-person interview is an alternative method of collecting survey data. Rather than asking respondents to read questionnaires and enter their own answers, researchers send interviewers to ask the questions orally and record respondents’ answers. Most interview surveys require more than one interviewer, although a researcher might undertake a small-scale interview survey by him- or herself.

**The Role of the Interviewer**

In-person interview surveys typically attain higher response rates than mail surveys do. Respondents seem more reluctant to turn down an interviewer who is standing on their doorstep than to throw away a mail questionnaire. A properly designed and executed interview survey ought to achieve a completion rate of at least 80–85 percent.

The presence of an interviewer generally decreases the number of “don’t know” and “no answer” responses. If minimizing such responses is important to the study, the interviewer can be instructed to probe for answers (“If you had to pick one of the answers, which do you think would come closest to your feelings?”).

The interviewer can also help respondents with confusing questionnaire items. If the respondent clearly misunderstands the intent of a question, the interviewer can clarify matters and thereby obtain a relevant response. Such clarifications must be strictly controlled, however, through formal specifications.

Finally, the interviewer can observe as well as ask questions. For example, the interviewer can make observations about the quality of the dwelling, the presence of various possessions, the respondent’s ability to speak English, and the respondent’s general reactions to the study. Interviewers for the BCS and for a 1983 supplement to the NCVS made detailed observations of the physical conditions surrounding each respondent’s home. Later analysis compared interviewer ratings of such things as litter and graffiti with respondent perceptions of the same problems (Maxfield, 1987).

Survey research is, of necessity, based on an unrealistic stimulus–response theory of cognition and behavior. That is, it is based on the assumption that a questionnaire item will mean the same thing to every respondent, and every given response must mean the same thing when given by different respondents. Although this is an impossible goal, survey questions are drafted to approximate the ideal as closely as possible. The interviewer also plays a role in this ideal situation. The interviewer’s presence should not affect a respondent’s perception of a question or the answer given. The interviewer, then, should be a neutral medium through which questions and answers are transmitted. If this goal is met, different interviewers will obtain the same responses from a given respondent, an example of reliability in measurement (see Chapter 5).

**General Rules for Interviewing**

The way interviews ought to be conducted will vary somewhat by the survey population and survey content. Nevertheless, some general guidelines apply to most interviewing situations, including sample surveys and more specialized interviewing.

**Appearance and Demeanor**  As a general rule, the interviewer should dress in a fashion similar to that of the people he or she will be interviewing. To the extent that the interviewer’s dress and grooming differ from that of the respondents, it should be in the direction of cleanliness, neatness, and modesty. Although middle-class standards of dress and grooming are not accepted by all sectors of American society, they remain the norm and are likely
to be acceptable to the largest number of respondents.

The importance of considering the target population cannot be overemphasized. “Establishment” attire should be the norm when interviewing most criminal justice officials. However, the middle-class norm in clothing may foster suspicion among street people or incarcerated individuals.

In demeanor, interviewers should be pleasant if nothing else. Because they will be prying into the respondent’s personal life and thoughts, they must communicate a genuine interest in getting to know the respondent without appearing to spy. They must be relaxed and friendly without being too casual or clinging. Good interviewers also have the ability to quickly determine the kind of person the respondent feels most comfortable with or most enjoys talking to, and they adapt accordingly.

**Familiarity with the Questionnaire** The interviewer must be able to read the questionnaire items to respondents without stumbling over words and phrases. A good model for interviewers is the actor reading lines in a play or film. The interviewer must read the questions as though they are part of a natural conversation, but that “conversation” must precisely follow the language set down in the question.

By the same token, the interviewer must be familiar with the specifications for administering the questionnaire. Inevitably, some questions will not exactly fit a given respondent’s situation, and the interviewer must determine how those questions should be interpreted in that situation. The specifications provided to the interviewer should include adequate guidelines in such cases, but the interviewer must know the organization and contents of the specifications well enough to refer to them efficiently.

**Probing for Responses** Probes are frequently required to elicit responses to open-ended questions. For example, to a question about neighborhood crime problems, the respondent might simply reply, “Pretty bad.” The interviewer could obtain an elaboration on this response through a variety of probes. Sometimes, the best probe is silence; if the interviewer sits quietly with pencil poised, the respondent will probably fill the pause with additional comments. Appropriate verbal probes are “How is that?” and “In what ways?” Perhaps the most generally useful probe is “Anything else?”

In every case, however, it is imperative that the probe be completely neutral. The probe must not in any way affect the nature of the subsequent response. If we anticipate that a given question may require probing for appropriate responses, we should write one or more useful probes next to the item in the questionnaire. This practice has two important advantages. First, it allows for more time to devise the best, most neutral probes. Second, it ensures that all interviewers will use the same probes as needed. Thus, even if the probe is not perfectly neutral, the same stimulus is presented to all respondents. This is the same logical guideline as for question wording. Although a question should not be loaded or biased, it is essential that every respondent be presented with the same question, even if a biased one.

**Coordination and Control** Whenever more than one interviewer will administer a survey, it is essential that the efforts be carefully coordinated and controlled. Two ways to ensure this control are by (1) training interviewers and (2) supervising them after they begin work.

Whether the researchers will be administering a survey themselves or paying a professional firm to do it for them, they should be attentive to the importance of training interviewers. The interviewers usually should know what the study is all about. Even though the interviewers may be involved only in the data collection phase of the project, they should understand what will be done with the information they gather and what purpose will be served.

There may be some exceptions to this, however. For example, in their follow-up study of
child abuse victims and controls, Cathy Spatz Widom and associates (Widom, Weiler, and Cotler, 1999) did not inform the professional interviewers who gathered data that their interest was in the long-term effects of child abuse. This safeguard was used to avoid even the slightest chance that interviewers’ knowledge of the study focus would affect how they conducted interviews.

Obviously, training should ensure that interviewers understand the questionnaire. Interviewers should also be clear on procedures to select respondents from among household members. And interviewers should recognize circumstances in which substitute sample elements may be used in place of addresses that no longer exist, families who have moved, or persons who simply refuse to be interviewed.

Training should include practice sessions in which interviewers administer the questionnaire to one another. The final stage of the training should involve some “real” interviews conducted under conditions like those in the actual survey.

While interviews are being conducted, it is a good idea to review questionnaires as they are completed. This may reveal questions or groups of questions that respondents do not understand. Alternatively, reviewing completed questionnaires can signal that an interviewer is encountering difficulties.

Computer-Assisted In-Person Interviews

Just as e-mail and web-based surveys apply new technology to the gathering of survey data through self-administration, laptop and handheld computers are increasingly being used to conduct in-person interviews. Different forms of computer-assisted interviewing (CAI) offer major advantages in the collection of survey data. At the same time, CAI has certain disadvantages that must be considered. We’ll begin by describing an example of how this technology was adopted in the BCS, one of the earliest uses of CAI in a general-purpose crime survey.

CAI in the British Crime Survey

Previous waves of the BCS, a face-to-face interview survey, asked respondents to complete a self-administered questionnaire about drug use, printed as a small booklet that was prominently marked “Confidential.” Beginning with the 1994 survey, respondents answered self-report questions on laptop computers. The BCS includes two related versions of CAI. In computer-assisted personal interviewing (CAPI), interviewers read questions from computer screens and then key in respondents’ answers. For self-report items, interviewers hand the computers to subjects, who then key in the responses themselves.

This approach is known as computer-assisted self-interviewing (CASI). In addition, CASI as used in the BCS is supplemented with audio instructions—respondents listen to interview prompts on headphones connected to the computer. After subjects key in their responses to self-report items, the answers are scrambled so the interviewer cannot access them. Notice how this feature of CASI enhances the researcher’s ethical obligation to keep responses confidential.

Malcolm Ramsay and Andrew Percy (1996) report that CASI has at least two benefits. First, respondents seem to sense a greater degree of confidentiality when they respond to questions on a computer screen as opposed to questions on a written form. Second, in its first few years of use the laptop computers were something of a novelty that stimulated respondents’ interest; this was especially true for younger respondents.

Examining results from the BCS reveals that CASI techniques produced higher estimates of illegal drug use than those revealed in previous surveys. Table 9.1 compares self-reported drug use from the 1998 BCS (Ramsey and Partridge, 1999) with results from the 1992 BCS (Mott and Mirrlees-Black, 1995), in which respondents answered questions in printed booklets. We present results for only three drugs here, together with tabulations about the use of
any drug. For each drug, the survey measured lifetime use (“Ever used?”) and use in the past 12 months. Notice that rates of self-reported use were substantially higher in 1998 than in 1992, with the exception of “semeron” use, reported by very few respondents in 1992 and none in 1998. If you’ve never heard of semeron, you’re not alone. It’s a fictitious drug, included in the list of real drugs to detect untruthful or exaggerated responses. If someone confessed to using semeron, his or her responses to other self-reported items would be suspect. Notice in Table 9.1 that CASI use in 1998 reduced the number of respondents who admitted using a drug that doesn’t exist.

Table 9.1  Self-Reported Drug Use, 1992 and 1998 British Crime Survey

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Marijuana or cannabis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ever used?</td>
<td>24</td>
<td>42</td>
</tr>
<tr>
<td>Used in previous 12 months?</td>
<td>12</td>
<td>23</td>
</tr>
<tr>
<td>Amphetamines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ever used?</td>
<td>9</td>
<td>20</td>
</tr>
<tr>
<td>Used in previous 12 months?</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Semeron</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ever used?</td>
<td>0.3</td>
<td>0.0</td>
</tr>
<tr>
<td>Used in previous 12 months?</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Any drug</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ever used?</td>
<td>28</td>
<td>49</td>
</tr>
<tr>
<td>Used in previous 12 months?</td>
<td>14</td>
<td>25</td>
</tr>
</tbody>
</table>


Advantages and Disadvantages  Different types of CAI offer a number of advantages for in-person interviews. The BCS and other surveys that include self-report items indicate that CAI is more “productive” in that self-reports of drug use and other offending tend to be higher. In 1999, the National Household Survey on Drug Abuse shifted completely to CAI (Wright, Barker, Gfroerer, and Piper, 2002). Other advantages include the following:

- Responses can be quickly keyed in and automatically reformatted into data files for analysis.
- Complex sequences of contingency questions can be automated. Instead of printing many examples of “If answer is yes, go to question 43a; if no . . . ,” computer-based questionnaires automatically jump to the next appropriate question contingent on responses to earlier ones.
- CAI offers a way to break up the monotony of a long interview by shifting from verbal interviewer prompts to self-interviewing, with or without an audio supplement.
- Questionnaires for self-interviewing can be programmed in different languages, readily switching to the language appropriate for a particular respondent.
- Audio-supplemented CASI produces a standardized interview, avoiding any bias that might emerge from interviewer effects.
- Audio supplements, in different languages, facilitate self-interviews of respondents who cannot read.

At the same time, CAI has certain disadvantages that preclude its use in many survey applications:

- Although CAI reduces costs in data processing, it requires more up-front investment in programming questionnaires, skip sequences, and the like.
- Automated skip sequences for contingency questions are great, but if something goes wrong with the programmed questionnaire,
all sorts of subsequent problems are possible. As Emma Forster and Alison McCleery (1999) point out, such question-routing mistakes might mean whole portions of a questionnaire are skipped. Whereas occasional random errors are possible with pen-and-paper interviews, large-scale systematic error can occur with CAI technology.

- Proofreading printed questionnaires is straightforward, but it can be difficult to audit a computerized questionnaire. Doing so might require special technical skills.
- It can be difficult to print and archive a complex questionnaire used in CAI. This was a problem with early applications of CAI technology, but improvements in software are helping to solve it.
- Batteries of laptops run down, and computers and software are more vulnerable to malfunctions and random weirdness than are stacks of printed questionnaires.

In sum, CAI can be costly and requires some specialized skills. As a result, these and related technologies are best suited for use by professional survey researchers or research centers that regularly conduct large-scale in-person interviews. We will return to this issue in the concluding section of this chapter.

**Telephone Surveys**

*Telephone surveys are fastest and relatively low cost.*

Telephone surveys have many advantages that make them a popular method. Probably the greatest advantages involve money and time. In a face-to-face household interview, a researcher may drive several miles to a respondent’s home, find no one there, return to the research office, and drive back the next day—possibly finding no one there again. It’s cheaper and quicker to call.

Interviewing by telephone, researchers can dress any way they please, and it will have no effect on the answers respondents give. It may be possible to probe into more sensitive areas, although that is not necessarily the case. People are, to some extent, more suspicious when they can’t see the person asking them questions—perhaps a consequence of telemarketing and salespeople conducting bogus surveys prior to making sales pitches.

Telephone surveys can give a researcher greater control over data collection if several interviewers are engaged in the project. If all the interviewers are calling from the research office, they can get clarification from the supervisor whenever problems occur, as they inevitably do. Alone in the field, an interviewer may have to wing it between weekly visits with the interviewing supervisor.

A related advantage is rooted in the growing diversity of U.S. cities. Because many major cities have growing immigrant populations, interviews may need to be conducted in different languages. Telephone interviews are usually conducted from a central site, so that one or more multilingual interviewers can be quickly summoned if an English-speaking interviewer makes contact with, say, a Spanish-speaking respondent. In-person interview surveys present much more difficult logistical problems in handling multiple languages. And mail surveys require printing and distributing questionnaires in different languages.

Another important factor in the increasing use of telephone surveys is personal safety. Dillman (1978:4) described perceptions of risk in in-person interviews:

> Interviewers must be able to operate comfortably in a climate in which strangers are viewed with distrust and must successfully counter respondents’ objections to being interviewed. Increasingly, interviewers must be willing to work at night to contact residents in many households. In some cases, this necessitates providing protection for interviewers working in areas of a city in which a definite threat to the safety of individuals exists.
Concerns for safety can work two ways to hamper face-to-face interviews. First, potential respondents may refuse to be interviewed because they fear the stranger-interviewer. Second, the interviewers themselves may be in danger. Telephone surveys eliminate risks to interviewers and reduce the concerns of respondents who may be uneasy about opening their doors to strangers.

Telephone interviewing has its problems, however. Telephone surveys are limited by definition to people who own telephones. Years ago, this method produced a substantial social class bias by excluding poor people. Over time, however, the telephone has become a standard fixture in almost all American homes. The U.S. Census Bureau estimates that 95.5 percent of all households now have telephones, so the earlier class bias has been substantially reduced (U.S. Bureau of the Census, 2006:Table 1117). The NCVS, traditionally an in-person interview, has increased its use of telephone interviews as part of the crime survey’s redesign.

At the same time, phone surveys are much less suitable for individuals not living in households. Homeless people are obvious examples; those who live in institutions are also difficult to reach out and touch via telephone. Patricia Tjaden and Nancy Thoennes (2000) cite single-adult households and people living in rural or inner-city areas as targets least likely to have telephone coverage.

A related sampling problem involves unlisted numbers. If the survey sample is selected from the pages of a local telephone directory, it totally omits all those people who have requested that their numbers not be published. Also, as we mentioned in Chapter 8, recent movers and transient residents are not well represented in published telephone directories. This potential bias is eliminated through random-digit dialing (RDD), a technique that has advanced telephone sampling substantially.

RDD samples use computer algorithms to generate lists of random telephone numbers—usually, the last four digits. This procedure gets around the sampling problem of unlisted telephone numbers but may substitute an administrative problem. Randomly generating phone numbers produces numbers that are not in operation or that serve a business establishment or pay phone. In most cases, businesses and pay phones are not included in the target population; dialing these numbers and learning that they’re out of scope will take time away from producing completed interviews with the target population.

Weisel (1999) offers an excellent description of RDD samples for use in community crime surveys. Among the increasingly important issues in RDD samples is the growing number of telephone numbers in use for cell phones, pagers, and home access to the Internet. This means that the rate of ineligible telephone numbers generated through RDD is increasing. Weisel’s rule of thumb is that ineligible numbers account for an estimated 60 percent of phone numbers produced by typical RDD procedures. Thus, if RDD is used, researchers should plan on making five calls to reach two households.

The ease with which people can hang up is, of course, another shortcoming of telephone surveys. Once a researcher is inside someone’s home for an interview, that person is unlikely to order the researcher out of the house in mid-interview. It’s much easier to terminate a telephone interview abruptly, saying something like “Whoops! Someone’s at the door. Gotta go” or simply hanging up.

The most difficult challenge with telephone surveys involves the explosion of telemarketing. The volume of junk phone calls rivals that of junk mail, and salespeople often begin their pitch by describing a “survey.” The viability of legitimate surveys is now hampered by the proliferation of bogus surveys, which are actually sales campaigns disguised as research. As if that weren’t bad enough, telemarketing has become so annoying that many people simply hang up whenever they hear a strange voice announcing some institutional affiliation.
Partly as a consequence, residential phone customers seek to exercise greater control over incoming calls through caller identification (ID), phone number blocking, and other new services. Such devices enable phone users to screen calls before picking up the receiver, making it more difficult for legitimate survey researchers to make contact. Research by Michael Link and Robert Oldendick (1999) indicates that caller ID does reduce interview completion rates.

Following action in several states, the Federal Trade Commission approved new regulations restricting telemarketing by enabling individuals to place their phone numbers on a national do-not-call registry (“Telemarketing Sales Rule,” 2003). The rule went into effect in October 2003 after withstanding vigorous legal challenges by the telemarketing industry. The impact of this regulation remains to be seen. On the one hand, it has the potential to make life easier for legitimate telephone survey researchers. However, because conducting research is identified as a legitimate activity under the Telemarketing Sales Rule, it may induce more unscrupulous firms to disguise their sales calls as surveys.

Computer-Assisted Telephone Interviewing

Much of the growth in telemarketing has been fueled by advances in computer and telecommunications technology. Computers generate and dial phone numbers (in some cases, computers even control recorded sales pitches); operators enter customer and order information in forms displayed on computer screens. Beginning in the 1980s, much of the same technology came to be widely used in telephone surveys, referred to as computer-assisted telephone interviewing (CATI). Here’s how it works.

Interviewers wearing telephone headsets sit at computer workstations. Computer programs dial sampled phone numbers, which can be either generated through RDD or extracted from a database of phone numbers compiled from some source. As soon as phone contact is made, the computer screen displays an introduction (“Hello, my name is . . . calling from the Survey Research Center at Ivory Tower University”) and the first question to be asked, often a query about the number of residents who live in the household. As interviewers key in answers to each question, the computer program displays a new screen that presents the next question, until the end of the interview is reached.

CATI systems offer several advantages over procedures in which an interviewer works through a printed interview schedule. Speed is one obvious plus. Forms on computer screens can be filled in more quickly than can paper forms. Typing answers to open-ended questions is much faster than writing them by hand. And CATI software immediately formats responses into a data file as they are keyed in, which eliminates the step of manually transferring answers from paper to computer.

Perhaps you’ve occasionally marveled at newspaper stories that report the results of a nationwide opinion poll the day after some major speech or event. The speed of CATI technology, coupled with RDD, makes these instant poll reports possible.

Accuracy is also enhanced by CATI systems in several ways. First, CATI programs can be designed to accept only valid responses to a given questionnaire item. For example, if valid responses to respondent “gender” are f for female and m for male, the computer will accept only those two letters, emitting a disagreeable noise and refusing to proceed if something else is keyed in. Second, the software can be programmed to automate contingency questions and skip sequences, thus ensuring that the interviewer skips over inappropriate items and rapidly gets to the next appropriate question. This can be especially handy in a victim survey, in which affirmative answers to screening questions automatically bring up detailed questions about each crime incident.
Comparison of the Three Methods

Cost, speed, and question content are issues to consider in selecting a survey method.

We’ve now examined three ways of collecting survey data: self-administered questionnaires, in-person interviews, and telephone surveys. We have also considered some recent advances in each mode of administration. Although we’ve touched on some of the relative advantages and disadvantages of each, let’s take a minute to compare them more directly.

Self-administered questionnaires are generally cheaper to use than interview surveys. Moreover, for self-administered e-mail or web-based surveys, it costs no more to conduct a national survey than a local one. Obviously, the cost difference between a local and a national in-person interview survey is considerable. The cost of telephone surveys continues to decline, as Internet-based calling services eliminate long-distance toll charges. Mail surveys typically require a small staff. One person can conduct a reasonable mail survey, although it is important not to underestimate the work involved. Web-based surveys can be very inexpensive, particularly because they combine data collection and tabulation.

Up to a point, cost and speed are inversely related. In-person interview surveys can be completed very quickly if a large pool of interviewers is available and funding is adequate to pay them. In contrast, if a small number of people are conducting a larger number of face-to-face interviews, costs are generally lower, but the survey takes much longer to complete. Telephone surveys that use CATI technology are easily the fastest. In comparing different modes of self-administered surveys, Nojin Kwak and Barry Radler (2002) report faster responses from web-based surveys, but higher completion rates from mailed questionnaires.

Self-administered surveys may be more appropriate to use with especially sensitive issues if the surveys offer complete anonymity. Respondents are sometimes reluctant to report controversial or deviant attitudes or behaviors in interviews, but they may be willing to respond to an anonymous self-administered questionnaire. However, the successful use of computers for self-report items in the BCS and the National Household Survey on Drug Abuse suggests that interacting with a machine can promote more candid responses. This is supported by experimental research comparing different modes of questionnaire administration (Tourangeau and Smith, 1996).

Interview surveys have many advantages, too. For example, in-person or telephone surveys are more appropriate when respondent literacy may be a problem. Interview surveys also result in fewer incomplete questionnaires. Although respondents may skip questions in a self-administered questionnaire, interviewers are trained not to do so. CAI offers a further check on this in telephone and in-person surveys.

Self-administered questionnaires may be more effective for sensitive issues, but interview surveys are definitely more effective in dealing with complicated ones. Interviewers can explain complex questions to respondents and use visual aids that are not possible in mail or phone surveys.

The 1984 BCS contained a good example of such techniques that did not depend on computer technology. The survey included a series of questions about attitudes toward neighborhood watch programs. Pretests of certain questionnaire items revealed that many people were unfamiliar with neighborhood watch. In conducting the actual survey, interviewers handed respondents a card that presented a brief description of neighborhood watch. Respondents were asked a battery of questions about their opinions of such programs after they had read the card.

In general, this may be called an “information–opinion” technique, in which each
respondent is first exposed to a standard description of some object and is then asked to give an opinion about the object. This technique cannot be used effectively in self-administered questionnaires because respondents might skip over a lengthy written description and begin to answer questions right away.

In-person interviews, especially with computer technology, can also help reduce response sets. Respondents (like students?) eventually become bored listening to a lengthy series of similar types of questions. It’s easier to maintain individuals’ interest by changing the kind of stimulation they are exposed to. A mix of questions verbalized by a person, presented on a computer screen, and heard privately through earphones is more interesting for respondents and reduces fatigue.

As mentioned earlier, interviewers who question respondents face to face are also able to make important observations aside from responses to questions asked in the interview. In a household interview, they may summarize characteristics of the neighborhood, the dwelling unit, and so forth. They may also note characteristics of the respondents or the quality of their interaction with the respondents—whether the respondent had difficulty communicating, was hostile, seemed to be lying, and so on. Finally, when the safety of interviewers is an issue, a mail or phone survey may be the best option.

Ultimately, researchers must weigh all these advantages and disadvantages of the three methods against research needs and available resources.

**Strengths and Weaknesses of Survey Research**

*Surveys tend to be high on reliability and generalizability, but validity can often be a weak point.*

Like other modes of collecting data in criminal justice research, surveys have strengths and weaknesses. It is important to consider these in deciding whether the survey format is appropriate for a specific research purpose.

Surveys are particularly useful in describing the characteristics of a large population. The NCVS has become an important tool for researchers and public officials because of its ability to describe levels of crime. A carefully selected probability sample, in combination with a standardized questionnaire, allows researchers to make refined descriptive statements about a neighborhood, a city, a nation, or some other large population.

Standardized questionnaires have an important advantage in regard to measurement. Earlier chapters discussed the ambiguous nature of concepts: They ultimately have no real meanings. One person’s view about, say, crime seriousness or punishment severity is quite different from another’s. Although we must be able to define concepts in ways that are most relevant to research goals, it’s not always easy to apply the same definitions uniformly to all subjects. Nevertheless, the survey researcher is bound to the requirement of having to ask exactly the same questions of all subjects and having to impute the same intent to all respondents giving a particular response.

At the same time, survey research has its weaknesses. First, the requirement for standardization might mean that we are trying to fit round pegs into square holes. Standardized questionnaire items often represent the least common denominator in assessing people’s attitudes, orientations, circumstances, and experiences. By designing questions that are at least minimally appropriate to all respondents, we may miss what is most appropriate to many respondents. In this sense, surveys often appear superficial in their coverage of complex topics.

Similarly, survey research cannot readily deal with the specific contexts of social life. Although questionnaires may provide information in this area, the survey researcher seldom develops a feel for the total life situation in which respondents are thinking and acting. This is in contrast to the experiences of
Chapter 9  Survey Research and Other Ways of Asking Questions

the participant observer (Chapter 10). Many researchers are dissatisfied with standard survey questions about important criminal justice concepts such as fear of crime, routine behavior, and crime prevention. NCVS procedures for counting series victimizations remain inadequate even after 30 years’ experience with the survey.

Although surveys are flexible in the sense mentioned earlier, they cannot be changed once interviewing has begun. Studies that involve direct observation can be modified as field conditions warrant, but surveys typically require that an initial study design remain unchanged throughout.

Using surveys to study crime and criminal justice policy presents special challenges. The target population frequently includes lower-income, transient persons who are difficult to contact through customary sampling methods. For example, homeless persons are excluded from any survey that samples households, but people who live on the street no doubt figure prominently as victims and offenders. Maxfield (1999) describes how new data from the National Incident-Based Reporting System suggest that a number of “nonhousehold-associated” persons are systematically undercounted by sampling procedures used in the NCVS. Research by Census Bureau staff has tried to document how many individuals might be missed in household-based surveys, finding that young males and minorities appear to be most undercounted (Martin, 1999).

Crime surveys such as the NCVS and the BCS have been deficient in getting information about crimes of violence when the victim and offender have some prior relationship. This is particularly true for domestic violence, although Michael Rand and Callie Rennison report some advantages of large surveys in this regard (2005).

Underreporting of domestic violence appears to be due, in part, to the very general nature of large-scale crime surveys. Catriona Mirrlees-Black (1995:8) of the British Home Office summarizes the trade-offs of using survey techniques to learn about domestic violence:

Measuring domestic violence is difficult territory. The advantage of the BCS is that it is based on a large nationally representative sample, has a relatively high response rate, and collects information on enough incidents to provide reliable details of their nature. One disadvantage is that domestic violence is measured in the context of a crime survey, and some women may not see what happened to them as “crime,” or be reluctant to do so. Also, there is little time to approach the topic “gently.” A specially designed questionnaire with carefully selected interviewers may well have the edge here.

In recent years, both the NCVS and the BCS have been revised to produce better measures of domestic and intimate violence. Estimates of domestic violence increased in the 1996 wave of the BCS, and researchers think that the increase reflects a greater willingness by respondents to discuss domestic violence with interviewers (Mirrlees-Black, Mayhew, and Percy, 1996).

Most recently, the use of computer-assisted self-interviewing in the BCS has produced higher estimates of victimization prevalence among women as well as the first measurable rates of domestic violence victimization for males (Mirrlees-Black, 1999).

We have mentioned that national crime surveys cannot be used to estimate the frequency of victimization in local areas like individual cities. For this reason, the Census Bureau conducted a series of victim surveys in cities in the 1970s. City crime surveys suffer problems of their own, however. Because sampling procedures are almost always based on some sample frame of city residents, city-specific surveys are not able to measure crimes that involve non-resident victims. For example, an RDD survey of Key West, Florida, residents could not possibly count the robbery of a visitor who came
down from Cincinnati. This problem still exists in the resurrection of city surveys by the BJS and the COPS office (Smith, Steadman, Minton, and Townsend, 1999). City-level surveys are based on samples of residents and so cannot measure incidents that affect nonresident victims.

Survey research is generally weaker on validity and stronger on reliability. In comparison with field research, for instance, the artificiality of the survey format puts a strain on validity. As an illustration, most researchers agree that fear of crime is not well measured by the standard question, “How safe do you feel, or would you feel, out alone in your neighborhood at night?” Survey responses to that question are, at best, approximate indicators of what we have in mind when we conceptualize fear of crime.

Reliability is a different matter. By presenting all subjects with a standardized stimulus, survey research goes a long way toward eliminating unreliability in observations made by the researcher.

However, even this statement is subject to qualification. Critics of survey methods argue that questionnaires for standard crime surveys and many specialized studies embody a narrow, legalistic conception of crime that cannot reflect the perceptions and experiences of minorities and women. Survey questions typically are based on male views and do not adequately tap victimization or fear of crime among women (Straus, 1999; Tjaden and Thoennes, 2000). Concern that survey questions might mean different things to different respondents raises important questions about reliability and about the generalizability of survey results across subgroups of a population.

As with all methods of observation, a full awareness of the inherent or probable weaknesses of survey research may partially resolve them. Ultimately, though, we are on the safest ground when we can use several different research methods to study a given topic.

Other Ways of Asking Questions

*Specialized interviews and focus groups are alternative ways of gathering question-based data.*

Sample surveys are perhaps the best-known application of asking questions as a data-gathering strategy for criminal justice research. Often, however, more specialized interviewing techniques are appropriate.

Specialized Interviewing

No precise definition of the term *survey* enables us to distinguish a survey from other types of interview situations. As a rule of thumb, a sample survey (even one that uses nonprobability sampling methods) is an interview-based technique for generalizing to a larger population, using a standard questionnaire. In contrast, specialized interviewing focuses on the views and opinions of only those individuals who are interviewed, and frequently uses more flexible interview methods.

Let’s say we are interested in how mental health professionals view different drug treatment programs for prison inmates. One approach is to conduct a sample survey of psychologists who work in state correctional facilities in which each sampled psychologist completes a structured questionnaire concerning drug treatment programs. This approach will enable us to generalize to the population of state prison psychologists.

Another approach is to study one or two correctional institutions (or some small number) intensively. We might interview a psychologist in each institution and present questions about various approaches to drug treatment therapy. In all likelihood, we will not use a highly structured questionnaire, but rather use a list of questions or topics we wish to discuss with each subject. And we will treat the interview as more of a directed conversation than a formal interview. Of course, we cannot general-
ize from interviews with one or two prison psychologists to any larger population. However, we will gain an understanding (and probably a more detailed one) of how staff psychologists in specific institutions feel about different drug treatment programs.

Specialized interviewing asks questions of a small number of subjects, typically using an interview schedule that is much less structured than that in sample surveys. Michael Quinn Patton (2001) distinguishes two variations of specialized interviews. The less structured alternative is to prepare a general interview guide that includes the issues, topics, or questions the researcher wishes to cover. Issues and items are not presented to respondents in any standardized order. The interview guide is more like a checklist than an interview schedule, ensuring that planned topics are addressed at some point in the interview. The standardized open-ended interview, in contrast, is more structured, using specific questions arranged in a particular order. The researcher presents each respondent with the same questions in the same sequence (subject to any contingency questions). The questions are open ended, but their format and presentation are standardized.

To underscore the flexibility of specialized interviewing, Patton describes how the two approaches can be used in combination (2001:347):

A conversational strategy can be used within an interview guide approach, or you can combine a guide approach. . . . This combined strategy offers the interviewer flexibility in probing and in determining when it is appropriate to explore certain subjects in greater depth, or even to pose questions about new areas of inquiry that were not originally anticipated in the interview instrument’s development.

Open-ended questions are ordinarily used because they capture rich detail better. The primary disadvantage of open-ended questions—having to categorize responses—is not a problem in specialized interviewing because of the small numbers of subjects and because researchers are more interested in describing than in generalizing.

Specialized interviewing can be incorporated into any research project as a supplementary source of information. If, for example, we are interested in the effects of determinate sentencing on prison populations, we can analyze data from the Census of State Adult Correctional Facilities, conducted by the BJS. We might also interview a small number of corrections administrators, perhaps asking them to react to our data analysis. Evaluation studies and other applied research projects frequently use specialized interviewing techniques, alone or in combination with other sources of data. For example, Nanci Plouffe and Rana Sampson (2004) interviewed auto thieves to learn why they tended to operate in certain kinds of shopping centers near the California-Mexico border. More generally, Scott Decker (2005) describes how specialized interviews of offenders can enhance applied research in crime prevention research.

Focus Groups

Like sample surveys, focus group techniques were refined by market research firms in the years following World War II. As the name implies, market research explores questions about the potential for sales of consumer products. Because a firm may spend millions of dollars developing, advertising, and distributing some new item, market research is an important tool to test consumer reactions before large sums of money are invested in a product.

Surveys have two disadvantages in market research. First, a nationwide or large-scale probability survey can be expensive. Second, it may be difficult to present advertising messages or other product images in a survey format. Focus groups have proved to be more suitable for many market research applications. In
recent years, focus groups have commonly been used as substitutes for surveys in criminal justice and other social science research.

In a focus group, 8–15 people are brought together in a room to engage in a guided group discussion of some topic. Although focus groups cannot be used to make statistical estimates about a population, members are nevertheless selected to represent a target population. Richard Krueger and Mary Anne Casey (2000) and D. W. Stewart and P. N. Shamdasani (1990) describe focus groups, their applications, and their advantages and disadvantages in detail.

For example, the location of community correctional facilities such as work-release centers and halfway houses often prompts a classic “Not in my backyard!” (NIMBY) response from people who live in neighborhoods where proposed facilities will be built. Recognizing this, a mayor who wants to find a suitable site without annoying neighborhood residents (voters) is well advised to convene a focus group that includes people who live in areas near possible facility locations. A focus group can test the “market acceptability” of a work-release center, which might include the best way to package and sell the product. Such an exercise might reveal that an appeal to altruism (“We all have to make sacrifices in the fight against crime”) is much less effective in gaining support than an alternative sales pitch that stresses potential economic benefits (“This new facility will provide jobs for neighborhood residents”).

Generalizations from focus groups to target populations cannot be precise; however, a study by V. M. Ward and associates (Ward, Bertrand, and Brown, 1991) found that focus group and survey results can be quite consistent under certain conditions. They conclude that focus groups are most useful in two cases: (1) when precise generalization to a larger population is not necessary and (2) when focus group participants and the larger population they are intended to represent are relatively homogeneous. So, for example, a focus group is not appropriate to predict how all city residents will react to a ban on handgun ownership. But a focus group of registered handgun owners could help evaluate a proposed city campaign to buy back handguns.

Focus groups may also be used in combination with survey research in one of two ways. First, a focus group can be valuable in questionnaire development. When researchers are uncertain how to present items to respondents, a focus group discussion about the topic can generate possible item formats. For instance, James Nolan and Yoshio Akiyama (1999) studied police routines for making records of hate crimes. In a general sense, they knew what concepts they wanted to measure but were unsure how to operationalize them. They convened five focus groups in different cities, groups that included police administrators, mid-level managers, patrol officers, and civilian employees, to learn about different perspectives on hate crime recording. Analyzing focus group results, Nolan and Akiyama prepared a self-administered questionnaire that was sent to a large number of individuals in four police departments.

Second, after a survey has been completed and preliminary results tabulated, focus groups may be used to guide the interpretation of some results. After a citywide survey in which we find, for example, that recent immigrants from Southeast Asian countries are least supportive of community policing, we might conduct a focus group of Asian residents to delve more deeply into their concerns.

Focus groups are flexible and can be adapted to many uses in basic and applied research. Keep in mind, however, two key elements expressed in the name of this data collection technique. Focus means that researchers present specific questions or issues for directed discussion. Having a free-for-all discussion about hate crime, for example, would not have yielded much useful insight for Nolan and Akiyama to develop a survey instrument. Group calls our attention to potential participants in the focused discussions. Like market researchers, we should select participants from a specific target
population that relates to our research questions. If we’re interested in how residents of a specific neighborhood will feel about opening a work-release center, we should select group participants who live in the target neighborhood or in one very much like it.

**Should You Do It Yourself?**

*Anyone can do a mail or simple telephone survey, but many times it’s better to use professional survey researchers.*

The final issue we address in this chapter is who should conduct surveys. Drawing a sample, constructing a questionnaire, and either conducting interviews or distributing self-administered instruments are not especially difficult. Equipped with the basic principles we have discussed so far in this book, you could complete a modest in-person or telephone survey yourself. Mail surveys of large numbers of subjects are entirely possible, especially with present-day computer capabilities.

At the same time, the different tasks involved in completing a survey require a lot of work and attention to detail. We have presented many tips for constructing questionnaires, but our guidelines barely scratch the surface. Many books describe survey techniques in more detail, and a growing number focus specifically on telephone or mail techniques. (See the list of Additional Readings at the end of this chapter.) In many respects, however, designing and executing a survey of even modest size can be challenging.

Consider the start-up costs involved in person or telephone interview surveys of any size. Finding, training, and paying interviewers are time consuming and potentially costly and require some degree of expertise. The price of computer equipment continues downward, but a CATI setup or supply of laptops and associated software for interviewers still represents a substantial investment that cannot easily be justified for a single survey.

If interview surveys are beyond a researcher’s means, he or she might fall back on a mail or web-based survey. Few capital costs are involved; most expenses are in consumables such as envelopes, stamps, and stationery. One or two persons can orchestrate a mail survey reasonably well at minimal expense. Perhaps a consultant could be hired to design a web-based survey at modest cost. But consider two issues.

First, the business of completing a survey involves a great deal of tedious work. In mail surveys, for example, questionnaires and cover letters must be printed, folded, or stuffed into envelopes, stamped, and delivered (finally!) to the post office. None of this is much fun. The enjoyment starts when completed questionnaires start to trickle in. It’s rewarding to begin the actual empirical research, and the excitement may get the researcher past the next stretch of tedium: going from paper questionnaires to actual data. So it’s possible for one individual to do a mail survey, but the researcher must be prepared for lots of work; even then, it will be more work than expected. Web-based surveys have their own trade-offs. Economizing on up-front programming costs entails the risk of being swamped by unusable electronic questionnaires.

The second issue is more difficult to deal with and is often overlooked by researchers. We have examined at some length the advantages and disadvantages of the three methods of questionnaire administration. Some methods are more or less appropriate than others are for different kinds of research questions. If a telephone or an in-person interview survey is best for the particular research needs, conducting a mail or web-based survey will be a compromise, perhaps an unacceptable one. But the researcher’s excitement at actually beginning the research may lead him or her to overlook or minimize problems with doing a mail survey on the cheap, in much the same way that researchers often are not in a position to recognize ethical problems with their own work,
PUTTING IT ALL TOGETHER

ASKING QUESTIONS

Research on racial profiling has used surveys and other ways of collecting data by asking questions. One nationally representative survey has been conducted several times. We will say a bit more about the driver survey conducted in New Jersey. Researchers in North Carolina conducted several surveys and focus groups to assess how drivers and the police view each other.

Police-Public Contact Survey

The Police-Public Contact Survey (PPCS) has its roots in comprehensive anticrime legislation passed in 1994. The law required the Bureau of Justice Statistics (BJS) to collect data on the use of excessive force by police. From the outset it was decided that this would be best accomplished by adding questions to the National Crime Victimization Survey (NCVS). Almost as an afterthought, Jan Chaiken, then director of BJS, decided to include pilot questions asking specifically about traffic stops. As a result, the PPCS became one of the first sources of survey-based information about traffic stops.

Though limited in some respects, using the NCVS as a research platform in this way has several strengths. First, the NCVS has a large sample size, so it produces large numbers of people with some police contact. In the 2002 survey, about 15,700 of 77,000 respondents reported a police contact. Second, not all traffic stops produce a citation, so a survey sample can reveal something about the “dark figure” of unrecorded traffic stops. Third, the race of respondents is recorded, so it’s possible to compare experiences of minority drivers to other drivers. Finally, considerable detail about the nature of the traffic stop can be recorded, including police demeanor, and whether or not respondents believed they were treated fairly. All of this is from the perspective of drivers, offering a very different source of information than is obtained from police records of traffic stops.

Tollbooth Survey of Drivers

In earlier chapters we mentioned the brief survey of drivers on the New Jersey Turnpike. The mechanics of the survey are especially interesting. Interviewers were stationed in tollbooths, at a time when all vehicles had to stop and pay cash tolls when exiting the turnpike. Cars were sampled as we described in Chapter 8. Here is how Lange, Johnson, and Voas describe the interview procedure:

Drivers of selected vehicles would hand the toll collector their Turnpike ticket, and then the surveyor would lean out the booth window and offer the driver $5 to answer a few brief questions. The driver was assured that the survey was voluntary and confidential. A $5 bill was placed inside an envelope that had a contact telephone number on it so drivers could learn more about the survey if they wished. (2005:199)

Drivers were asked where they entered the turnpike, making it possible to associate respondents with specific turnpike segments. They were then asked their age, and racial/ethnic identification from the categories white, black, Hispanic/Latino, Asian, American Indian, or other; follow-up questions further classified those who responded Hispanic/Latino or Asian. Gender was recorded by observation. Interviewers also recorded vehicle license plate numbers and state

Results from the 2002 survey are summarized by Durose, Schmitt, and Langan (2004). Note, that these are national estimates. As such they offer useful information about the range of experience throughout the country. But because people are policed by thousands of organizations, any findings about race disparities could not be traced to individual organizations. National findings about police–public contacts have about as much validity as national findings about victimization. They illustrate the broadest patterns of experience, but do not accurately portray patterns of practice for individual agencies.
of registration. This was a hand-held-computer-assisted interview. Questionnaires were programmed into a PDA, and responses tapped in directly. This technique ensured survey results could be directly uploaded for analysis.

**North Carolina Surveys and Focus Groups**

Like the PPCS, a statewide telephone survey of North Carolina drivers yielded information on driver behavior, the circumstances surrounding traffic stops, and the character of interactions with police. Unlike the PPCS, it was possible to compare driver experiences across law enforcement organizations. In particular, Smith and associates report that drivers had generally more favorable attitudes about experience with North Carolina State Highway Patrol compared with police from local agencies. In Chapter 8 we described how a sample of known speeders was included in the survey. This made it possible to test the criterion validity of self-report items. Because the survey sought information about people’s experience in traffic stops, its target population was North Carolina drivers. Therefore a sampling frame was readily available from the state agency that kept driver records. And because telephone numbers were included in these records, a telephone survey was the method of choice.

Multiple focus groups were conducted with North Carolina State Highway Patrol officers and with drivers. Smith and colleagues structured some focus groups so that all participants were African American or all were white. This was to promote open, candid discussion about sensitive topics. Researchers cite focus group literature that advises making respondents feel as comfortable as possible in support of this policy. Six police focus groups of six to nine troopers each were conducted. Six citizen focus groups of 10 people each were convened in large cities throughout North Carolina. Participants in the citizens’ groups were selected by a research firm to represent ages between 24 and 60, and included at least four women.

Participants in focus groups of African American citizens felt police generally did a good job, but many also had some personal experience of perceived discrimination. Impolite or demeaning treatment by police was interpreted as signs of bias or discrimination. Drivers had sophisticated views about what sorts of stops were linked to profiling; traffic stops for minor offenses, like rolling through a stop sign, were especially suspect. An excerpt from one of the groups of African American citizens illustrates the kinds of insights that are difficult to get with large-scale surveys:

“I see him (the officer) just looking around my car and what not, and he said, ‘You know the reason why I pulled you over, don’t you?’ He said, ‘Your tag has September ’98 on it.’ I said, ‘It’s only May.’ He was like, ‘That’s right. My fault.’ He gave me my driver’s license and registration back and said, ‘You don’t have any guns in the car do you?’ I said, ‘No, it’s at home with the dope.’” (2003:359)

Focus groups with troopers also produced candid observations. Most were concerned about increased attention to the racial distribution of people they stopped, feeling that they had to “balance the books” (2003:289). You may recognize how that would increase the role race plays in decisions to stop. These focus group interviews were held in June 2001. The views that two troopers expressed about profiling sound familiar in the post-9/11 world:

I just want to say that we are all trained and I do not care what law enforcement agency you are in, you are trained to profile and you act on that. It is inherent. This business requires that.

I think stereotyping and violations go together. If you look at a certain group, stereotype a certain group and in that group you deem it to be, have more violations, then naturally you will work more at stopping whatever . . . (2003:287)

Surveys are best at getting a large number of responses to mostly closed-ended questions. Focus groups are valuable sources of richer detail that can explain what people do and how they feel. Surveys are better at tabulating experience with traffic stops. Focus groups are better at explaining how people feel about their experiences.
as we saw in Chapter 3. Doing a mail survey because it’s all you can afford does not necessarily make the mail survey worth doing.

The alternative to doing it yourself is to contract with a professional survey research firm or a company that routinely conducts surveys. Most universities have a survey research center or institute, often affiliated with a sociology or political science department. For example, the Eagleton Center for Public Interest Polling at Rutgers University conducts a variety of research and other surveys (http://eagletonpoll.rutgers.edu/index.php). Such institutes are usually available to conduct surveys for government organizations as well as university researchers, and they can often do so very economically. Private research firms are another possibility. Most have the capability to conduct all types of surveys as well as focus groups.

Using a professional survey firm or institute has several advantages. Chapter 8 described the basic principles of sampling, but actually drawing a probability sample can be complex. Even the BJS/COPS do-it-yourself guidebook for community crime surveys (Weisel, 1999) counsels police departments and others to consult with experts in drawing RDD samples. Professional firms regularly use sampling frames that can represent city, state, and national samples, or whatever combination is appropriate.

We have emphasized the importance of measurement throughout this book. Researchers should develop conceptual and operational definitions and be attentive to all phases of the measurement process. However, constructing an actual questionnaire requires attention to details that may not always be obvious to researchers. Survey firms are experienced in preparing standard demographic items, batteries of matrix questions, and complex contingency questions with appropriate skip sequences.

Although it is often best for researchers to discuss specific concepts and even to draft questions, professional firms offer the considerable benefit of experience in pulling it all together. This is not to say that a researcher should simply propose some ideas for questions and then leave the details to the pros. Working together with a survey institute or market research firm to propose questionnaire items, review draft instruments, evaluate pretests, and make final modifications is usually the best approach.

Perhaps the chief benefit of contracting for a survey is that professional firms have a pool of trained interviewers or the equipment to conduct computer-assisted telephone interviews. Survey research centers and other professional organizations have the latest specialized equipment, software, and know-how to take advantage of advances in all forms of CAI. Furthermore, such companies can more readily handle such administrative details as training interviewers, arranging travel for in-person surveys, coordinating mail surveys, and providing general supervision. This frees researchers from much of the tedium of survey research, enabling them to focus on more substantive issues.

Researchers must ultimately decide whether to conduct a survey themselves or to contract with a professional firm. And the decision is best made after carefully considering the pros and cons of each approach. Too often, university faculty assume that students can get the job done while overlooking the important issues of how to maintain quality control and whether a survey is a worthwhile investment of students’ time. Similarly, criminal justice practitioners may believe that agency staff can handle a mail survey or conduct phone interviews from the office. Again, compromises in the quality of results, together with the opportunity costs of diverting staff from other tasks, must be considered. The do-it-yourself strategy may seem cheaper in the short run, but it often becomes a false economy when attention turns to data analysis and interpretation.

We’ll close this section with an apocryphal story about a consultant’s business card; the card reads: “Fast! Low cost! High quality! Pick
any two.” It’s best to make an informed choice that best suits your needs.

✪ Main Points

- Survey research, a popular social research method, involves the administration of questionnaires in a systematic way to a sample of respondents selected from some population.
- Survey research is especially appropriate for descriptive or exploratory studies of large populations, but surveys have many other uses in criminal justice research.
- Surveys are the method of choice for obtaining self-reported offending data. Continuing efforts to improve self-report surveys include using confidential computer-assisted personal interviews.
- Questions may be open ended or closed ended. Each technique for formulating questions has advantages and disadvantages.
- Short items in a questionnaire are usually better than long ones.
- Bias in questionnaire items encourages respondents to answer in a particular way or to support a particular point of view. It should be avoided.
- Questionnaires may be administered in three basically different ways: self-administered questionnaires, face-to-face interviews, and telephone interviews. Each mode of administration can be varied in a number of ways.
- Computers can be used to enhance each type of survey. Computer-assisted surveys have many advantages, but they often require special skills and equipment.
- It is generally advisable to plan follow-up mailings for self-administered questionnaires, sending new questionnaires to respondents who fail to respond to the initial appeal.
- The essential characteristic of interviewers is that they be neutral; their presence in the data collection process must not have any effect on the responses given to questionnaire items.
- Surveys conducted over the telephone are fast and flexible.
- Each method of survey administration has a variety of advantages and disadvantages.
- Survey research has the weaknesses of being somewhat artificial and potentially superficial. It is difficult to gain a full sense of social processes in their natural settings through the use of surveys.
- Specialized interviews with a small number of people and focus groups are additional ways of collecting data by asking questions.
- Although the particular tasks required to complete a survey are not especially difficult, researchers must carefully consider whether to conduct surveys themselves or contract with a professional organization.

✪ Key Terms

- Closed-ended questions
- Open-ended questions
- Computer-assisted interviewing
- Questionnaire

✪ Review Questions and Exercises

1. Find a questionnaire on the Internet. Bring the questionnaire to class and critique it. Critique other aspects of the survey design as well.
2. For each of the open-ended questions listed, construct a closed-ended question that could be used in a questionnaire.
   a. What was your family’s total income last year?
   b. How do you feel about shock incarceration, or “boot camp” programs?
   c. How do people in your neighborhood feel about the police?
   d. What do you feel is the biggest problem facing this community?
   e. How do you protect your home from burglary?
3. Prepare a brief questionnaire to study perceptions of crime near your college or university. Include questions asking respondents to describe a nearby area where they either are afraid to go after dark or think crime is a problem. Then use your questionnaire to interview at least 10 students.
4. A recent evaluation of a federal program to support community policing included sending questionnaires to a sample of about 1200 police chiefs. Each questionnaire included a number of items asking about specific features of community policing and whether they were being used in the department. Almost all the police chiefs had someone else complete the questionnaire. What’s the unit of analysis in this survey?
What problems might result from having an individual complete such a questionnaire?

Online Sessions for Chapter 9

Online Sessions present further examples of selected material covered in each chapter. Your instructor may choose to use these for in-class discussion (as I do with my students) or you can access them at home. To ensure that you always have access to live, correct links for the websites described here, we provide the necessary links on the companion website for Research Methods for Criminal Justice and Criminology at http://www.thomsonedu.com/criminaljustice/maxfield. Once at the companion website, select this specific chapter, click on “Chapter Resources,” then click on “Web Links.”

More on Web-Based Surveys

Web Survey Methodology (WebSM) presents information on a variety of topics in connection with conducting web-based surveys. A series of links to available software is presented. Though the site is maintained by a loose coalition of organizations in Europe, it offers much of interest to researchers generally. http://www.websm.org/

Free Online Survey Tools

The web now includes hundreds of firms that will conduct web-based surveys, along with services that enable you to prepare, distribute, and compile results yourself. Though it may seem to counter our advice to use professionals whenever possible, you can learn quite a lot about the mechanics of web surveys by doing one. Fortunately, a number of sites offer varying levels of access ranging from free surveys of very modest scope to more costly subscriptions for large surveys. One example is SurveyMonkey.com, where free subscriptions allow users to prepare surveys of 10 questions each, and receive up to 100 responses. Registration is simple. Questionnaire design templates are available. Or you can start from scratch. Try it, or link to another site that offers free surveys of limited scope. http:// surveymonkey.com/

Questionnaires from BJS

Most BJS publications on the NCVS or other surveys include copies of questionnaires. In addition, the BJS website includes links to questionnaires used in all major surveys it sponsors. The second URL links to materials about the community victim surveys sponsored by the COPS office and BJS. You can even download software for customizing the questionnaire. That would be a useful in-class exercise. Finally, go to the third URL for a link to the NCVS interviewing manual for field representatives. Almost 900 pages in length, this will give you a feel for what it’s like to conduct interviews for a large national survey.

http://www.ojp.usdoj.gov/bjs/quest.htm
http://www.ojp.usdoj.gov/bjs/abstract/cvs.htm
http://www.ojp.usdoj.gov/bjs/cvict.htm

Interviewing Offenders

Among the resources available at the Center for Problem-Oriented Policing (POP Center) is Scott Decker’s guide to interviewing offenders. The focus is on applied research, such as how learning more about offender techniques can help develop crime prevention strategies. See the publication, “Theft of and from Autos in Parking Facilities in Chula Vista, California,” for an example of a questionnaire used to interview offenders. Both can be found on the POP Center website.

http://popcenter.org/Tools/toolinterviewingOffenders.htm
http://popcenter.org/Library/researcherprojects/TheftCars.pdf

Question Bank and Survey Guidelines

The Social Research Update website has a very useful article on general resources for survey research available online. These are grouped into four categories: questionnaires available and the Centre for Applied Social Surveys, organizations that conduct surveys, data providers, and general resources.

http://www.soc.surrey.ac.uk/sru/SRU41.html

Additional Readings


Dillman, Don A., Mail and Internet Surveys: The Tailored Design Method, 2nd ed. (New York: Wiley, 1999). This update of a classic reference on self-administered surveys includes a variety of web-based techniques. Dillman makes many good suggestions for improving response rates.

General Accounting Office, Using Structured Interviewing Techniques (Washington, DC: General
Accounting Office, 1991). This is another useful handbook in the GAO series on evaluation methods. In contrast to Patton (below), the GAO emphasizes getting comparable information from respondents through structured interviews. This is a very useful step-by-step guide.


Patton, Michael Quinn, Qualitative Research and Evaluation Methods, 3rd ed. (Thousand Oaks, CA: Sage, 2001). This is a thorough discussion of specialized interviewing. Patton’s advice will also be useful in constructing questionnaires for surveys in general.

Weisel, Deborah, Conducting Community Surveys: A Practical Guide for Law Enforcement Agencies (Washington, DC: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, and Office of Community Oriented Police Services, 1999). Another practical guide, this brief publication is prepared for use by public officials, not researchers. As such, it’s a very good description of the nuts and bolts of doing telephone surveys.
Field Research

The techniques described in this chapter focus on observing life in its natural habitat—going where the action is and watching. We’ll consider how to prepare for the field, how to observe, how to make records of what is observed, and how to recognize the relative strengths and weaknesses of field research.

Introduction 283
Topics Appropriate to Field Research 284
The Various Roles of the Observer 286
Asking Questions 288
Gaining Access to Subjects 290
Gaining Access to Formal Organizations 290
Gaining Access to Subcultures 293
Selecting Cases for Observation 294
Purposive Sampling in Field Research 295
Recording Observations 297
Cameras and Sound Recorders 298
Field Notes 299
Structured Observations 300
Linking Field Observations and Other Data 302
Conducting a Safety Audit 304
Illustrations of Field Research 304
Shoplifting 305
How Many People Wear Seat Belts? 307
Bars and Violence 309
Strengths and Weaknesses of Field Research 310
Validity 311
(continued)
Introduction

Field research is often associated with qualitative techniques, though many other applications are possible.

We turn now to what may seem like the most obvious method of making observations: field research. If researchers want to know about something, why not simply go where it’s happening and watch it happen?

Field research encompasses two different methods of obtaining data: (1) making direct observation and (2) asking questions. This chapter concentrates primarily on observation, although we will briefly describe techniques for specialized interviewing in field studies.

Most of the observation methods discussed in this book are designed to produce data appropriate for quantitative (statistical) analysis. Surveys provide data to calculate things like the percentage of crime victims in a population or the mean value of property lost in burglaries. Field research may yield qualitative data—observations not easily reduced to numbers—in addition to quantitative data. For example, a field researcher who is studying burglars may note how many times subjects have been arrested (quantitative) as well as whether individual burglars tend to select certain types of targets (qualitative).

Qualitative field research is often a theory- or hypothesis-generating activity as well. In many types of field studies, researchers do not have precisely defined hypotheses to be tested. Field observation may be used to make sense out of an ongoing process that cannot be predicted in advance. This process involves making initial observations, developing tentative general conclusions that suggest further observations, making those observations, revising the prior conclusions, and so forth.

For example, Ross Homel and associates (Homel, Tomsen, and Thommeny, 1992) conducted a field study of violence in bars in Sydney, Australia, and found that certain situations tended to trigger violent incidents. Subsequent studies tested a series of hypotheses about the links between certain situations and violence (Homel and Clark, 1994), and how interior design was related to aggression in dance clubs (Macintyre and Homel, 1996). Later research by James Roberts (2002) expanded these findings by examining management and serving practices New Jersey bars and clubs. Barney Glaser and Anselm Strauss (1967) referred to this process as “grounded theory.” Rather than following the deductive approach to theory building described in Chapter 2, grounded theory is based on (or grounded in) experience, usually through observations made in the field.

Field studies in criminal justice may also produce quantitative data that can be used to test hypotheses or evaluate policy innovations. Typically, qualitative exploratory observations help define the nature of some crime problem and suggest possible policy responses. Following the policy response, further observations are made to assess the policy’s impact. For example, we briefly described the situational crime prevention approach in Chapter 2. The first and last of the five steps in a situational crime prevention project illustrate the dual
uses of observation for problem definition and hypothesis testing (Clarke, 1997:5):

1. Collect data about the nature and dimensions of the specific crime problem.

5. Monitor results and disseminate experience.

By now, especially if you have experience as a criminal justice professional, you may be thinking that field research is not much different from what police officers and many other people do every day—make observations in the field and ask people questions. Police may also collect data about particular crime problems, take action, and monitor results. So what’s new here?

Compared with criminal justice professionals, researchers tend to be more concerned with making generalizations and then using systematic field research techniques to support those generalizations. For example, consider the different goals and approaches used by two people who might observe shoplifters: a retail store security guard and a criminal justice researcher. The security guard wishes to capture a thief and prevent the loss of shop merchandise. Toward those ends, he or she adapts surveillance techniques to the behavior of a particular suspected shoplifter. The researcher’s interests are different; perhaps she or he estimates the frequency of shoplifting, describes characteristics of shoplifters, or evaluates some specific measure to prevent shoplifting. In all likelihood, researchers use more standardized methods of observation aimed toward a generalized understanding.

In a sense, we all do field research whenever we observe or participate in social behavior and try to understand it, whether at a corner tavern, in a doctor’s waiting room, or on an airplane. When we report our observations to others, we are reporting our field research efforts.

This chapter examines field research methods in some detail, providing a logical overview and suggesting some specific skills and techniques that make scientific field research more useful than the casual observation we all engage in. As we cover the various applications and techniques of field research, it’s useful to recall the distinction we made, way back in Chapter 1, between ordinary human inquiry and social scientific research. Field methods illustrate how the common techniques of observation that we all use in ordinary inquiry can be deployed in systematic ways.

**Topics Appropriate to Field Research**

*When conditions or behavior must be studied in natural settings, field research is usually the best approach.*

One of the key strengths of field research is the comprehensive perspective it gives the researcher. This aspect of field research enhances its validity. By going directly to the phenomenon under study and observing it as completely as possible, we can develop a deeper and fuller understanding of it. This mode of observation, then, is especially (though not exclusively) appropriate to research topics that appear to defy simple quantification. The field researcher may recognize nuances of attitude, behavior, and setting that escape researchers using other methods.

For example, Clifford Shearing and Philip Stenning (1992:251) describe how Disney World employs subtle but pervasive mechanisms of informal social control that are largely invisible to millions of theme park visitors. It is difficult to imagine any technique other than direct observation that could produce these insights:

Control strategies are embedded in both environmental features and structural relations. In both cases control structures and activities have other functions which are highlighted so that the control function is overshadowed. For example, virtually every pool, fountain, and flower garden serves both as an aesthetic object and to direct visitors away from, or to-
wards, particular locations. Similarly, every Disney employee, while visibly and primarily engaged in other functions, is also engaged in the maintenance of order.

Many of the different uses of field observation in criminal justice research are nicely summarized by George McCall. Comparing the three principal ways of collecting data—observing, asking questions, and consulting written records—McCall (1978:8–9) stated that observation is most appropriate for obtaining information about physical or social settings, behaviors, and events.

Observation is not the only way of getting data about settings, behaviors, and events. Information about the number of households on a block (setting) may be found in a city clerk’s office; crime surveys routinely ask about victimization (events) and whether crimes were reported to police (behavior). However, in many circumstances, field observation is the preferred method.

Field research is especially appropriate for topics that can best be understood within their natural settings. Surveys may be able to measure behaviors and attitudes in somewhat artificial settings, but not all behavior is best measured this way. For example, field research is a superior method for studying how street-level drug dealers interpret behavioral and situational cues to distinguish potential customers, normal street traffic, and undercover police officers. It would be difficult to study these skills through a survey.

Field research on actual crimes involves obtaining information about events. McCall (1978) pointed out that observational studies of vice—such as prostitution and drug use—are much more common than observational studies of other crimes, largely because these behaviors depend at least in part on being visible and attracting customers. One notable exception is research on shoplifting. A classic study by Terry Baumer and Dennis Rosenbaum (1982) had two goals: (1) to estimate the incidence of shoplifting in a large department store and (2) to assess the effectiveness of different store security measures. Each objective required devising some measure of shoplifting, which Baumer and Rosenbaum obtained through direct observation. Samples of persons were followed by research staff from the time they entered the store until they left. Observers, posing as fellow shoppers, watched for any theft by the person they had been assigned to follow. We’ll have more to say about this study later in the chapter.

Many aspects of physical settings are probably best studied through direct observation. The prevalence and patterns of gang graffiti in public places could not be reliably measured through surveys, unless the goal was to measure perceptions of graffiti. The work of Oscar Newman (1972, 1996), Ray Jeffery (1977), and Patricia and Paul Brantingham (1991) on the relationship between crime and environmental design depends crucially on field observation of settings. If opportunities for crime vary by physical setting, then observation of the physical characteristics of a setting is required.

An evaluation of street lighting as a crime prevention tool in two areas of London illustrates how observation can be used to measure both physical settings and behavior. Kate Painter (1991, 1996) was interested in the relationships between street lighting, certain crime rates (measured by victim surveys), fear of crime, and nighttime mobility. Improvements in street lighting were made in selected streets; surveys of pedestrians and households in the affected areas were conducted before and after the lighting improvements. Survey questions included items about victimization, perceptions of crime problems and lighting quality, and reports about routine nighttime behavior. Although the pretest and posttest survey items could have been used to assess changes in attitudes and behavior associated with improved lighting, field observations provided better measures of behavior. Painter conducted systematic counts of pedestrians in areas both before and after street lighting was enhanced. Observations like this are better measures of
such behavior than are survey items because people often have difficulty recalling routine actions such as how often they walk through some area after dark.

Painter’s research also included provisions to observe physical settings. First, light levels (measured in lux units) were assessed in the experimental areas before and after changes in streetlighting equipment (1991:176). Second, interviewers who conducted household interviews made observations of physical settings in the areas around sampled households.

The Various Roles of the Observer

Field observer roles range from full participation to fully detached observation.

The term field research is broader and more inclusive than the common term participant observation. Field researchers need not always participate in what they are studying, although they usually will study it directly at the scene of the action. As Catherine Marshall and Gretchen Rossman (1995:60) point out:

The researcher may plan a role that entails varying degrees of “participantness”—that is, the degree of actual participation in daily life. At one extreme is the full participant, who goes about ordinary life in a role or set of roles constructed in the setting. At the other extreme is the complete observer, who engages not at all in social interaction and may even shun involvement in the world being studied. And, of course, all possible complementary mixes along the continuum are available to the researcher.

The full participant, in this sense, may be a genuine participant in what he or she is studying (for example, a participant in a demonstration against capital punishment)—or at least pretend to be a genuine participant. In any event, if you are acting as a full participant, you let people see you only as a participant, not as a researcher.

That raises an ethical question: Is it ethical to deceive the people we are studying in the hope that they will confide in us as they would not confide in an identified researcher? Do the interests of science—the scientific values of the research—offset any ethical concerns?

Related to this ethical consideration is a scientific one. No researcher deceives his or her subjects solely for the purpose of deception. Rather, it is done in the belief that the data will be more valid and reliable, that the subjects will be more natural and honest if they do not know the researcher is doing a research project. If the people being studied know they are being studied, they might reject the researcher or modify their speech and behavior to appear more respectable than they otherwise would. In either case, the process being observed might radically change.

On the other side of the coin, if we assume the role of complete participant, we may affect what we are studying. To play the role of participant, we must participate, yet our participation may affect the social process we are studying. Additional problems may emerge in any participant observation study of active criminals. Legal and physical risks, mentioned in Chapter 3, present obstacles to the complete participant in field research among active offenders or delinquents.

Finally, complete participation in field studies of criminal justice institutions is seldom possible. Although it is common for police officers to become criminal justice researchers, practical constraints on the official duties of police present major obstacles to simultaneously acting as researcher and police officer. Similarly, the responsibilities of judges, prosecutors, probation officers, and corrections workers are not normally compatible with collecting data for research.

Because of these considerations—ethical, scientific, practical, and safety—field researchers most often choose a different role. The
role participant-as-observer participates with the group under study but makes it clear that he or she is also undertaking research. If someone has been convicted of some offense and been placed on probation, for example, that might present an opportunity to launch a study of probation officers.

McCall (1978) suggested that field researchers who study active offenders may comfortably occupy positions around the periphery of criminal activity. Acting as a participant in certain types of leisure activities, such as frequenting selected bars or dance clubs, may be appropriate roles. This approach was used by Dina Perrone (2006) in her research on drug use in New York dance clubs (see the box in Chapter 3). Furthermore, McCall described how making one’s role as a researcher known to criminals and becoming known as a “right square” is more acceptable to subjects than an unsuccessful attempt to masquerade as a colleague. There are dangers in this role also, however. The people being studied may shift their attention to the research project, and the process being observed may no longer be typical. Conversely, a researcher may come to identify too much with the interests and viewpoints of the participants. This is referred to as “going native” and results in a loss of detachment necessary for social science.

The observer-as-participant identifies him or herself as a researcher and interacts with the participants in the course of their routine activities but makes no pretense of actually being a participant. Many observational studies of police patrol are examples of this approach. Researchers typically accompany police officers on patrol, observing routine activities and interactions between police and citizens. Spending several hours in the company of a police officer also affords opportunities for unstructured interviewing.

Although the researcher’s role in the eyes of police is exclusively observational, people with whom police come into contact may assume that an observer is actually a plainclothes officer. As a consequence, observation of citizens in their encounters with police tends to be less contaminated by the presence of a researcher than is observation of police themselves.

McCall (1978:88) noted that going native is a common tendency among field workers, especially in observational studies of police. For example, observers may become more sympathetic toward officers’ behavior and toward the views police express about the people they encounter. Observers may even actively assist officers on patrol. In either case, going native is often the product of the researcher’s efforts to be accepted by police and the natural tendency of police to justify their actions to observers.

The complete observer, at the other extreme, observes some location or process without becoming a part of it in any way. The subjects of study might not even realize they are being studied because of the researcher’s unobtrusiveness. An individual making observations while sitting in a courtroom is an example. Although the complete observer is less likely to affect what is being studied and less likely to go native than the complete participant, he or she may also be less able to develop a full appreciation of what is being studied. A courtroom observer, for example, witnesses only the public acts that take place in the courtroom, not private conferences between judges and attorneys.

McCall (1978:45) pointed out an interesting and often unnoticed trade-off between the role observers adopt and their ability to learn from what they see. If their role is covert (complete participation) or detached (complete observation), they are less able to ask questions to clarify what they observe. As complete participants, they take pains to conceal their observations and must exercise care in querying subjects. Similarly, complete observation means that it is generally not possible to interact with the persons or things being observed.

Researchers have to think carefully about the trade-off. If it is most important that subjects not be affected by their role as observer,
then complete participation or observation is preferred. If being able to ask questions about what they observe is important, then some role that combines participation and observation is better.

More generally, the appropriate role for observers hinges on what they want to learn and how their inquiry is affected by opportunities and constraints. Different situations require different roles for researchers. Unfortunately, there are no clear guidelines for making this choice; field researchers rely on their understanding of the situation, their judgment, and their experience. In making a decision, researchers must be guided by both methodological and ethical considerations. Because these often conflict, deciding on the appropriate role may be difficult. Often, researchers find that their role limits the scope of their study.

**Asking Questions**

*Field researchers frequently supplement observations by interviewing subjects.*

Field research might involve going where the action is and simply watching and listening. Researchers can learn a lot merely by being attentive to what's going on. Field research can also involve more active inquiry. Sometimes, it’s appropriate to ask people questions and record their answers. Richard Wright and Trevor Bennett (1990) describe observing and interviewing as two complementary ways to study offenders.

We examined interviewing in Chapter 9 on survey research. The interviewing normally done in connection with field observation falls into the category of specialized interviewing.

Field research interviews are usually much less structured than survey interviews. At one extreme, an unstructured interview is essentially a conversation in which the interviewer establishes a general direction for the conversation and pursues specific topics raised by the respondent. Ideally, the respondent does most of the talking. Michael Patton (2001) refers to this type as an “informal conversational interview,” which is especially well suited to in-depth probing.

Unstructured interviews are most appropriate when researchers have little knowledge about a topic and when it’s reasonable for them to have a casual conversation with a subject. This is a good strategy for interviewing active criminals. Unstructured interviews are also appropriate when researchers and subjects are together for an extended time, such as a researcher accompanying police on patrol.

In other field research situations, interviews will be somewhat more structured. The conversational approach may be difficult to use with officials in criminal justice or other agencies, who will respond best (at least initially) to a specific set of open-ended questions. This is because it is usually necessary to arrange appointments to conduct field research interviews with judges, prosecutors, bail commissioners, and other officials. Having arranged such an appointment, it would be awkward to initiate a casual conversation in hopes of eliciting the desired information.

At the same time, one of the special strengths of field research is its flexibility in the field. Even during structured interviews with public officials, the answers evoked by initial questions should shape subsequent ones. It isn’t enough merely to ask preestablished questions and record the answers. The researcher needs to ask a question, hear the answer, interpret its meaning relative to the general inquiry, and then frame another question either to explore the earlier answer in more depth or to redirect the respondent’s attention to an area more relevant to the inquiry. In short, the researcher needs to be able to listen, think, and talk almost simultaneously.

For example, Michael Maxfield and W. Carsten Andresen (2002) studied the use of video recording equipment to document traffic stops by New Jersey State Police. Very little published research existed on either the use of video equipment by police or even the ways in which
they go about doing traffic enforcement. Part of the research involved semistructured field interviews with police commanders, supervisors, and officers. Figure 10.1 shows the interview guide Maxfield and Andresen used with supervisors. This guide was just that—a guide. Some supervisors were friendly and wanted to talk, so the interview became more of a conversation that eventually yielded answers to the queries in the guide. Others were wary, probably because the agency had been subject to criticism over alleged discriminatory enforcement. Interviews with such supervisors were brief, and they followed the guide very closely.

The discussion of probes in Chapter 9 offers useful guidelines for getting more in-depth answers without biasing later answers. Researchers must learn the skill of being a good listener, of being more interested than interesting. They learn to say things like “How is that?” “In what ways?” “How do you mean that?” and “What would be an example of that?” They learn to look and listen expectantly and to let the person they are interviewing break the silence.

At its best, a field research interview is much like normal conversation. Because of this, it is essential to keep reminding ourselves that we are not having a normal conversation. In normal conversations, each of us wants to come across as an interesting, worthwhile person. Often, we don’t really hear each other because we’re too busy thinking of what we’ll say next. As an interviewer, the desire to appear interesting is counterproductive to the task. We need to make the other person seem interesting by being interested ourselves. By the way, if you do this, people will actually regard you as a great conversationalist.

Like other aspects of field research, interviewing improves with practice. Fortunately, it is something you can practice any time you want. Practice on your friends.
Gaining Access to Subjects

Arranging access to subjects in formal organizations or those in subcultures begins with an initial contact.

Suppose you decide to undertake field research on a community corrections agency in a large city. Let’s assume that you do not know a great deal about the agency and that you will identify yourself as a researcher to staff and other people you encounter. Your research interests are primarily descriptive: You want to observe the routine operations of the agency in the office and elsewhere. In addition, you want to interview agency staff and persons who are serving community corrections sentences. This section will discuss some of the ways you might prepare before you conduct your interviews and direct observations.

As usual, you are well advised to begin with a search of the relevant literature, filling in your knowledge of the subject and learning what others have said about it.

Gaining Access to Formal Organizations

Any research on a criminal justice institution, or on persons who work either in or under the supervision of an institution, normally requires a formal request and approval. One of the first steps in preparing for the field, then, is to arrange access to the community corrections agency.

Obtaining initial approval can be confusing and frustrating. Many criminal justice agencies in large cities have a complex organization, combining a formal hierarchy with a bewildering variety of informal organizational cultures. For example, criminal courts are highly structured organizations in which a presiding judge may oversee court assignments and case scheduling for judges and many support personnel. At the same time, courts are chaotic organizations in which three constellations of professionals—prosecutors, defense attorneys, and judges—episodically interact to process large numbers of cases.

To further complicate field research in criminal justice agencies, the obvious strategy of gaining approval from a single executive—such as a corrections commissioner or police chief—does not guarantee that operations staff at lower levels will cooperate. Researchers are often required to negotiate access at multiple levels. For example, in a project to study auto theft in New Jersey Maxfield has worked with police chiefs, county prosecutors, the state attorney general, executives from insurance agencies, and members of other organizations throughout the state. However, seeking to interview a staff person in a diversionary program for juvenile auto thieves in one New Jersey county, Maxfield was referred to a judge, who then sought approval from the county’s chief judge, who subsequently forwarded the request to a state agency that oversees New Jersey courts. Approval was gained after about 11 months.

Continuing with the example of your research on community corrections, the best strategy in gaining access to virtually any other formal criminal justice organization is to use a four-step procedure: sponsor, letter, phone call, and meeting. Our discussion of these steps assumes that you will begin your field research by interviewing the agency executive director and gaining that person’s approval for subsequent interviews and observations.

Sponsor The first step is to find a sponsor—a person who is personally known to and respected by the executive director. Ideally, a sponsor will be able to advise you on a person to contact, that person’s formal position in the organization, and her or his informal status, including her or his relationships with other key officials. Such advice can be important in helping you initiate contact with the right person while avoiding people who have bad reputations.

For example, you may initially think that a particular judge who is often mentioned in newspaper stories about community corrections would be a useful source of information. However, your sponsor might advise you that
the judge is not held in high regard by prosecutors and community corrections staff. Your association with this judge would generate suspicion on the part of other officials whom you might eventually want to contact and may frustrate your attempts to obtain information.

By the way, finding the right sponsor is often the most important step in gaining access. It may, in fact, require a couple of extra steps, because you might first need to ask a professor whether she or he knows someone. You could then contact that person (with the sponsorship of your professor) and ask for further assistance.

For purposes of illustration, we will assume that your professor is knowledgeable, well connected, and happy to act as your sponsor. Your professor confirms your view that it is best to begin with the executive director of community corrections.

**Letter** Next, write a letter to the executive director. Your letter should have three parts: introduction, brief statement of your research purpose, and action request. See Figure 10.2 for an example. The introduction begins by naming your sponsor, thus immediately establishing your mutual acquaintance. This is

---

Jane Adams  
Executive Director  
Chaos County Community Corrections  
Anxiety Falls, Colorado  
1 May 2006

Dear Ms. Adams:

My colleague, Professor Marcus Nelson, suggested I contact you for assistance in my research on community corrections. I will be conducting a study of community corrections programs and wish to include the Chaos County agency in my research.

Briefly, I am interested in learning more about the different types of sentences that judges impose in jurisdictions with community corrections agencies. As you know, Colorado’s community corrections statute grants considerable discretion to individual counties in arranging locally administered corrections programs. Because of this, it is generally believed that a wide variety of corrections programs and sentences have been developed throughout the state. My research seeks to learn more about these programs as a first step toward developing recommendations that may guide the development of programs in other states. I also wish to learn more about the routine administration of a community corrections program such as yours.

I would like to meet with you to discuss what programs Chaos County has developed, including current programs and those that were considered but not implemented. In addition, any information about different types of community corrections sentences that Chaos County judges impose would be very useful. Finally, I would appreciate your suggestions on further sources of information about community corrections programs in Chaos County and other areas.

I will call your office at about 10:00 a.m. on Monday, May 8, to arrange a meeting. If that time will not be convenient, or if you have any questions about my research, please contact me at the number below.

Thanks in advance for your help.

Sincerely,

Alfred Nobel  
Research Assistant  
Institute for Advanced Studies  
(201) 555-1212

---

**Figure 10.2** Sample Letter for Sponsor
a key part of the process; if you do not name a sponsor, or if you name the wrong sponsor, you might get no further.

Next, you describe your research purpose succinctly. This is not the place to give a detailed description as you would in a proposal. If possible, keep the description to one or two paragraphs, as in Figure 10.2. If a longer description is necessary to explain what you will be doing, you should still include only a brief description in your introductory letter, referring the reader to a separate attachment in which you summarize your research.

The action request describes what immediate role you are asking the contact person to play in your research. You may simply be requesting an interview, or you may want the person to help you gain access to other officials. Notice how the sample in Figure 10.2 mentions both an interview and “suggestions on further sources of information about community corrections.” In any case, you will usually want to arrange to meet or at least talk with the contact person. That leads to the third step.

**Phone Call**  You probably already know that it can be difficult to arrange meetings with public officials (and often professors) or even to reach people by telephone. You can simplify this task by concluding your letter with a proposal for this step: arranging a phone call. The example in Figure 10.2 specifies a date and approximate time when you will call. To be safe, specify a date about a week from the date of your letter. Notice also the request that the executive director call you if some other time will be more convenient.

When you make the call, the executive director will have some idea of who you are and what you want. She will also have had the opportunity to contact your sponsor if she wants to verify any information in your introductory letter.

The actual phone call should go smoothly. Even if you are not able to talk with the executive director personally, you will probably be able to talk to an assistant and make an appointment for a meeting (the next step). Again, this will be possible because your letter described what you eventually want—a meeting with the executive director—and established your legitimacy by naming a sponsor.

**Meeting**  The final step is meeting with or interviewing the contact person. Because you have used the letter–phone call–meeting procedure, the contact person may have already taken preliminary steps to help you. For example, because the letter in Figure 10.2 indicates that you wish to interview the executive director about different types of community corrections sentences, she may have assembled some procedures manuals or reports in preparation for your meeting.

This procedure generally works well in gaining initial access to public officials or other people who work in formal organizations. Once initial access is gained, it is up to the researcher to use interviewing skills and other techniques to elicit the desired information. This is not as difficult as it might seem to novice (or apprentice) researchers for a couple of reasons.

First, most people are at least a bit flattered that their work, ideas, and knowledge are of interest to a researcher. And researchers can take advantage of this with the right words of encouragement. Second, criminal justice professionals are often happy to talk about their work with a knowledgeable outsider. Police, probation officers, and corrections workers usually encounter only their colleagues and their clients on the job. Interactions with colleagues become routinized and suffused with office politics, and interactions with clients are common sources of stress for them. Talking with an interested and knowledgeable researcher is often seen as a pleasant diversion.

By the same token, Richard Wright and Scott Decker (1994) report that most members of their sample of active burglars were both happy and flattered to discuss the craft of burglary. Because they were engaged in illegal
activities, burglars had to be more circumspect about sharing their experiences, unlike the way many people talk about events at work. As a result, burglars enjoyed the chance to describe their work to interested researchers who both promised confidentiality and treated them “as having expert knowledge normally unavailable to outsiders” (1994:26).

Gaining Access to Subcultures

Research by Wright and Decker illustrates how gaining access to subcultures in criminal justice—active criminals, deviants, juvenile gangs, and inmates—requires tactics that differ in some respects from those used to meet with public officials. Letters, phone calls, and formal meetings are usually not appropriate for initiating field research among active offenders. However, the basic principle of using a sponsor to gain initial access operates in much the same way, although the word informant is normally used to refer to someone who helps make contact with subcultures.

Informants may be people whose job involves working with criminals—for example, police, juvenile caseworkers, probation officers, attorneys, and counselors at drug clinics. Lawyers who specialize in criminal defense work can be especially useful sources of information about potential subjects. Francis Gant and Peter Grabosky (2001) contacted a private investigator to help locate car thieves and people working in auto-related businesses who were reputed to deal in stolen parts.

As McCall (1978:31) noted, obtaining initial access to criminals is not usually difficult, but chains of referrals may be necessary:

If a researcher wants to make contact with, say, a bootlegger, he thinks of a person he knows who is closest to the social structure of bootlegging. Perhaps that person will be a police officer, a judge, a liquor store owner, a crime reporter, or a recently arrived Southern migrant. If he doesn’t personally know a judge or a crime reporter, he surely knows someone (his own lawyer . . . ) who does and who would be willing to introduce him. By means of a very short chain of such referrals, the researcher can obtain an introduction to virtually any type of criminal.

For example, in their study of New York heroin users, Bruce Johnson and associates (1985:16, 195–97) initially recruited former addicts they contacted through methadone treatment centers. These persons, employed as research field-workers, used their network of social contacts in two Harlem neighborhoods to identify active heroin users who would serve as research subjects. And the first group of research subjects identified additional heroin users and criminals, who themselves became subjects.

Wright and Decker (1994) were fortunate to encounter a former offender who was well connected with active criminals. Playing the role of sponsor, the ex-offender helped researchers in two related ways. First, he referred them to other people, who, in turn, found active burglars willing to participate in the study. Second, he was well known and respected among burglars, and his sponsorship of the researchers made it possible for them to study a group of naturally suspicious subjects.

A different approach for gaining access to subcultures is to hang around places where criminals hang out. Wright and Decker rejected that strategy as a time-consuming and uncertain way to find burglars, in part because they were not sure where burglars hung out. In contrast, Bruce Jacobs (1996) initiated contact with street-level drug dealers by hanging around and being noticed in locations known for crack availability. Consider how this tactic might make sense for finding drug dealers, whose illegal work requires customers. In contrast, the offense of burglary is more secretive, and it’s more difficult to imagine how one would find an area known for the presence of burglars.

Whatever techniques are used to identify subjects among subcultures, it is generally not
possible to produce a probability sample, and so the sample cannot be assumed to represent some larger population within specified confidence intervals. It is also important to think about potential selection biases in whatever procedures are used to recruit subjects. Although we can’t make probability statements about samples of active offenders, such samples may be representative of a subculture target population.

Selecting Cases for Observation

This brings us to the more general question of how to select cases for observation in field research. The techniques used by Wright and Decker, as well as by many other researchers who have studied active criminals, combine the use of informants and what is called snowball sampling. As we mentioned in Chapter 8, with snowball sampling, initial research subjects (or informants) identify other persons who might also become subjects, who, in turn, suggest more potential subjects, and so on. In this way, a group of subjects is accumulated through a series of referrals.

Wright and Decker’s (1994) study provides a good example. The ex-offender contacted a few active burglars and a few “street-wise noncriminals,” who, in turn, referred researchers to additional subjects, and so on. This process is illustrated in Figure 10.3, which shows the chain of referrals that accumulated a snowball sample of 105 individuals.

Starting at the top of Figure 10.3, the ex-offender put researchers in contact with two subjects directly (001 and 003), a small-time criminal, three street-wise noncriminals, a crack addict, a youth worker, and someone serving probation. Continuing downward, the small-time criminal was especially helpful, identifying 12 subjects who participated in the study (005, 006, 008, 009, 010, 021, 022, 023, 025, 026, 030, 032). Notice how the snowball effect continues, with subject 026 identifying subjects 028 and 029. Notice also that some subjects were themselves “nominated” by more than one source. In the middle of the bottom row in Figure 10.3, for example, subject 064 was mentioned by subjects 060 and 061.

There are, of course, other ways of selecting subjects for observation. Chapter 8 discussed the more conventional techniques involved in probability sampling and the accompanying logic. Although the general principles of representativeness should be remembered in field research, controlled sampling techniques are often not possible.

As an illustration, consider the potential selection biases involved in a field study of deviants. Let’s say we want to study a small number of drug dealers. We have a friend who works in the probation department of a large city who is willing to introduce us to people convicted of drug dealing and sentenced to probation. What selection problems might result from studying subjects identified in this way? How might our subjects not be representative of the general population of drug dealers? If we work our way backward from the chain of events that begins with a crime and ends with a criminal sentence, the answers should become clear.

First, drug dealers sentenced to probation may be first-time offenders or persons convicted of dealing small amounts of “softer” drugs. Repeat offenders and “kingpin” cocaine dealers will not be in this group. Second, it is possible that people initially charged with drug dealing were convicted of simple possession through a plea bargain; because of our focus on people convicted of dealing, our selection procedure will miss this group as well. Finally, by selecting dealers who have been arrested and convicted, we may be gaining access only to those less skilled dealers who got caught. More skilled or experienced dealers may be less likely to be arrested in the first place; they may be different in important ways from the dealers we wish to study. Also, if dealers in street drug markets are more likely to be arrested than dealers who work through social networks of friends and acquaintances, a sample based on arrested dealers could be biased in more subtle ways.

To see why this raises an important issue in selecting cases for field research, let’s return
again to the sample of burglars studied by Wright and Decker. Notice that their snowball sample began with an ex-offender and that they sought out active burglars. An alternative approach would be to select a probability or other sample of convicted burglars, perhaps in prison or on probation. But Wright and Decker rejected this strategy for sampling because of the possibility that they would overlook burglars who had not been caught. After accumulating their sample, the researchers were in a position to test this assumption by examining arrest records for their subjects. Only about one-fourth of the active burglars had ever been convicted of burglary; an additional one-third had been arrested for burglary but not convicted. More than 40 percent had no burglary arrests, and 8 percent had never been arrested for any offense (1994:12).

Putting all this together, Wright and Decker concluded that about three-fourths of their subjects would not have been eligible for inclusion if the researchers had based their sample on persons convicted of burglary. Thus, little overlap exists between the population of active burglars and the population of convicted burglars.

### Purposive Sampling in Field Research

Sampling in field research tends to be more complicated than in other kinds of research. In many types of field studies, researchers attempt to observe everything within their field of study; thus, in a sense, they do not sample...
at all. In reality, of course, it is impossible to observe everything. To the extent that field researchers observe only a portion of what happens, then, what they do observe is a de facto sample of all the possible observations that might have been made. We can seldom select a controlled sample of such observations. But we can keep in mind the general principles of representativeness and interpret our observations accordingly.

The ability to systematically sample cases for observation depends on the degree of structure and predictability of the phenomenon being observed. This is more of a general guideline than a hard-and-fast rule. For example, the actions of youth gangs, burglars, and auto thieves are less structured and predictable than those of police officers. It is possible to select a probability sample of police officers for observation because the behavior of police officers in a given city is structured and predictable in many dimensions. Because the population of active criminals is unknown, however, it is not possible to select a probability sample for observation.

This example should call to mind our discussion of sampling frames in Chapter 8. A roster of police officers and their assignments to patrol sectors and shifts could serve as a sampling frame for selecting subjects to observe. No such roster of gang members, burglars, and auto thieves is available. Criminal history records could serve as a sampling frame for selecting persons with previous arrests or convictions, subject to the problems of selectivity we have mentioned.

Now consider the case in which a sampling frame is less important than the regularity of a process. The regular, predictable passage of people on city sidewalks makes it possible to systematically select a sample of cases for observation. There is no sampling frame of pedestrians, but studies such as Painter’s research (1991) on the effects of street lighting can depend on the reliable flow of passers by who may be observed.

<table>
<thead>
<tr>
<th>Sampling Dimension</th>
<th>Variation in</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>Behavior and characteristics</td>
</tr>
<tr>
<td>Space</td>
<td>Behavior</td>
</tr>
<tr>
<td>Time, micro</td>
<td>Behavior by time of day, day of week</td>
</tr>
<tr>
<td></td>
<td>Lighting by time of day</td>
</tr>
<tr>
<td></td>
<td>Business, store, entertainment activities by time of day, day of week</td>
</tr>
<tr>
<td>Time, macro</td>
<td>Behavior by season, holiday</td>
</tr>
<tr>
<td></td>
<td>Entertainment by season, holiday</td>
</tr>
<tr>
<td>Weather</td>
<td>Behavior by weather</td>
</tr>
</tbody>
</table>

In an observational study such as Painter’s, we might also make observations at a number of different locations on different streets. We could pick the sample of locations through standard probability methods, or, more likely, we could use a rough quota system, observing wide streets and narrow ones, busy streets and quiet ones, or samples from different times of day. In a study of pedestrian traffic, we might also observe people in different types of urban neighborhoods—comparing residential and commercial areas, for example.

Kipke and associates (1998) combined purposive and probability sampling techniques for finding street youths. First, places where street youth hung out were identified by observation, and by interviewing young people using shelters. Then probability samples of individuals at those locations were selected by counting people who appeared to be ages 12–24 and randomly selecting subjects for interviews. An appendix to the authors’ report presents samples of forms to record youth observed at different places, and to sample subjects from different sized groups.

Table 10.1 summarizes different sampling dimensions that might be considered in planning field research. The behavior of people, together with the characteristics of people and
places, can vary by population group, location, time, and weather. We have already touched on the first two in this chapter; now we will briefly discuss how sampling plans might consider time and weather dimensions.

People tend to engage in more out-of-door activities in fair weather than in wet or snowy conditions. In northern cities, people are outside more when the weather is warm. Any study of outdoor activity should therefore consider the potential effects of variation in the weather. For example, in Painter’s study of pedestrian traffic before and after improvements in street lighting, it was important to consider weather conditions during the times observations were made.

Behavior also varies by time, presented as micro- and macrodimensions in Table 10.1. City streets in a central business district are busiest during working hours; more people are in residential areas at other times. And, of course, people do different things on weekends than during the work week. Seasonal variation, the macro time dimension, may also be important in criminal justice research. Daylight is longer in summer months, which affects the amount of time people spend outdoors. Shopping peaks from Thanksgiving to Christmas, increasing the number of shoppers, who along with their automobiles may become targets for thieves. A study of assault in Dallas by Keith Harries and associates (Harries, Stadler, and Zdorkowski, 1984) found variation in assault rates by time of year and day of the week, with assaults tending to peak in summer months and on weekends. Marcus Felson and associates (1996) describe variations in the rhythms of activity at New York’s Port Authority bus terminal. Time of day, day of week, season, and weather affect the facility’s use by bus passengers, transients, and a variety of offenders.

In practice, controlled probability sampling is seldom used in field research. Different types of purposive samples are much more common. Patton (2001:230) describes a broad range of approaches to purposive sampling and offers a useful comparison of probability and purposive samples:

The logic and power of probability sampling derive from statistical probability theory. A random and statistically representative sample permits confident generalization from a sample to a larger population. . . . The logic and power of purposeful sampling lies in selecting information-rich cases for study in depth. (emphasis in original)

Nonetheless, if researchers understand the principles and logic of more formal sampling methods, they are likely to produce more effective purposive sampling in field research.

In field research, it is important to bear in mind questions that correspond to two stages of sampling. First, to what extent are the total situations available for observation representative of the more general class of phenomena we wish to describe and explain? For example, are the three juvenile gangs being observed representative of all gangs? Second, are actual observations within those total situations representative of all the possible observations? For example, have we observed a representative sample of the members of the three gangs? Have we observed a representative sample of the interactions that took place? Even when controlled probability sampling methods are impossible or inappropriate, the logical link between representativeness and generalizability still holds. Careful purposive sampling can produce observations that represent the phenomena of interest. However, we cannot make probabilistic statements of representativeness like those examined in Chapter 8.

Recording Observations

Many different methods are available for collecting and recording field observations.

Just as there is great variety in the types of field studies we might conduct, we have many options for making records of field observations.
In conducting field interviews, for example, researchers almost certainly write notes of some kind, but they might also electronically record interviews. Video recording may be useful in field interviews to capture visual images of dress and body language. Photographs or videos can be used to make records of visual images such as a block of apartment buildings before and after some physical design change or to serve as a pretest for an experimental neighborhood cleanup campaign. This technique was used by Robert Sampson and Stephen Raudenbush (1999) in connection with probability samples of city blocks in Chicago. Videotapes were made of sampled blocks, and the recordings were then viewed to assess physical and social conditions in those areas.

We can think of a continuum of methods for recording observations. At one extreme is traditional field observation and note taking with pen and paper, such as we might use in field interviews. The opposite extreme includes various types of automated and remote measurement, such as videos, devices that count automobile traffic, or computer tabulations of mass transit users. In between is a host of methods that have many potential applications in criminal justice research.

Of course, the methods selected for recording observations are directly related to issues of measurement—especially how key concepts are operationalized. Thinking back to our discussion of measurement in Chapter 5, you should recognize why this is so. If, for example, we are interested in policies to increase nighttime pedestrian traffic in some city, we might want to know why people do or do not go out at night and how many people stroll around different neighborhoods. Interviews—in connection with field research or a survey—can determine people’s reasons for going out or not; video recordings of passersby can provide simple counts. By the same token, a traffic-counting device can produce information about the number of automobiles that pass a particular point on the road, but it cannot measure what the blood alcohol levels of drivers are, whether riders are wearing seat belts, or how fast a vehicle is traveling.

Automated devices to record observations have many potential uses. In earlier chapters we have described research on electronic monitoring (ELMO) technology that records the physical location of individuals. In addition to the ELMO telephone technology—studied by Terry Baumer and associates (Baumer, Maxfield, and Mendelsohn, 1993)—other useful technological aids include devices for making videos of offenders in their home, devices that allow voice pattern recognition, and global satellite tracking devices similar to those attached to animals whose movements are of interest to zoologists.

**Cameras and Sound Recorders**

Video cameras may be used in public places to record relatively simple phenomena, such as the passage of people or automobiles, or more complex social processes. For several years, London police have monitored traffic conditions at dozens of key intersections using video cameras mounted on building rooftops. In fact the 2007 Road Atlas for Britain includes the locations of stationary video cameras on its maps. Since 2003, video cameras have monitored all traffic entering central London as part of an effort to reduce traffic. The license plates of vehicles that do not register paying a toll are recorded, and violation notices sent to owners. Ronald Clarke (1996) studied speeding in Illinois, drawing on observations automatically recorded by cameras placed at several locations throughout the state.

In an effort to develop ways to control aggressive panhandling in New York, George Kelling made videotapes of “squeegee people,” who wipe the windshields of motorists stopped at intersections and then demand payment for their services. Kelling taped naturally occurring interactions between squeegee people and motorists, and he also staged contacts that involved plainclothes police in unmarked cars.
The video recordings made it possible to study the specific tactics used by squeegee people to intimidate motorists and the motorists’ reaction to unwanted windshield service (Kelling and Coles, 1996:141–43).

Still photographs may be appropriate to record some types of observations, such as graffiti or litter. Photos have the added benefit of preserving visual images that can later be viewed and coded by more than one person, thus facilitating interrater reliability checks. For example, if we are interested in studying pedestrian traffic on city streets, we might gather data about what types of people we see and how many there are. As the number and complexity of our observations increase, it becomes more difficult to reliably record how many males and females we see, how many adults and juveniles, and so on. Taking photographs of sampled areas will enable us to be more confident in our measurements and will also make it possible for another person to check on our interpretation of the photographs.

In addition to their use in interviews, audio recorders are useful for dictating observations. For example, a researcher interested in patterns of activity on urban streets can dictate observations while riding through selected areas in an automobile. It is possible to dictate observations in an unstructured manner, describing each street scene as it unfolds. Or a voice recorder can be used more like an audio checklist, with observers noting specified items seen in preselected areas.

**Field Notes**

Even audio recorders and cameras cannot capture all the relevant aspects of social processes. Most field researchers make some records of observations as written notes, perhaps in a field journal. Field notes should include both empirical observations and interpretations of them. They should record what we “know” we have observed and what we “think” we have observed. It is important, however, that these different kinds of notes be identified for what they are. For example, we might note that person X approached and handed something to person Y—a known drug dealer—that we think this was a drug transaction, and that we think person X was a new customer.

We can anticipate some of the most important observations before we begin the study; others will become apparent as our observations progress. Our note taking will be easier if we prepare standardized recording forms in advance. In a study of nighttime pedestrian traffic, for example, we might anticipate the characteristics of pedestrians who are most likely to be useful for analysis—age, gender, ethnicity, and so forth—and prepare a form on which the actual observations can easily be recorded. Or we might develop a symbolic shorthand in advance to speed up recording. In studying participation at a meeting of a community crime prevention group, we might want to construct a numbered grid of the different sections of the meeting room; then we could record the locations of participants easily, quickly, and accurately.

None of this advance preparation should limit recording of unanticipated events and aspects of the situation. Quite the contrary: The speedy handling of anticipated observations gives the researcher more freedom to observe the unanticipated.

Every student is familiar with the process of taking notes. Good note taking in field research requires more careful and deliberate attention and involves some specific skills. Three guidelines are particularly important.

First, don’t trust your memory any more than you have to; it’s untrustworthy. Even if you pride yourself on having a photographic memory, it’s a good idea to take notes, either during the observation or as soon afterward as possible. If you are taking notes during the observation, do it unobtrusively because people are likely to behave differently if they see you writing down everything they say or do.

Second, it’s usually a good idea to take notes in stages. In the first stage, you may need
to take sketchy notes (words and phrases) to keep abreast of what’s happening. Then remove yourself and rewrite your notes in more detail. If you do this soon after the events you’ve observed, the sketchy notes will help you recall most of the details. The longer you delay, the less likely you are to recall things accurately and fully. James Roberts (2002), in his study of aggression in New Jersey nightclubs, was reluctant to take any notes while inside clubs, so he retired to his car to make sketchy notes about observations, then wrote them up in more detail later.

Third, you will inevitably wonder how much you should record. Is it really worth the effort to write out all the details you can recall right after the observation session? The basic answer is yes. In field research, you can’t really be sure what’s important and unimportant until you’ve had a chance to review and analyze a great volume of information, so you should record even things that don’t seem important at the time. They may turn out to be significant after all. In addition, the act of recording the details of something “unimportant” may jog your memory on something that is important.

Structured Observations
Field notes may be recorded on highly structured forms in which observers mark items in much the same way a survey interviewer marks a closed-ended questionnaire. For example, Steve Mastrofski and associates (1998:11) describe how police performance can be recorded on field observation questionnaires:

Unlike ethnographic research, which relies heavily on the field researcher to make choices about what to observe and how to interpret it, the observer using [structured observation] is guided by . . . instruments designed by presumably experienced and knowledgeable police researchers.

Training for such efforts is extensive and time consuming. But Mastrofski and associates compared structured observation to closed-ended questions on a questionnaire. If researchers can anticipate in advance that observers will encounter a limited number of situations in the field, those situations can be recorded on structured observation forms. And, like closed-ended survey questions, structured observations have higher reliability.

In a long-term study, Ralph Taylor (1999) developed forms to code a range of physical characteristics in a sample of Baltimore neighborhoods. Observers recorded information on closed-ended items about housing layout, street length and width, traffic volume, type of non-residential land use, graffiti, persons hanging out, and so forth. Observations were completed in the same neighborhoods in 1981 and 1994.

Because structured field observation forms often resemble survey questionnaires, the use of such forms has the benefit of enabling researchers to generate numeric measures of conditions observed in the field. The Bureau of Justice Assistance (1993:43) has produced a handbook containing guidelines for conducting structured field observations, called environmental surveys. The name is significant because observers record information about the conditions of a specified environment:

[Environmental] surveys seek to assess, as systematically and objectively as possible, the overall physical environment of an area. That physical environment comprises the buildings, parks, streets, transportation facilities, and overall landscaping of an area as well as the functions and conditions of those entities.

Environmental surveys have come to be an important component of problem-oriented policing and situational crime prevention. For example, Figure 10.4 is adapted from an environmental survey form used by the Philadelphia Police Department in drug enforcement initiatives. Environmental surveys are conducted to plan police strategy in drug enforcement in small areas and to assess changes in conditions following targeted enforcement.
Notice that the form can be used to record both information about physical conditions (street width, traffic volume, streetlights) and counts of people and their activities.

Like interview surveys, environmental surveys require that observers be carefully trained in what to observe and how to interpret it. For example, the instructions that accompany the environmental survey excerpted in Figure 10.4 include guidance on coding abandoned automobiles:

Count as abandoned if it appears non-drivable (i.e., has shattered windows, dismantled body parts, missing tires, missing license plates). Consider it abandoned
if it appears that it has not been driven for some time and that it is not going to be for some time to come.

Other instructions provide details on how to count drivable lanes, what sorts of activities constitute “playing” and “working,” how to estimate the ages of people observed, and so on.

**Linking Field Observations and Other Data**

Criminal justice research sometimes combines field research with surveys or data from official records. Although criminal justice research may utilize field methods or sample surveys exclusively, a given project will often collect data from several sources. This is consistent with general advice about using appropriate measures and data collection methods. Simply saying “I am going to conduct an observational study of youth gangs” restricts the research focus at the outset to the kinds of information that can be gathered through observation. Such a study may be useful and interesting, but a researcher is better advised to consider what data collection methods are necessary in any particular study.

For example, in their Baltimore study, Ralph Taylor and associates (Taylor, Shumaker, and Gottfredson, 1985) were interested in the effects of neighborhood physical characteristics on residents’ perceptions of crime problems. Collecting data on physical characteristics required field observation, but sample surveys were necessary for gathering data on perceptions. So Taylor and colleagues conducted a survey of households in the neighborhoods where observations were made. Finally, they compared census data and crime reports from Baltimore police records with information from observations and household surveys.

These three sources of data provided different types of measures that, in combination, yielded a great deal of information about neighborhoods and neighborhood residents in the study sites. Surveys measured respondents’ perceptions and beliefs about crime and other problems. Police records represented crimes reported to police by area residents. Field observations yielded data on physical characteristics and conditions of neighborhoods as rated by outside observers. Census data made it possible to control for the effects of socioeconomic status variables such as income, employment status, and housing tenure.

Finally, Taylor (1999) conducted follow-up field observations and interviews in 1994—13 years after initial data collection. He found that physical conditions had declined since the first series of observations in 1981 but that residents in selected neighborhoods did not perceive that conditions were worse. Thus, Taylor was able to compare changes in conditions observed by researchers in the field to what survey respondents reported to interviewers. The different methods of data collection yielded different results because they measured different dimensions of neighborhoods—field observations measured conditions, whereas surveys measured perceptions.

A long-term research project on community policing in Chicago similarly draws on data from surveys, field observation, and police records (Chicago Community Policing Evaluation Consortium, 2003). As just one example, researchers studied what sorts of activities and discussions emerged at community meetings in 130 of the city’s 270 police beats that covered residential areas. Observers attended meetings, making detailed notes and completing structured observation forms. One section of the form, shown in Figure 10.5, instructed observers to make notes of specific types of neighborhood problems that were discussed at the meeting (Bennis, Skogan, and Steiner, 2003). Here is an excerpt from the narrative notes that supplemented this section of the form:

They . . . had very serious concerns in regard to a dilapidated building in their block that was being used for drug sales. The drug seller’s people were also squat-
In addition, observers distributed questionnaires to community residents and police officers attending each meeting. Items asked how often people attended beat meetings, what sorts of other civic activities they pursued, and whether they thought various other issues were problems in their neighborhood. As an example, the combination of field observation data and survey questionnaires enabled researchers to assess the degree of general social activism among those who attended beat meetings.

Field research can also be conducted after a survey. For example, a survey intended to measure fear of crime and related concepts might ask respondents to specify any area near their residence that they believe is particularly dangerous. Follow-up field visits to the named areas could then be conducted, during which observers record information about physical characteristics, land use, numbers of people present, and so forth.

The box titled “Conducting a Safety Audit” describes how structured field observations were combined with a focus group discussion to assess the scope of environmental design changes in Toronto, Canada.

The flexibility of field methods is one reason observation and field interviews can readily be incorporated into many research projects. And field observation often provides a much richer understanding of a phenomenon that is imperfectly revealed through a survey questionnaire.
In their study of home detention, Terry Baumer and Robert Mendelsohn (1990) supplemented interviews with persons sentenced to electronic monitoring by actually wearing the devices themselves for several days. Their participant observation enabled them to experience punitive features of home detention that would have been difficult to appreciate any other way.

Illustrations of Field Research

Examples illustrate different applications of field research to study shoplifting, traffic, and violence in bars.

Before concluding this chapter on field research, let’s examine some illustrations of the method in action. These descriptions will
provide a clearer sense of how researchers use field observations and interviews in criminal justice research.

**Shoplifting**

We have briefly mentioned the research on shoplifting by Baumer and Rosenbaum (1982). A fuller discussion of this unusual study illustrates much of the potential of field research, some problems that may be encountered, and the clever approaches to these problems adopted by Baumer and Rosenbaum.

This study was undertaken with two objectives: (1) to estimate the prevalence of shoplifting and (2) to assess the effectiveness of store security in identifying shoplifters. As we noted earlier, it is difficult to obtain reliable counts of offenses such as shoplifting and...
so-called victimless crimes. Shoplifting is usually counted only when an offense is witnessed by police or store security personnel. Store inventory records could be used to estimate losses by customer theft, but as Baumer and Rosenbaum note, this obscures the difference between shoplifting by customers and employee theft. The approach they adopted was direct observation.

McCall (1978:20) described three difficulties in observational studies of crime. First, most offenses are relatively rare and unpredictable; an observer might spend weeks or months in the field without witnessing a single burglary, for example. Second, unless they are carefully concealed, field observers, by their very presence, are likely to deter criminals anxious to keep their actions from public view. Finally, there is some danger to field workers who attentively observe crimes in progress while taking careful field notes.

Baumer and Rosenbaum reasoned that shoplifting is largely immune from these problems. First, because shoplifting by definition takes place in only certain specific locations, observers can focus their attention there. Baumer and Rosenbaum selected a large department store in downtown Chicago as the site of their research. The second and third problems mentioned by McCall are also largely absent because shoplifters commit their offenses in the presence of other shoppers; the crime is one of stealth, not confrontation. Field workers could readily adopt the participant observation role of shopper with no more fear for their own safety than that experienced by other shoppers.

Field observers could therefore be used in much the same way undercover retail security workers are deployed in many stores: Pose as a shopper and watch for thefts by other shoppers. When a theft is witnessed, that’s shoplifting.

What about selecting subjects for observation? And how can simple counts of observed thefts be used to estimate the prevalence of shoplifting? “Shoplifting prevalence” was defined as the number of shoppers who stole something during their visit to the store divided by the total number of shoppers. This meant that it was necessary to obtain counts of shoppers as well as shoplifters. Because shopping is a predictable activity (if it were not, there would be no shops!), it is possible to use systematic sampling methods to select people to observe.

Baumer and Rosenbaum chose a variety of days and times to conduct observations. During these times, observers and field supervisors were stationed at each entrance to the department store. Field supervisors counted everyone who entered the store and assigned observers to follow selected individuals from the time they entered until they left. Subjects were selected by systematic sampling in which, say, every 20th person who entered during the observation period was followed by a field observer. Dividing the number of thefts witnessed by the number of shoppers observed yielded an estimate of shoplifting prevalence. For instance, if observers were assigned to follow 500 people, and they witnessed thefts by 20 of them, the prevalence of shoplifting was 4 percent.

Counts of observed thefts also provided one way to assess store security. Comparing the number of thefts witnessed by participant observers to the number detected by store security staff was one obvious way to evaluate store security. Baumer and Rosenbaum devised another method, however, one that solved a potential problem with their approach to counting shoplifting. The problem may have already occurred to you.

If you were an observer in this research, how confident would you be in your ability to detect shoplifting? Stated somewhat differently, how reliable would your observations be? Retail thieves are at least somewhat careful to conceal their crime. As a participant observer, you would have to balance your diligence in watching people to whom you were assigned (observation) with care to not appear too interested in what your fellow shoppers were doing (participation).
Baumer and Rosenbaum used an ingenious adaptation of double-blind experimental methods—discussed in Chapter 7—to assess the reliability of field observers. Unknown to field observers, other research staff were employed as confederate shoplifters. Their job was to enter the store and steal something. Field supervisors, knowing when confederates would be entering the store, periodically instructed observers to follow one of these persons. Confederates committed known thefts and made it possible to determine whether observers were able to detect the incident.

In this way, the researchers could measure the reliability of field observations. If an observer detected 85 percent of the known thefts, that observer’s count of shoplifting by real shoppers could be considered 85 percent accurate. Reliability figures were used to adjust estimates of shoplifting prevalence. For example, if the prevalence estimate for all observers was 4 percent and all observers were 85 percent reliable, then the adjusted prevalence rate was 4.7 percent (0.04/0.85).

One of the lessons from this study is that simple observation is often not so simple. This is especially true when researchers are trying to observe people doing something they wish to conceal from observers. Our next example offers similar guidance in planning observations of people who are doing something required by law.

How Many People Wear Seat Belts?
Although nations in Europe and other parts of the world have required drivers to use seat belts for many years, laws mandating seat belt use were rare in the United States until the 1980s. Problems in enforcing such laws, as well as general resistance to regulation, are among the reasons these laws were late in coming. These concerns promoted interest in the degree of compliance with mandated seat belt use.

Indiana University’s Transportation Research Center (TRC) conducted pilot studies in Indiana to develop methods of observing seat belt use. One of several research reports issued by TRC (Cornwell, Doherty, Mitter, and Drayer, 1989) illustrates sampling issues in field observations and offers a good example of why reliability is so important in field research.

The first decision faced by TRC researchers was how to conduct the observations. They opted for stationary observers posted at road-sides rather than mobile observers riding around in cars. The primary reason for this choice was the need to calculate the rates of seat belt use, expressed as the number of people seen to be wearing belts divided by the number of cars observed. Keeping track of the number of vehicles observed is easier when they can be counted as they pass by some stationary point. Trying to count cars in view while riding in a car yourself is more difficult.

Where to place stationary observers was the next decision. Cornwell and associates describe how detecting seat belt use in moving vehicles is difficult. For this reason, most observers were stationed at controlled intersections—those with a stop sign or traffic light. Because interstate highways do not have intersections, other observers were posted at entrance ramps where cars were traveling more slowly.

Because it is impossible to know the number of autos in operation throughout Indiana at any given time, TRC researchers could not select a probability sample of vehicles to observe. Instead, they used systematic procedures to sample observations on three dimensions that might be associated with seat belt use: time of day, roadway type, and observation site. In addition to considerations about site type, Cornwell and colleagues wished to stratify sites by density of auto ownership. Finding that this was highly correlated with population, they divided Indiana counties into three strata based on population and selected sites within each stratum. Small counties were oversampled to ensure that different geographic areas of the state were represented.

Roadway type included interstate highways, U.S. and state roads, and local streets. Observations were made for each day of the week and at different times of day to represent the various
trip purposes. For example, weekday observations at 6:30–7:30 A.M. represented blue-collar home-to-work trips, later morning times were for white-collar commuters, and midday hours represented lunch breaks.

Standing by the roadside and trying to determine whether drivers and front-seat passengers are wearing safety belts was easier than observing shoplifters but still presented some difficulties. Cornwell and associates describe training procedures and steps they took to maximize the accuracy of observations. All observers completed a training period in which they worked under the supervision of an experienced fieldworker. If you had been a neophyte observer for the seat belt study, here are some of the things you would have been told about your role:

“First, you realize, of course, that we are complete observers, not participant observers. We don’t try to conceal ourselves because we would not be able to see the cars very well if we hid behind signs or something. Sure, drivers can see us standing here with our clipboards, and they may wonder what we’re doing. We’ve posted a sign that says, ‘Traffic Safety Research in Progress,’ but there’s no way they can know that we’re checking seat belt use. So try to relax and don’t feel self-conscious about standing on the side of the road in your orange vest watching cars.

“Second, remember that you should check only passenger cars and station wagons. Trucks, buses, recreational vehicles, and taxicabs are not covered by the law. Also, this is Indiana, not California, so you won’t see too many older cars that don’t have shoulder belts. Even if you do, don’t try to observe them because they aren’t covered by the law either. And remember, we’re interested only in front-seat occupants; ignore the back seat.

“Most importantly, record only safety belts that you can see. If you see a belt in use, great; mark a Y for yes on the coding form. If you see a belt that is NOT in use, mark N for no. But if you can’t see that someone is wearing a belt and you can’t actually see a safety belt that the person is not wearing, mark a U for uncertain. I know you’re a good observer; we’re all good observers. But sometimes cars go by so fast that you can’t see whether a person is wearing a belt. It happens to all of us. Just be sure you mark U when you’re uncertain.

“Now, I have been doing this for about 6 months and I’ve learned some things that they didn’t tell me when I started. We have run into some problems that nobody thought about when they first designed this study. For example, you’re lucky to be starting in May, when the weather is good and the sun is up. In wintertime, people are wearing heavy coats, and it’s much harder to tell whether they are wearing their safety belt. On cold mornings, the windows on cars can be fogged up, making it hard to see inside. Also, when I started working the 6:30–7:30 A.M. shift in January, it was still dark and we couldn’t see anything. We had to revise the time-of-day sampling plan because of that.

“From our site here, we have four lanes of roadway to observe. Concentrate on the traffic lanes closest to us. You will be able to see the cars in those lanes better than you can see those in the far lanes. When traffic is stopped at the light it’s not too bad, but when cars are moving you won’t be able to observe everybody. Just remember, it’s better to observe a smaller number of cars where you can confidently code Y or N than it is to try to see more cars but have to mark U because you can’t be sure. There is no reason to expect that drivers in the far lanes will be any different than those in the near lanes. Code cars in all lanes when possible, but when in doubt focus on those closest to you.”
The National Highway Traffic Safety Administration (NHTSA) now conducts observation studies of safety belt use in most states. Donna Glassbrenner (2005) presents an example of annual tabulations. For further details on how these observations were conducted, see her technical report (Glassbrenner, 2002), which also describes how observation procedures have evolved over the years.

Bars and Violence

Researchers in the first two examples conducted systematic observations for specific purposes and produced quantitative estimates of shoplifting prevalence and seat belt use. Field research is commonly used in more qualitative studies as well, in which precise quantitative estimates are neither available nor needed. A fascinating study of violence in Australian bars by Ross Homel and associates (Homel, Tomsen, and Thommeny, 1992) provides an example.

Anyone with any sort of experience in or knowledge about crime will be at least casually familiar with the relationship between drinking and violence. Psychologists and medical researchers have found a physiological disinhibiting effect of alcohol that can lead to aggression and subsequent violence. In a large proportion of homicides, either the victim or the offender (or both) had been drinking. Barroom brawls are known to most people by reputation if not by experience.

Homel and associates set out to learn how various situational factors related to public drinking might promote or inhibit violence in Australian bars and nightclubs. Think for a moment about how you might approach their research question: “whether alcohol consumption itself contributes in some way to the likelihood of violence, or whether aspects of the drinkers or of the drinking settings are the critical factors” (1992:681). Examining police records might reveal that assault reports are more likely to come from bars than from, say, public libraries. Or a survey might find that self-reported bar patrons were more likely to have witnessed or participated in violence than respondents who did not frequent bars or nightclubs. But neither of these approaches will yield measures of the setting or situational factors that might provoke violence. Field research can produce direct observation of barroom settings and is well suited to addressing the question framed by Homel and associates.

Researchers began by selecting 4 “high-risk” and 2 “low-risk” sites based on Sydney police records and preliminary scouting visits. These 6 sites were visited five or more times, and an additional 16 sites were visited once in the course of scouting.

Visits to bars were conducted by pairs of observers who stayed 2–6 hours at each site. Their role is best described as complete participant because they were permitted one alcoholic drink per hour and otherwise played the role of bar patron. Observers made no notes while on-site. As soon as possible after leaving a bar, they wrote up separate narrative accounts. Later, at group meetings of observers and research staff, the narrative accounts were discussed and any discrepancies resolved. Narratives were later coded by research staff to identify categories of situations, people, and activities that seemed to be associated with the likelihood of violence.

These eventually included physical and social atmosphere, drinking patterns, characteristics of patrons, and characteristics of staff. The researchers began their study by assuming that some thing or things distinguished bars in which violence was common from those in which it was less common. After beginning their field work, however, Homel and associates (1992:684) realized that circumstances and situations were the more important factors:

During field research it soon became apparent that the violent premises are for most of the time not violent. Violent occasions in these places seemed to have characteristics that clearly marked them out from nonviolent times. . . . This unexpectedly helped us refine our ideas about
the relevant situational variables, and to some extent reduced the importance of comparisons with the premises selected as controls.

In other words, the research question was partly restated. What began as a study to determine *why* some bars in Sydney were violent was revised to determine *what situations* seemed to contribute violence.

This illustrates one of the strengths of field research—the ability to make adjustments while in the field. You may recognize this as an example of inductive reasoning. Learning that even violent clubs were peaceful most of the time, Homel and associates were able to focus observers’ attention on looking for more specific features of the bar environment and staff and patron characteristics. Such adjustments “on the fly” would be difficult, if not impossible, if you were doing a survey.

Altogether, field observers made 55 visits to 23 sites, for a total of about 300 hours of field observation. During these visits, observers witnessed 32 incidents of physical violence. Examining detailed field notes, researchers attributed violent incidents to a variety of interrelated factors.

With respect to patrons, violence was most likely to break out in bars frequented by young, working-class males. However, these personal characteristics were deemed less important than the flow of people in and out of bars. Violent incidents were often triggered when groups of males entered a club and encountered other groups of males they did not know.

Physical features mattered little unless they contributed to crowding or other adverse characteristics of the social atmosphere. Chief among social features associated with violence were discomfort and boredom. A crowded, uncomfortable bar with no entertainment spelled trouble.

Drinking patterns made a difference; violent incidents were most common when bar patrons were very drunk. More importantly, certain management practices seemed to produce more drunk patrons. Fewer customers were drunk in bars that had either a restaurant or a snack table. Bars with high cover charges and cheap drinks produced a high density of drunk patrons and violence. The economics of this situation are clear: If you must pay to enter a bar that serves cheap drinks, you’ll get more for your money by drinking a lot.

The final ingredient found to contribute to violence was aggressive bouncers. “Many bouncers seem poorly trained, obsessed with their own machismo (relating badly to groups of male strangers), and some of them appear to regard their employment as a license to assault people” (1992:688). Rather than reducing violence by rejecting unruly patrons, bouncers sometimes escalated violence by starting fights.

Field observation was necessary to identify what situations produce violence in bars. No other way of collecting data could have yielded the rich and detailed information that enabled Homel and associates (1992:688) to diagnose the complex relationships that produce violence in bars:

Violent incidents in public drinking locations do not occur simply because of the presence of young or rough patrons or because of rock bands, or any other single variable. Violent occasions are characterized by subtle interactions of several variables. Chief among these are groups of male strangers, low comfort, high boredom, high drunkenness, as well as aggressive and unreasonable bouncers and floor staff. (1992:688)

**Strengths and Weaknesses of Field Research**

*Validity is usually a strength of field research, but reliability and generalizability are sometimes weaknesses.*

As we have seen, field research is especially effective for studying the subtle nuances of behavior and for examining processes over time. For
these reasons, the chief strength of this method is the depth of understanding it permits.

Conducting field studies of behavior is often more appropriate than trying to measure behavior through surveys. Counts of seat belt use or theft by shoppers obtained through observation are not subject to the effects of social desirability that we might expect in survey questions about those behaviors.

Flexibility is another advantage of field research. Researchers can modify their research design at any time. Moreover, they are always prepared to engage in qualitative field research if the occasion arises, whereas launching a survey requires considerable advance work.

Field research can be relatively inexpensive. Other research methods may require costly equipment or a large research staff, but field research often can be undertaken by one researcher with a notebook and a pen. This is not to say that field research is never expensive. The shoplifting research, seat belt study, and race profiling research described in our running example required many trained observers. Expensive recording equipment may be needed, or the researcher may wish to travel to Australia to replicate the study by Homel and associates.

Field research has its weaknesses, too. First, qualitative field studies seldom yield precise descriptive statements about a large population. Observing casual discussions among corrections officers in a cafeteria, for example, does not yield trustworthy estimates about prison conditions. Nevertheless, it could provide important insights into some of the problems facing staff and inmates in a specific institution.

Second, field observation can produce systematic counts of behaviors and reasonable estimates for a large population of behaviors beyond those actually observed. However, because it is difficult to know the total population of given phenomena—shoppers or drivers, for example—precise probability samples cannot normally be drawn. In designing a quantitative field study or assessing the representativeness of some other study, researchers must think carefully about the density and predictability of what will be observed. Then they must decide whether sampling procedures are likely to tap representative instances of cases they will observe.

More generally, the advantages and disadvantages of different types of field studies can be considered in terms of their validity, reliability, and generalizability. As we have seen, validity and reliability are both qualities of measurements. Validity concerns whether measurements actually measure what they are supposed to, not something else. Reliability is a matter of dependability: If researchers make the same measurement again and again, will they get the same result? Note that some examples we described in this chapter included special steps to improve reliability. Finally, generalizability refers to whether specific research findings apply to people, places, and things not actually observed. Let’s see how field research stacks up in these respects.

Validity
Survey measurements are sometimes criticized as superficial and weak on validity. Observational studies have the potential to yield measures that are more valid. With respect to qualitative field research, “being there” is a powerful technique for gaining insights into the nature of human affairs.

Recall our discussion in Chapter 9 of some of the limitations of using survey methods to study domestic violence. An alternative is a field study in which the researcher interacts at length with victims of domestic violence. The relative strengths of each approach are nicely illustrated in a pair of articles that examine domestic violence in England. Chapter 9 quoted from Catriona Mirrlees-Black’s (1995) article on domestic violence as measured in the British Crime Survey. John Hood-Williams and Tracey Bush (1995) provide a different perspective through their study published in the same issue of the Home Office Research Bulletin.

Tracey Bush lived in a London public housing project (termed “housing estate” in England) for about 5 years. This enabled her
to study domestic violence in a natural setting: “The views of men and women on the estate about relationships and domestic violence have been gathered through the researcher’s network of friends, neighbours, acquaintances, and contacts” (Hood-Williams and Bush, 1995:11). Through long and patient field work, Bush learned that women sometimes “normalize” low levels of violence, seeing it as an unfortunate but unavoidable consequence of their relationship with a male partner. When violence escalates, victims may blame themselves. Victims may also remain in an abusive relationship in hopes that things will get better:

She reported that she wanted the companionship and respect that she had received at the beginning of the relationship. It was the earlier, nonviolent man, whom she had met and fallen in love with, that she wanted back. (1995:13)

Mirrlees-Black (1995) notes that measuring domestic violence is “difficult territory” in part because women may not recognize assault by a partner as a crime. Field research such as that by Hood-Williams and Bush offers an example of this phenomenon and helps us understand why it exists.

Validity is a particular strength of field research. Measurements based on surveys or on simple counts of some phenomenon often give an incomplete picture of the fundamental concept of interest. Survey responses to questions about domestic violence victimization, however carefully phrased, are limited in their ability to count incidents that victims don’t recognize as crimes. More importantly, survey methods cannot yield the rich understanding of domestic violence and its context that Tracey Bush discovered in her 5 years of field research.

In field research, validity often refers to whether the intended meaning of the things observed or people interviewed has been accurately captured. In the case of interviews, Joseph Maxwell (2005) suggests getting feedback on the measures from the people being studied. For example, Wright and Decker (1994) conducted lengthy semistructured interviews with their sample of burglars. The researchers recognized that their limited understanding of the social context of burglary may have produced some errors in interpreting what they learned from subjects. To guard against this, Wright and Decker had some of their subjects review what they thought they had learned:

As the writing proceeded, we read various parts of the manuscript to selected members of our sample. This allowed us to check our interpretations against those of insiders and to enlist their help in reformulating passages they regarded as misleading or inaccurate. . . . The result of using this procedure, we believe, is a book that faithfully conveys the offender’s perspective on the process of committing residential burglaries. (1994:33–34)

This approach is possible only if subjects are aware of the researcher’s role as a researcher. In that case, having informants review draft field notes or interview transcripts can be an excellent strategy for improving validity.

Reliability
Qualitative field research can have a potential problem with reliability. Suppose you characterize your best friend’s political orientations based on everything you know about him or her. There’s certainly no question that your assessment of that person’s politics is at least somewhat idiosyncratic. The measurement you arrive at will appear to have considerable validity. We can’t be sure, however, that someone else will characterize your friend’s politics the same way you do, even with the same amount of observation.

Field research measurements—even in-depth ones—are also often very personal. If, for example, you wished to conduct a field study of bars and honky-tonks near your campus, you might judge levels of disorder on a Friday night to be low or moderate. In contrast, older adults
might observe the same levels of noise and commotion and rate levels of disorder as intolerably high. How we interpret the phenomena we observe depends very much on our own experiences and preferences.

The reliability of quantitative field studies can be enhanced by careful attention to the details of observation. We have seen examples of this in studies of shoplifting and seat belt use. Environmental surveys can likewise promote reliable observations by including detailed instructions on how to classify what is observed. Reliability can be strengthened by reviewing the products of field observations. Homel and associates sought to increase the reliability of observers’ narrative descriptions by having group discussions about discrepancies in reports by different observers.

In a more general sense, reliability will increase as the degree of interpretation required in making actual observations decreases. Participant observation or unstructured interviews may require a considerable degree of interpretation on the part of the observer, and most of us draw on our own experiences and backgrounds in interpreting what we observe. At another extreme, electronic devices and machines can produce very reliable counts of persons who enter a store or of cars that pass some particular point. Somewhere in the middle are field workers who observe shoplifters, motorists, or pedestrians and tabulate some specific behavior.

**Generalizability**

One of the chief goals of social science is generalization. We study particular situations and events to learn about life in general. Usually, nobody is interested in the specific subjects observed by the researcher. Who cares, after all, about the 18 people who told NCVS interviewers about their stolen bicycles? We are interested only if their victimization experiences can be generalized to all U.S. households.

Generalizability can be a problem for qualitative field research. It crops up in two forms. First, the personal nature of the observations and measurements made by the researcher can produce results that will not necessarily be replicated by another independent researcher. If the observation depends in part on the individual observers, it is more valuable as a source of particular insight than as a general truth. You may recognize the similarity between this and the more general issue of reliability.

Second, because field researchers get a full and in-depth view of their subject matter, they can reach an unusually comprehensive understanding. By its very comprehensiveness, however, this understanding is less generalizable than results based on rigorous sampling and standardized measurements.

For example, in connection with an investigation of racial profiling, Maxfield has conducted observational research with the New Jersey State Police. This has taken several forms, but one recent experience involved learning about radar speed enforcement on a 50-mile segment of the New Jersey Turnpike. Maxfield accompanied troopers on a thorough tour of this segment, identifying where radar units were routinely stationed (termed “fishing holes” by troopers). In his field work, he also examined physical characteristics of the roadway; patterns of in- and out-of-state travel; and areas where entrance ramps, slight upward grades, and other features affected vehicle speed. Finally, he gained extensive information on priorities and patterns in speed enforcement—learning what affects troopers’ decisions to stop certain vehicles.

As a result, Maxfield has detailed knowledge about that 50-mile segment of the New Jersey Turnpike. How generalizable is that knowledge? In one sense, learning about “fishing holes” in very specific terms can help identify such sites on other roads. And learning how slight upward grades can slow traffic in one situation may help us understand traffic on other upward grades. But a detailed, idiosyncratic understanding of 50 miles of highway is just that—idiosyncratic. Knowing all there is to know about a straight, largely level stretch
Field research has been an important element in studies of racial profiling for two reasons. First, field research has provided measures of driver behavior that are not dependent on police records. As we have seen in earlier chapters, it is important to compare police records of stops to some other source of information. Second, field research has provided insights into traffic enforcement, an area of policing not much studied by researchers. Field research has also covered the wide range of applications from highly structured counting to less structured field observation and interviews.

Field Measures of Speed
Studies of racial profiling in three states used highly structured techniques to measure the speed of vehicles. The most sophisticated equipment was used by Lange and associates in New Jersey. Here’s how the authors described their setup:

The digital photographs were captured by a TC-2000 camera system, integrated with an AutoPatrol PR-100 radar system, provided by Transcore, Inc. The equipment, other than two large strobe lights, was mounted inside an unmarked van, parked behind pre-existing guide rails along the Turnpike. The camera and radar sensor pointed out of the van’s back window toward oncoming traffic. The two strobe lights were mounted on tripods behind the van and directed toward oncoming traffic. Transcore’s employees operated the equipment. (2005:202)

Equipment was programmed to photograph every vehicle exceeding the speed limit by 15 or more miles per hour. Operators also photographed and timed samples of 25–50 other vehicles per hour. Elsewhere we have described the other element of observation—coding the appearance of driver race from photographs.

Pennsylvania researchers also used radar to measure the speed of vehicles in selected locations throughout the state. Their procedures were less automated, relying on teams of two observers in a car parked on the side of sampled roadways. Undergraduate students at Pennsylvania State University served as observers. They were trained by Pennsylvania state police in the use of radar equipment, completing the same classroom training that was required of troopers. Additional training for observers was conducted on samples of roadways by the project director. State police were trained to operate radar equipment, but not to combine it with systematic field observation of driver characteristics. That was an important research task however. Engel et al. (2004) describe training and field procedures in detail. Their simple field observation form is included in an appendix to their report (p. 312).

Smith and associates tried but rejected stationary observation as a technique for recording

The situation is likely to be different in smaller specialty shops and the department stores in suburban malls that have replaced most urban locations. Similarly, conclusions about seat belt use in Indiana may not apply in New York or Idaho. At the time, Indiana’s law called for a $25 fine for not wearing a seat belt and
speed and observing drivers. They cited the high speed of passing vehicles and glare from windows as problems they encountered. Instead, a research team used mobile observation techniques—observing drivers and timing cars that passed them. Radar was also considered and rejected because it was feared that vehicles having radar detectors, said to be common in North Carolina, would slow down when nearing the research vehicle. Worse, Smith and associates report that truck drivers quickly broadcast word of detected radar, thus eroding the planned unobtrusive measure.

As you can see, the observational component of research in these three states varied quite a lot. Reading the detailed reports from each study offers valuable insights into the kinds of things field researchers must consider.

Observing New Jersey State Police
Other research in New Jersey used less structured field observation techniques. This was because the research purpose was less structured—learning about the general nature of traffic enforcement on New Jersey highways. Andresen, Kelling, and Maxfield were interested in the mechanics of making traffic stops, and what kinds of things troopers considered in deciding which vehicles to stop. Researchers have long accompanied municipal police on patrol and a number of studies have documented their efforts. But, as Andresen points out, only a handful of studies have examined traffic enforcement, and even fewer considered state police.

To study video recording cameras in state police cars, Maxfield and Andresen (2002) rode with state police and watched the equipment in use. They learned that sound quality of recordings was often poor, for a variety of reasons associated with microphones and wireless transmission. It was initially hoped that video records might make it possible to classify the race of drivers, but after watching in-car video monitors the researchers confirmed that poor image quality undermined the potential reliability of that approach. The Rutgers University researchers expected that troopers would be on their best behavior. But they did witness actions by troopers to avoid recording sound and/or images on a few occasions. Even though people behave differently when accompanied by researchers, it’s not uncommon for police to let their guard down a little.

Andresen accompanied troopers on 57 patrols overall, conducting unstructured interviews during the several hours he spent with individual troopers. He adopted the common practice of using an interview guide, a list of simple questions he planned to ask in the field. He took extensive notes while riding, and repeatedly told troopers they could examine his notes. Andresen observed over 150 traffic stops, writing field notes to document who was involved, reasons for the stop, what actions troopers took, and post-stop comments from troopers. He reports that most troopers seemed to enjoy describing their work. And, as you might imagine, troopers’ commentary about traffic enforcement was very interesting.

Semistructured interviews with 57 troopers and field notes from 150 traffic stops produced a large volume of written material. We’ll say a bit more about how Andresen summarized that material in Chapter 11.

could be invoked only if a driver was stopped for some other reason. Compliance may be different in states that have stiffer penalties and where police are permitted to stop drivers for not wearing seat belts.

At the same time, some field studies are less rooted in the local context of the subject under study. Wright and Decker (1994) studied burglars in St. Louis, and it’s certainly reasonable to wonder whether their findings apply to residential burglars in St. Petersburg, Florida. The actions and routines of burglars might be affected by local police strategies, differences in the age or style of dwelling units, or even the type and amount of vegetation screening buildings from the street. However, Wright and Decker draw
general conclusions about how burglars search for targets, what features of dwellings signal vulnerability, how opportunistic knowledge can trigger an offense, and what strategies exist for fencing stolen goods. It’s likely that their findings about the technology and incentives that affect St. Louis burglars apply generally to residential burglars in other cities.

In reviewing reports of field research projects, it’s important to determine where and to what extent the researcher is generalizing beyond her or his specific observations to other settings. Such generalizations may be in order, but it is necessary to judge that. Nothing in this research method guarantees it. As we’ve seen, field research is a potentially powerful tool for criminal justice research, one that provides a useful balance to surveys.

**Main Points**

- Field research is a data collection method that involves the direct observation of phenomena in their natural settings.
- Field observation is usually the preferred data collection method for obtaining information about physical or social settings, behavior, and events.
- Field research in criminal justice may produce either qualitative or quantitative data. Grounded theory is typically built from qualitative field observations. Or observations that can be quantified may produce measures for hypothesis testing.
- Observations made through field research can often be integrated with data collected from other sources. In this way, field observations can help researchers interpret other data.
- Asking questions through a form of specialized interviewing is often integrated with field observation.
- Field researchers may or may not identify themselves as researchers to the people they are observing. Being identified as a researcher may have some effect on what is observed.
- Preparing for the field involves negotiating or arranging access to subjects. Specific strategies depend on whether formal organizations, subcultures, or something in between are being studied.
- Controlled probability sampling techniques are not usually possible in field research. But in some circumstances, researchers can combine probability and nonprobability sampling.
- Snowball sampling is a method for acquiring an ever-increasing number of sample observations. One participant is asked to recommend others for interviewing, and each of these other participants is asked for more recommendations.
- If field observations will be made on a phenomenon that occurs with some degree of regularity, purposive sampling techniques can be used to select cases for observation.
- Alternatives for recording field observations range from video, audio, and other equipment to unstructured field notes. In between are observations recorded on structured forms; environmental surveys are examples.
- Field notes should be planned in advance to the greatest extent possible. However, note taking should be flexible enough to make records of unexpected observations.
- Compared with surveys, field research measurements generally have more validity but less reliability, and field research results cannot be generalized as safely as those based on rigorous sampling and standardized questionnaires.

**Key Terms**

Environmental survey  
Snowball sampling

**Review Questions and Exercises**

1. Think of some group or activity you participate in or are familiar with. In two or three paragraphs, describe how an outsider might effectively go about studying that group or activity. What should he or she read, what contacts should be made, and so on?

2. Review the box titled “Conducting a Safety Audit” by Gisela Bichler. Try conducting a safety audit on your campus or in an area near your campus. If possible, supplement this with satellite photo images of the area where you conduct your audit.
3. Many police departments encourage citizen ride-alongs as a component of community policing. If this is the case for a police or sheriff’s department near you, take advantage of this excellent opportunity to test your observation and unstructured interviewing skills.

4. This question is linked to Exercise 3 in Chapter 9. After tabulating responses to your questions about nearby areas where respondents feel unsafe or believe crime to be a problem, plan and conduct field observations in those areas. First, visit one or more areas and make detailed field notes about the area’s characteristics, thinking especially about how those characteristics might be associated with perceptions of crime problems. Second, use your notes and observations to develop a form for making structured field observations. Third, visit more areas cited by your respondents and use your form to record area characteristics.

**Online Sessions for Chapter 10**

Online Sessions present further examples of selected material covered in each chapter. Your instructor may choose to use these for in-class discussion (as I do with my students) or you can access them at home. To ensure that you always have access to live, correct links for the websites described here, we provide the necessary links on the companion website for Research Methods for Criminal Justice and Criminology at http://www.thomsonedu.com/criminaljustice/maxfield. Once at the companion website, select this specific chapter, click on “Chapter Resources,” then click on “Web Links.”

**Counting Pedestrians** Two related features of Manhattan combine to produce routine counts of pedestrians in different neighborhoods. First, though it’s the nation’s largest city, New York is very much a pedestrian town. Second, retail shops and other businesses depend on walk-up customers attracted by walk-by displays. Areas of Manhattan that appeal to shoppers and tourists therefore regularly sponsor systematic counts of pedestrians passing by sampled points.

Here are links to pedestrian counts for three such areas: Times Square, Grand Central Station, and Downtown—the financial district and site of the World Trade Center. Each site is maintained by a business improvement district, coalitions of business owners that promote development and general well-being in their areas. The document for Grand Central presents details on how the counts are conducted, including the specific locations of observers. Note the enormous numbers of pedestrians in the Times Square area—over 500,000 in a 24-hour period in late August 2005 along Broadway between 45th and 48th Streets.

http://timessquarenyc.org/facts/facts_economic.html

http://www.grandcentralpartnership.org/pdfs/Pedestrian_Counts_Dec%202005.pdf

http://www.downtownny.com/dobusiness/retail/locate/numbers/pedestrian/

**Rating Street Cleanliness** The website for this book includes the running example from previous editions—“Why did crime go down in New York City?” Officials in New York and many researchers pointed to changes in policing, notably the evidence-driven use of “Compstat” as a management tool. The use of timely data to assess performance has now been adopted by virtually all city agencies and departments in New York.

Very detailed and disaggregated performance measures are available on the Web, as we saw in the exercises for Chapter 5. The link for this exercise will take you to a document that describes how street cleanliness is measured by observation. Here you will find: (1) a numerical scale for cleanliness, ranging from acceptably clean (1.0–1.2) to filthy (1.8–3.0); (2) photos to illustrate seven categories of clean and filthy streets; (3) procedures for making observations and filling out scoring sheets. From the scorecard home page, click on the link: “Learn More About Scorecard.” After reviewing the document, you will be able to launch your own street cleanliness rating program.


**Where Are the Speed Cameras?** Different types of traffic cameras are ubiquitous in Great Britain. Red-light cameras record the license plate of vehicles running traffic lights. Cameras monitor cars entering central London during weekdays as part of a traffic control program. Speed cameras combine radar and photo technology to record the
license plate of cars exceeding posted speed limits. One of the manufacturers of speed cameras shows camera locations throughout England and Wales, the first URL. If you explore different regions listed on the website you’ll find some maps show traffic and accident data for areas covered by the cameras. The second URL is an example for the area around Cambridge, northeast of London. This is a good example of collecting data by automated observation, and combining it with data from other sources.

http://www.speedcheck.co.uk/yourArea.htm

http://www.cambssafetycameras.co.uk/

Observing High-Risk Parking Lots In the course of research on car crime in northern New Jersey, Maxfield learned of a wave of headlight thefts. Beginning in 2002, thefts of high-intensity headlights from a certain car make and model became very common. A surveillance video of a theft showed thieves taking two lights in about 40 seconds. Some people were repeatedly victimized, including a Newark police officer whose headlights were stolen on two occasions. But the all-time repeat victim lived in a Bloomfield, New Jersey, apartment complex; this person’s headlights were stolen five times in about 13 months. If you examine an aerial photo of the complex and its parking lot, you’ll see why this was an attractive location for thieves to repeatedly strike.

Using a mapping program that also includes satellite photos (for example, http://maps.google.com), type in this address: 188 Watchung Ave, Bloomfield, NJ. Zoom in a couple of clicks and you will see the pointer in front of an L-shaped building at the corner of Watchung Avenue and Broughton Avenue. That’s where our repeat victim lived and parked his car. Zoom in closer and you will see the parking lot behind the building, partly screened by trees. Next to the building’s driveway on Watchung Avenue (to the east, or right) is the southbound entrance to the Garden State Parkway, an expressway leading to Newark about six miles to the south. You can probably see why this was an ideal place for headlight thieves.

More generally, the availability of satellite photos makes it possible to “observe” substantial detail in just about any part of the country. Other web-based services show photos all over the world.

As another exercise, view the parking lot for the New Jersey Transit train station in Princeton Junction. Type in: 2 Wallace Road, Princeton Junction, NJ. Switch to satellite view and zoom in a couple of clicks. You will see hundreds of cars parked on both sides of the tracks in sprawling parking lots. These are commuter cars, belonging to people traveling the 40 miles or so to New York and Newark. Since they are commuters, the cars remain unguarded for 8 or more hours per day. Not surprisingly, the lot has become a hot spot for car theft.

Now, consider the photo in a slightly more analytic way. It offers a way to approximate the population at risk of car theft—the number of vehicles parked in station lots. Because the lot is full, it’s reasonable to assume that demand for parking is high and it will probably always be full on weekdays.

✪ Additional Readings

Bureau of Justice Assistance, A Police Guide to Surveying Citizens and Their Environment (Washington, DC: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Assistance, 1993). Intended for use in community policing initiatives, this publication is a useful source of ideas about conducting structured observations. Appendixes include detailed examples of environmental surveys. You can also download this publication in text form (no drawings) from the web (http://www.ncjrs.org/txtfiles/polc.txt).

Felson, Marcus, Crime and Everyday Life, 3rd ed. (Thousand Oaks, CA: Sage, 2002). We mentioned this book as an example of criminal justice theory in Chapter 2. Many of Felson’s explanations of how everyday life is linked to crime describe physical features of cities and other land use patterns. This entertaining book suggests many opportunities for conducting field research.

Miller, Joel, Profiling Populations Available for Stops and Searches (London: Home Office Policing and Reducing Crime Unit, 2000). Race-biased policing has been a concern in England for many years. This report presents a thorough description of observation to produce baseline measures of populations eligible to be stopped by police. Similar efforts have been underway in many U.S. states and cities, but published
reports of such studies had not been released as of spring 2003 (http://www.homeoffice.gov.uk/rds/policerspubs1.html, click on #131).

Patton, Michael Quinn, *Qualitative Research and Evaluation Methods*, 3rd ed. (Thousand Oaks, CA: Sage, 2001). We mentioned this book in Chapter 9 as a good source of guidance on questionnaire construction. Likewise, Patton offers in-depth information on observation techniques, along with tips on conducting unstructured and semistructured field interviews. Finally, Patton describes a variety of purposive sampling techniques for qualitative interviewing and field research.

Chapter 11

Agency Records, Content Analysis, and Secondary Data

We’ll examine three sources of existing data: agency records, content analysis, and data collected by other researchers. Data from these sources have many applications in criminal justice research.

Introduction 321
Topics Appropriate for Agency Records and Content Analysis 322
Types of Agency Records 323
Published Statistics 323
Nonpublic Agency Records 326
New Data Collected by Agency Staff 330
Improving Police Records of Domestic Violence 332
Units of Analysis and Sampling 332
Units of Analysis 333
Sampling 334
Reliability and Validity 335
Sources of Reliability and Validity Problems 336

Content Analysis 340
Units of Analysis and Sampling in Content Analysis 341
Coding in Content Analysis 343
Illustrations of Content Analysis 344

Secondary Analysis 347
Putting It All Together: Agency Records and Content Analysis 348

Sources of Secondary Data 350
Advantages and Disadvantages of Secondary Data 351

Main Points 352
Introduction

Agency records, secondary data, and content analysis do not require direct interaction with research subjects.

Except for the complete observer in field research, the modes of observation discussed so far require the researcher to intrude to some degree into whatever he or she is studying. This is most obvious with survey research. Even the field researcher, as we’ve seen, can change things in the process of studying them.

Other ways of collecting data do not involve intrusion by observers. In this chapter, we’ll consider three different approaches to using information that has been collected by others, often as a routine practice. First, a great deal of criminal justice research uses data collected by state and local agencies such as police, criminal courts, probation offices, juvenile authorities, and corrections departments. Federal organizations such as the FBI, the Bureau of Justice Statistics, the Federal Bureau of Prisons, and the National Institute of Corrections compile information about crime problems and criminal justice institutions. In addition, nongovernment organizations such as the National Center for State Courts and the American Prosecutors’ Research Institute collect data from members.

Government agencies gather a vast amount of crime and criminal justice data, probably rivaled only by efforts to produce economic and public health indicators. We refer to such information as “data from agency records.” In this chapter, we will describe different types of such data that are available for criminal justice research, together with the promise and potential pitfalls of using information from agency records.

Second, in content analysis, researchers examine a class of social artifacts—written documents or other types of messages. Suppose, for example, you want to contrast the importance of criminal justice policy and health care policy for Americans in 1992 and 2004. One option is to examine public opinion polls from these years. Another method is to analyze articles from newspapers published in each year. The latter is an example of content analysis: the analysis of communications.

Finally, information collected by others is frequently used in criminal justice research, which in this case involves secondary analysis of existing data. Investigators who conduct research funded by federal agencies such as the National Institute of Justice are usually obliged to release their data for public use. Thus, for example, if you are interested in establishing a system for using bail guidelines to make pretrial release decisions, you might analyze data collected by John Goldkamp and Doris Weiland (1993) in their research on judicial decision guidelines for bail. This chapter briefly describes some sources of secondary data, which are covered more fully on the companion website.

Before we begin to explore each of these sources of data in detail, be aware of one caution about obtrusive and unobtrusive measurement. When you go to the library to consult published statistics about persons under correctional supervision or to conduct secondary analysis of a prison inmate survey directed by the RAND Corporation, you are not interacting with research subjects. Do not, however, be too quick to describe such measurement as unobtrusive. Inmates interviewed by RAND researchers were certainly aware of their role in providing information for research, even if they had no idea that you would later examine the data. You may gather jail census figures from a published report, but those data were originally obtained from a survey questionnaire completed by one or more persons directly involved in the measurement process.

In a general sense, then, most data you obtain from agency records or research projects conducted by others are secondary data. Someone else gathered the original data, usually for purposes that differ from yours.
Topics Appropriate for
Agency Records and
Content Analysis

Agency records support a wide variety of research applications.

Data from agency records or archives may have originally been gathered in any number of ways, from sample surveys to direct observation. Because of this, such data may, in principle, be appropriate for just about any criminal justice research topic.

Published statistics and agency records are most commonly used in descriptive or exploratory studies. This is consistent with the fact that many of the criminal justice data published by government agencies are intended to describe something. For instance, the Bureau of Justice Statistics (BJS) publishes annual figures on prison populations. If we are interested in describing differences in prison populations between states or changes in prison populations from 1990 through 2005, a good place to begin is with the annual data published by BJS. Similarly, published figures on crimes reported to police, criminal victimization, felony court caseloads, drug use by high school seniors, and a host of other measures are available over time—for 25 years or longer in many cases.

Agency records may also be used in explanatory studies. Nancy Sinauer and colleagues (1999) examined medical examiner records for over 1000 female homicide victims in North Carolina to understand the relationship between female homicide and residence in urban versus rural counties. They found that counties on the outskirts of cities had higher female homicide rates than either urban or rural counties.

Agency records are frequently used in applied studies as well. Evaluations of new policies that seek to reduce recidivism might draw on arrest or conviction data for measures of recidivism. A study of drug courts as an alternative way to process defendants with substance abuse problems traced arrest records for experimental and control subjects (Gottfredson, Najaka, and Kearly, 2003). In another study, Lawrence Winner and associates (Winner, Lanza-Kaduce, Bishop, and Frazier, 1997) compared rearrest records for juveniles transferred to the adult court system with those who had been kept in the juvenile system. Sometimes, records obtained from private firms can be used in applied studies; Gisela Bichler and Ronald Clarke (1996) examined records of international telephone calls in their evaluation of efforts to reduce telephone fraud in New York City.

In a different type of applied study, Pamela Lattimore and Joanna Baker (1992) combined data on prison releases and capacities, reincarcerations, and general-population forecasts to develop a mathematical model that predicts future prison populations in North Carolina. This is an example of forecasting, in which past relationships among arrest rates and prison sentences for different age groups are compared with estimates of future population by age group. Assuming that past associations between age, arrest, and prison sentence will remain constant in future years, demographic models of future population can be used to predict future admissions to prison.

Topics appropriate to research using content analysis center on the important links between communication, perceptions of crime problems, individual behavior, and criminal justice policy. The prevalence of violence on fictional television dramas has long been a concern of researchers and public officials (Anderson and Bushman, 2002). Wesley Skogan and Michael Maxfield (1981) studied the link between mass media news stories about crime and people’s perceptions. Numerous attempts have been made to explore relationships between exposure to pornography and sexual assault (see, for example, Pollard, 1995). Mass media also plays an important role in affecting policy action by
public officials. Many studies have examined the influence of media agenda setting for criminal justice policy (see, for example, Chermak and Weiss, 1997; Fishman, 1980).

Research data collected by other investigators, through surveys or field observation, may be used for a broad variety of later studies. National Crime Victimization Survey (NCVS) data have been used by a large number of researchers in countless descriptive and explanatory studies since the 1970s. In a notably ambitious use of secondary data, Robert Sampson and John Laub (1993) recovered life history data on 500 delinquents and 500 nondelinquents originally collected by Sheldon and Eleanor Glueck in the 1940s. Taking advantage of theoretical and empirical advances in criminological research over the ensuing 40 years, Sampson and Laub produced a major contribution to knowledge of criminal career development in childhood.

Existing data may also be considered as a supplemental source of data. For example, a researcher planning to survey correctional facility administrators about their views on the need for drug treatment programs will do well to examine existing data on the number of drug users sentenced to prison terms. Or, if we are evaluating an experimental morale-building program in a probation services department, statistics on absenteeism will be useful in connection with the data our own research will generate.

This is not to say that agency records and secondary data can always provide answers to research questions. If this were true, much of this book would be unnecessary. The key to distinguishing appropriate and inappropriate uses of agency records, content analysis, and secondary data is understanding how these written records are produced. We cannot emphasize this point too strongly. Much of this chapter will underscore the importance of learning where data come from and how they are gathered.

### Types of Agency Records

Researchers use a variety of published statistics and nonpublic agency records.

Information collected by or for public agencies usually falls into one of three general categories: (1) published statistics, (2) nonpublic agency records routinely collected for internal use, and (3) new data collected by agency staff for specific research purposes. Each category varies in the extent to which data are readily available to the researcher and in the researcher's degree of control over the data collection process.

### Published Statistics

Most government organizations routinely collect and publish compilations of data. Examples are the Census Bureau, the FBI, the Administrative Office of U.S. Courts, the Federal Bureau of Prisons, and the BJS. Two of these organizations merit special mention. First, the Census Bureau conducts enumerations and sample surveys for several other federal organizations. Notable examples are the NCVS, Census of Children in Custody, Survey of Inmates in Local Jails, Correctional Populations in the United States, and Survey of Justice Expenditure and Employment.

Second, the BJS compiles data from several sources and publishes annual and special reports on most data series. For example, Criminal Victimization in the United States reports summary data from the NCVS each year. Table 11.1 presents a sample breakdown of victimization rates by race and home ownership from the report for 2004. The BJS also issues reports on people under correctional supervision. These are based on sample surveys and enumerations of jail, prison, and juvenile facility populations. Sample tabulations of females serving sentences in state and federal prison in 2004 are shown in Table 11.2. And the series Federal Criminal Case Processing reports detailed data on federal court activity.
The most comprehensive BJS publication on criminal justice data is the annual *Sourcebook of Criminal Justice Statistics*. Since 1972, this report has summarized hundreds of criminal justice data series, ranging from public perceptions about crime, to characteristics of criminal justice agencies, to tables on how states execute capital offenders. Data from private sources such as the Gallup Poll are included with statistics collected by government agencies. Most importantly, each year’s *Sourcebook* concludes with notes on data sources, appendixes summarizing data collection procedures for major series, and addresses of organizations that either collect or archive original data.

Compilations of published data on crime and criminal justice are readily available from many sources. For example, the *Sourcebook* is available on the web (http://www.albany.edu/sourcebook). Appendix D on the companion website presents a more comprehensive list, together with some guidelines on how to get more information from the BJS and other major sources. At this point, however, we want to suggest some possible uses, and limits, of what Herbert Jacob (1984:9) refers to as being “like the apple in the Garden of Eden: tempting but full of danger . . . [for] the unwary researcher.”

Referring to Tables 11.1 and 11.2, you may recognize that published data from series such as the NCVS or *Correctional Populations in the United States* are summary data, as discussed in Chapters 4 and 6. This means that data are presented in highly aggregated form and cannot be used to analyze the individuals from or about whom information was originally collected. For example, Table 11.2 shows that in 2004, 2789 women were serving sentences of 1 year or more in state and federal correctional institutions in New York. By comparing that figure with figures from other states, we can make some descriptive statements about prison populations in different states. We can also consult earlier editions of *Correctional Populations* to examine trends in prison populations over time or to compare rates of growth from state to state.

Summary data cannot, however, be used to reveal anything about individual correctional facilities, let alone facility inmates. Individual-level data about inmates and institutions are available from the Census Bureau in electronic form, but published tabulations present only summaries.

This is not to say that published data are useless to criminal justice researchers. Highly aggregated summary data from published statistical series are frequently used in descriptive, explanatory, and applied studies. For example, Eric Baumer and colleagues (Baumer, Lauritsen, Rosenfeld, and Wright, 1998) examined

### Table 11.1 Victimization Rates by Type of Crime, Form of Tenure, and Race of Head of Household, 2004 (rates per 1000 households)

<table>
<thead>
<tr>
<th>Household Crimes</th>
<th>Burglary</th>
<th>Motor Vehicle Theft</th>
<th>Theft</th>
<th>Total Number of Households</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race of Head of Household</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>27.6</td>
<td>7.6</td>
<td>121.6</td>
<td>95,605,550</td>
</tr>
<tr>
<td>Black</td>
<td>44.3</td>
<td>15.6</td>
<td>130.6</td>
<td>13,376,960</td>
</tr>
<tr>
<td>Home Ownership</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Owned</td>
<td>24.9</td>
<td>7.1</td>
<td>110.8</td>
<td>79,511,410</td>
</tr>
<tr>
<td>Rented</td>
<td>39.9</td>
<td>12.5</td>
<td>148.9</td>
<td>36,264,170</td>
</tr>
</tbody>
</table>

Source: Adapted from Bureau of Justice Statistics (2006:Tables 16, 56).
UCR, DAWN, and DUF data for 142 U.S. cities from 1984 to 1992. Interested in the relationship between crack cocaine use and robbery, burglary, and homicide rates, the researchers found that high levels of crack cocaine use in the population were found in cities that also had high rates of robbery. In contrast, burglary rates were lower in cities with high levels of crack cocaine use. In another study, Alfred Blumstein and associates (Blumstein, Cohen, and Rosenfeld, 1991) compared published UCR and NCVS data for burglary and robbery and concluded that the different measures of crime covaried relatively consistently from 1973 through 1985.

Ted Robert Gurr (1989) used published statistics on violent crime dating back to thirteenth-century England to examine how social and political events affected patterns of homicide through 1984. A long-term decline in homicide rates has been punctuated by spikes during periods of social dislocation, when:

significant segments of a population have been separated from the regulating institutions that instill and reinforce the basic Western injunctions against interpersonal violence. They may be migrants, demobilized veterans, a growing population of resentful young people for whom there is no social or economic niche, or badly educated young black men trapped in the decaying ghettos of an affluent society. (1989:48–49)

Published data can therefore address questions about highly aggregated patterns or trends—drug use and crime, the covariation in two estimates of crime, or epochal change in fatal violence. Published data also have the distinct advantage of being readily available; a web search or a trip to the library can quickly place several data series at your disposal. You can obtain copies of most publications by the BJS and other Justice Department offices on the Internet; most documents published since about 1994 are available electronically. Some formerly printed data reports are now available only in electronic format.

Electronic formats have many advantages. Data fields may be read directly into statistical or graphics computer programs for analysis. Many basic crime data can be downloaded from the BJS website in spreadsheet formats or read directly into presentation software. More importantly, complete data series are available in electronic formats. Although printed reports of the NCVS limit you to summary tabulations such as those in Table 11.1, electronic and optical media include the original survey data from 80,000 or more respondents.

Of course, before using either original data or tabulations from published sources, researchers must consider the issues of validity and reliability and the more general question of how well these data meet specific research purposes. It would make little sense to use only FBI data on homicides in a descriptive study of domestic violence because murder records measure only incidents of fatal violence. Or data from an annual prison census would not be appropriate for research on changes in sentences to community corrections programs.

<table>
<thead>
<tr>
<th>Region</th>
<th>Total 2004</th>
<th>Percentage Change 2003–2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northeast</td>
<td>8910</td>
<td>−2.2</td>
</tr>
<tr>
<td>Connecticut</td>
<td>1488</td>
<td>−3.9</td>
</tr>
<tr>
<td>Maine</td>
<td>125</td>
<td>0.8</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>741</td>
<td>4.7</td>
</tr>
<tr>
<td>New Hampshire</td>
<td>119</td>
<td>1.7</td>
</tr>
<tr>
<td>New Jersey</td>
<td>1470</td>
<td>−3.1</td>
</tr>
<tr>
<td>New York</td>
<td>2789</td>
<td>−4.3</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>1827</td>
<td>0.2</td>
</tr>
<tr>
<td>Rhode Island</td>
<td>208</td>
<td>−6.3</td>
</tr>
<tr>
<td>Vermont</td>
<td>143</td>
<td>5.9</td>
</tr>
<tr>
<td>Midwest</td>
<td>16,545</td>
<td>5.5</td>
</tr>
<tr>
<td>South</td>
<td>44,666</td>
<td>3.7</td>
</tr>
<tr>
<td>West</td>
<td>22,563</td>
<td>5.1</td>
</tr>
</tbody>
</table>

Source: Adapted from Harrison and Beck (2004:Table 5).
Nonpublic Agency Records

Despite the large volume of published statistics in criminal justice, those data represent only the tip of the proverbial iceberg. The FBI publishes the summary UCR, but each of the nation’s several thousand law enforcement agencies produces an incredible volume of data not routinely released for public distribution. The BJS publication *Correctional Populations in the United States* presents statistics on prison inmates collected from annual surveys of correctional facilities, but any given facility also maintains detailed case files on individual inmates. The volume *Court Caseload Statistics*, published by the National Center for State Courts, contains summary data on cases filed and disposed in state courts, but any courthouse in any large city houses paper or computer files on thousands of individual defendants. Finally, reports on the annual survey of expenditures and employment in criminal justice are sources of summary data on budgets and personnel, but every agency also maintains its own detailed records of expenditures and human resources.

Although we have labeled this data source “nonpublic agency records,” most criminal justice organizations will make such data available to criminal justice researchers. But obtaining agency records is not as simple as strolling into a library and asking for a BJS publication or clicking a download button on the Census Bureau’s web page. Obtaining access to nonpublic records involves many of the same steps we outlined in Chapter 10 for gaining entry for field research.

At the outset, we want to emphasize that the potential promise of agency records is not without cost. Jacob’s caution about the hidden perils to unwary researchers who uncritically accept published data applies to nonpublic agency records as well. On the one hand, we could devote an entire book to describing the potential applications of agency records in criminal justice research, together with advice on how to use and interpret such data. On the other hand, we can summarize that unwritten book with one piece of advice: Understand how agency records are produced. Restated slightly, researchers can find the road to happiness in using agency records by following the paper trail.

By way of illustrating this humble maxim, we present two types of examples. First, we describe two studies in which nonpublic agency records were used to reveal important findings about the etiology of crime and its spatial distribution. Second, we briefly review two studies in which the authors recognized validity and reliability problems but still were able to draw conclusions about the behavior of criminal justice organizations. At the end of this section, we summarize the promise of agency records for research and note cautions that must be exercised if such records are to be used effectively.

Child Abuse, Delinquency, and Adult Arrests

In earlier chapters, we described Cathy Spatz Widom’s research on child abuse as an example of a quasi-experimental design. For present purposes, her research illustrates the use of several different types of agency records.

Widom (1989) identified cases of child abuse and neglect by consulting records from juvenile and adult criminal courts in a large midwestern city. Unlike adult courts, juvenile court proceedings are not open to the public, and juvenile records may not be released. However, after obtaining institutional review board approval, Widom was granted access to these files for research purposes by court authorities. From juvenile court records, she selected 774 cases of neglect, physical abuse, or sexual abuse. Cases of extreme abuse were processed in adult criminal court, where charges were filed against the offender. Criminal court records yielded an additional 134 cases.

As described in Chapter 7, Widom constructed a comparison group of nonabused children through individual matching. Comparison subjects were found from two different sources of agency records. First, abused children who were ages 6–11 at the time of the abuse were matched to comparison subjects by consulting public school records. An abused
child was matched with a comparison child of the same sex, race, and age (within 6 months) who attended the same school. Second, children who were younger than 6 years old at the time of abuse were matched on similar criteria by consulting birth records and selecting a comparison subject born in the same hospital. These two types of public records and matching criteria—attending the same public school and being born in the same hospital—were used in an attempt to control for socioeconomic status. Although such criteria may be viewed with skepticism today, Widom (1989:360) points out that during this time period (1967–71), school busing was not used in the area where her study was conducted, and “elementary schools represented very homogeneous neighborhoods.” Similarly, in the late 1960s, hospitals tended to serve local communities, unlike contemporary medical-industrial complexes that advertise their services far and wide. Widom assumed that children born in the same hospital were more likely to be from similar socioeconomic backgrounds than children born in different hospitals. Though far from perfect, school and birth records enabled Widom to construct approximate matches on socioeconomic status, which illustrates our earlier advice to be creative while being careful. Shadish, Cook, and Campbell (2002:129) describe how school- and hospital-based controls can be used more generally in social research.

Widom’s research purpose was explanatory—to examine the link between early child abuse and later delinquency or adult criminal behavior. These two dependent variables were measured by consulting additional agency records for information on arrests of abused subjects and comparison subjects for the years 1971 through 1986. Juvenile court files yielded information on delinquency. Adult arrests were measured from criminal history files maintained by local, state, and national law enforcement agencies. In an effort to locate as many subjects as possible, Widom searched for current addresses and Social Security numbers in the files of state motor vehicle bureaus. Finally, “marriage license bureau records were searched to find married names for the females” (Widom 1989:361).

Findings revealed modest but statistically significant differences between abused and comparison subjects. As a group, abused subjects were more likely to have records of delinquency or adult arrests (Maxfield and Widom, 1996). However, Widom (1992) also found differences in these dependent variables by race and gender.

Now let’s consider two potential validity and reliability issues that might be raised by Widom’s use of nonpublic agency records. First, data from juvenile and adult criminal courts reveal only cases of abuse that come to the attention of public officials. Unreported and unsubstantiated cases of abuse or neglect are excluded, and this raises a question about the validity of Widom’s measure of the independent variable. Second, dependent variable measures are similarly flawed because juvenile and adult arrests do not reflect all delinquent or criminal behavior.

Recognizing these problems, Widom is careful to point out that her measure of abuse probably reflects only the most severe cases, those that were brought to the attention of public officials. She also notes that cases in her study were processed before officials and the general public had become more aware of the problem of child abuse. Widom (1989:365–66) understood the limits of official records and qualified her conclusions accordingly: “These findings, then, are not generalizable to unreported cases of abuse or neglect. Ours are also the cases in which agencies have intervened, and in which there is little doubt of abuse or neglect. Thus, these findings are confounded with the processing factor.”

Crime “Hot Spots” Law enforcement officials and criminal justice researchers have long been interested in the spatial concentration of crime. A growing number of police departments employ crime analysts to identify “hot
spots”—geographic areas and times of day that signal concentrations of various types of crime (Boba, 2003). Additionally, researchers have recognized that some individual people (and places) are repeatedly victimized and that these repeat victims play a major role in producing hot spots (Pease and Laycock, 1996).

Research and policy interest in hot spots and repeat victimization can be traced to an influential article by Lawrence Sherman and associates (Sherman, Gartin, and Buerger, 1989). Their analysis used a long-neglected measure produced in great volume by police departments. Calls for service (CFS) represent the initial reports of crime and other problems to police departments. Most large police departments record basic information about CFS—location, time, nature of complaint—on audiotapes. Telephone operators respond to most CFS by dispatching a patrol car, a process that is automatically recorded on a computer. This produces a source of data on incidents brought to the attention of police.

Truly astonishing numbers of crime victims, witnesses, or people with some sort of noncrime problem telephone police departments each year. Sherman and associates (1989:36) analyzed more than 300,000 such calls to the Minneapolis Police Department in 1986, proposing that “calls to the police provide the most extensive and faithful account of what the public tells the police about crime, with the specific errors and biases that that entails.” As we pointed out in Chapter 6, UCR and similar data on recorded crime are subject to validity and reliability problems that reflect decisions by police and complainants. CFS are recorded (automatically) before most such decisions are made and so are less subject to screening by police or the public. An additional advantage, and one crucial for Sherman’s research on the “criminology of place,” is that CFS data provide microlevel information on where incidents occur.

Because Sherman and colleagues (1989:37) were interested in the concentration of calls in certain locations, they required some measure of the number of distinct locations in Minneapolis. Consulting records from the city tax assessor’s office, the Administrative Engineering Service, and the Traffic Engineering Office, researchers concluded that it was not possible to obtain a precise count of locations. A denominator of places for computing CFS rates was estimated at 115,000, which included 109,000 street addresses and 6000 intersections.

In a slight digression, we point out that this approximation illustrates another point made by Jacob (1984:39). Given the many sources of potential error in published statistics and agency records, researchers who use such data should report rounded figures and thereby avoid the illusion of exaggerated accuracy that precise numbers imply. For example, there is no way of knowing whether Minneapolis has 107,037 or 111,252 street addresses because of varying definitions used by different agencies. It is therefore safer to state “about 109,000 street addresses” than to report either a specific estimate or a “precise” average of the two estimates (109,144.5), which falsely inflates the accuracy of data.

Addressing first the general question of concentration, Sherman and colleagues (1989:38) found that about 50 percent of calls for service occurred in just 3 percent of Minneapolis locations. Looking closer at types of incidents sheds light on the links between type of offense and type of place. For example, hot spots such as discount department stores and large parking lots produced large numbers of shoplifting reports and calls from motorists locked out of their cars. One hotel stood out with a large number of calls for burglary and violent crime but turned out to be more of a “cool spot” when standardized by the hotel’s average daily population of more than 3000 guests and employees. In contrast, the approximate robbery rate for a single bar that had an estimated daily population of about 300 was 83 per 1000 persons at risk.

With these interesting descriptive analyses in hand, Sherman and associates (1989:47–48)
next addressed the question of crime displacement, arguing that the potential for displacement varies by type of incident. Predatory crimes such as robbery and burglary are at least partly dependent on an opportunity structure that can be modified. Citing measures for reducing convenience store robberies, the researchers proposed that features of such places—physical design and staffing patterns—can deter motivated offenders. In contrast, incidents such as interpersonal violence and vice are less place dependent. Domestic violence can occur any place two people live. Prostitution or drug sales depend on market attractions between seller and buyer. Parties to the transaction are mobile and can easily relocate to a place where surveillance or other inhibiting factors are absent. Thus, there is a greater potential for displacement of crimes linked to intimate or market relationships, whereas prevention may be possible for incidents that are more dependent on characteristics of place.

This example is different from Widom's study in many respects; however, the researchers all recognized the potential promise and shortcomings of data from agency records—in each case, they were careful and creative. Despite the advantages of using CFS data, Sherman and associates describe three possible problems: the potential for duplicate records of the same incident; false reports, equivalent to false fire alarms; and misleading hot spots such as hospitals and police stations that produce secondary crime reports. After taking these problems into consideration, Sherman and colleagues (1989) concluded that the advantages of CFS data outweigh the disadvantages. Readers can form their own conclusions, aided by the careful description of the paper trail—how CFS data are produced—provided by Sherman and colleagues.

**Agency Records as Measures of Decision Making** Two studies mentioned in earlier chapters illustrate how flaws in agency record-keeping practices can be used to infer something about agency behavior. In earlier chapters we mentioned the classic study by Richard Mc Cleary and associates (Mc Cleary, Nienstedt, and Erven, 1982) as an illustration of measurement problems and the ways in which those problems threaten certain quasi-experimental designs. Recall that Mc Cleary and associates discovered that an apparent reduction in burglary rates was, in fact, due to changes in record-keeping practices. Assigning officers from a special unit to investigate burglaries revealed that earlier investigative procedures sometimes resulted in double counts of a single incident; the special unit also reduced misclassification of larcenies as burglaries. Mc Cleary and colleagues became suspicious of police records when they detected an immediate decline in burglary rates following the introduction of the special unit, a pattern that was not reasonable given the technology of burglary. Following the paper trail, they were able to discover how changes in procedures affected these measures. In the same article, they describe additional examples—how replacement of a police chief and changes in patrol dispatch procedures produced apparent increases in crime and calls for service. However, after carefully investigating how these records were produced, they were able to link changes in the indicators to changes in agency behavior rather than changes in the frequency of crime.

Terry Baumer and associates (Baumer, Maxfield, and Mendelsohn, 1993) discovered a similar phenomenon when they detected inconsistencies between court records of juvenile probation violations and other indicators of probation client behavior. (See the box titled “Home Detention” in Chapter 1.) After locating virtually no instances of probation violations in agency records, Baumer and associates found that arrest histories and data from electronic monitoring showed evidence of infractions. This led them to follow the paper trail through juvenile court, where they learned that juvenile probation staff were only minimally supervising their clients. Becoming more suspicious,
the authors uncovered other inconsistencies in program implementation by juvenile court staff. So, data intended to monitor the behavior of juvenile probationers instead showed that probation staff were not doing their jobs.

These two examples point to an important lesson in the use of agency records for criminal justice research: *Expect the expected.* If unexpected findings or patterns emerge, review data collection procedures once again before accepting the unexpected. In these two examples, researchers suspected record-keeping problems when data analysis produced results that were sharply discrepant with their expectations. Changing the way burglaries are investigated is unlikely to produce a pronounced, immediate decline in burglary. And it is implausible that an experimental program for juvenile offenders will be 99 percent successful. Such clues prompted researchers to inquire further how such suspicious indicators were produced.

**New Data Collected by Agency Staff**

Thus far, we have concentrated on the research uses of information routinely collected by or for public agencies. Such data are readily available, but researchers have little control over the actual data collection process. Furthermore, agency procedures and definitions may not correspond with the needs of researchers.

It is sometimes possible to use a hybrid source of data in which criminal justice agency staff collect information for specific research purposes. We refer to this as a “hybrid” source because it combines the collection of new data—through observation or interviews—with day-to-day criminal justice agency activities. Virtually all criminal justice organizations routinely document their actions, from investigating crime reports to housing convicted felons. By slightly modifying forms normally used for recording information, researchers may be able to have agency staff collect original data for them.

For example, let’s say we are interested in the general question of how many crimes reported to police involve nonresident victims. Reading about violent crime against international tourists in south Florida, we wonder how common such incidents are and decide to systematically investigate the problem. It doesn’t take us long to learn that no published data are available on the resident or nonresident status of Dade County crime victims. We next try the Miami and Miami-Dade Police Departments, suspecting that such information might be recorded on crime report forms. No luck here, either. Incident report forms include victim name and address, but we are told that police routinely record the local address for tourists, typically a hotel, motel, or guest house. Staff in the police crime records office inform us that officers sometimes write down something like “tourist, resident of Montreal” in the comments section of the crime report form, but they are neither required nor asked to do this.

Assuming we can gain approval to conduct our study from police and other relevant parties, we may be able to supplement the standard police report form. Adding an entry such as the following will do the trick:

Is complainant a resident of Dade County?  
[ ] yes  [ ] no

If “no,” record permanent address here:

____________________________________

A seemingly simple modification of crime report forms may not happen quite so easily. Approval from the department is, of course, one of the first requirements. And recall our discussion in Chapter 10 on the need to gain approval for research from criminal justice agency staff at all levels. Individual officers who complete crime report forms must be made aware of the change and told why the new item was added. We might distribute a memorandum that explains the reasons for adopting a new crime report form. It might also be a good idea to have supervisors describe the new form at roll call before each shift of officers heads out on patrol. Finally, we could review samples of the new forms over the first few days they are used to determine the extent to which officers are completing the new item.
Chapter 11  Agency Records, Content Analysis, and Secondary Data

Incorporating new data collection procedures into agency routine has two major advantages. First, and most obvious, having agency staff collect data for us is much less costly than fielding a team of research assistants. It is difficult to imagine how original data on the resident status of Dade County victims could be collected in any other way. Second, we have more control over the measurement process than we would by relying on agency definitions. Some Dade County officers might note information on victim residence, but most probably would not. Adding a specific question enhances the reliability of data collection. We might consider using an existing crime report item for “victim address,” but the tendency of officers to record local addresses for tourists would undermine measurement validity. A specific “resident/non-resident” item is a more valid indicator.

The box by Marie Mele, titled “Improving Police Records of Domestic Violence,” gives an example of this approach to gathering new data from agency records. You should recognize the importance of collaboration in this example. Less obvious, but equally important, is a lesson we will return to later in this chapter: Agency records are not usually intended for research purposes and, as a consequence, are not always well suited to researchers’ needs. Sometimes, as was the case for Marie Mele, some subtle tweaking can transform unusable mounds of paper files into computer-based record systems. In any event, researchers are well advised to be careful, not assuming that existing data will meet their needs, but to also be creative in seeking ways to improve the quality of agency data.

This approach to data collection has many potential applications. Probation officers or other court staff in many jurisdictions complete some type of presentence investigation on convicted offenders. A researcher might be able to supplement standard interview forms with additional items appropriate for some specific research interest. In their experimental study of intensive probation, Joan Petersilia and Susan Turner (1991) obtained the cooperation of probation staff to complete three data collection forms on each research subject. Intake forms yielded demographic and criminal history information. Review forms completed after 6 and 12 months documented the nature and types of services that probation staff delivered to experimental and control subjects (1991:621). Additional agency records provided data on probationer performance, but the supplementary data were needed to measure implementation of the intensive probation program.

The National Institute of Justice (NIJ) (1999) integrated new data collection into routine arrest procedures with its Arrestee Drug Abuse Monitoring (ADAM) system. As we saw in Chapter 6, samples of persons arrested in each participating city were selected four times each year and asked to voluntarily submit a urine sample for anonymous testing. Results were tabulated to estimate the prevalence of drug use among people arrested for different types of offenses.

The NIJ incrementally expanded the use of ADAM as a “research platform,” as illustrated in a study by Scott Decker and associates (Decker, Pennell, and Caldwell, 1997), which supplemented the interview questionnaire in 11 ADAM sites with questions about firearm availability and use. The researchers found no association between drug use and firearm use, but they did discover that gun use was common among juvenile males and especially widespread among admitted gang members. By piggybacking on ADAM in this way, Decker and associates were able to obtain information from more than 7000 subjects at very low cost, which illustrates the principal strength of enlisting agency staff in data collection efforts.

At the same time, having agencies collect original research data has some disadvantages. An obvious one is the need to obtain the cooperation of organizations and staff. The difficulty of this varies in direct proportion to the intrusiveness of data collection. Cooperation is less likely if some major additional effort is required of agency personnel or if data collection activities disrupt routine operations.
The potential benefit to participating agencies is a related point. If a research project or an experimental program is likely to economize agency operations or improve staff performance, as was the case for Marie Mele’s research, it will be easier to enlist their assistance.

Researchers have less control over the data collection process when they rely on agency staff. Petersilia (1989:442) points out that agency personnel have competing demands on their time and naturally place a lower priority on data collection than on their primary duties. If you were a probation officer serving a heavy caseload and were asked to complete detailed 6- and 12-month reports on services provided to individual clients, would you devote more attention to keeping up with your clients or filling out data collection forms?

**Units of Analysis and Sampling**

*Researchers must be especially attentive to the units of agency records, particularly in selecting samples for analysis.*

After determining that agency records will be suitable for some particular research purpose, several decisions and tasks remain. We will consider two of them briefly because each is covered in more detail elsewhere in the book: units of analysis and sampling.
few kinks, the final system was established, and I continued entering new incident reports for a period of time.

This is an admittedly brief description of a process that unfolded over several months. The process began as an attempt to obtain what we thought were existing data. Finding that suitable data were not available, Maxfield had sufficient access to the department to begin discussions of how to supplement existing record-keeping practices to tabulate repeat victimization. Many public agencies will accommodate researchers if such accommodation does not unduly burden the agency. Collecting new data for researchers does qualify as unduly burdensome for most public organizations. The key here was collaboration in a way that helped both the researchers and the host organization.

It’s obvious that my research benefited from having the department design and ultimately assume responsibility for tabulating repeat victimization. The department benefited in three ways. First, DVSAU staff recognized how a data file that identified repeat offenders and victims could produce information that would aid their investigations of incidents. I emphasize offenders here because, all other things being equal, police are more interested in offenders than they are in victims. So the database was designed to track people (offenders) of special interest to police and people (victims) of special interest to my research. Second, recognizing that setting up a new data system is especially difficult in a tradition-bound organization, I was able to ease the transition somewhat by initially entering data myself. One intentional by-product of this was to establish quality-control procedures during a shake-down period. As a criminal justice researcher, I knew that the reliability of data-gathering procedures was important and was able to establish reliable data entry routines for the department.

The DVSAU staff benefited in yet another way that was probably most important of all. This department had incorporated its version of Compstat for about 3 years. Two important components of Compstat are timely data and accountability. Each week, the DVSAU commander was required to present a summary of the unit’s activity and was held accountable for the performance of unit staff. Before the database was developed, the DVSAU commander spent several hours compiling data to prepare for each week’s Compstat meeting. After the database was operational, preparation time was reduced to minutes. In addition, the DVSAU commander was able to introduce new performance measures—repeat offenders and victims—that came to be valued by the department’s chief.

Units of Analysis

Archives and agency records may be based on units of analysis that are not suitable for particular research questions. If we are interested in studying individual probationers, for example, we need individual-level data about persons sentenced to probation. Summary data on the number of probationers served each week might not meet our research needs. Or if we wish to examine whether probationers convicted of drug offenses are supervised more closely than those convicted of assault, data on individuals sentenced to probation will have to be aggregated into categories based on convicted offense.

A general rule we mentioned in Chapter 4 bears repeating here: It is possible to move from individual to aggregate units of analysis, but not the other way around. Thus, we could aggregate records on individual probationers into groups that reflected convicted offense, but we could not disaggregate weekly reports to produce information about individuals. In any case, researchers using agency records must be attentive to the match or mismatch between the units of analysis required to address specific research questions and the level of aggregation represented in agency records.

Units of analysis can be especially troublesome in studies of criminal justice processes or studies of people moving through some institutional process. This is because criminal
justice agencies use different units of count in keeping records of their activities. Figure 11.1, adapted from a report prepared by officials in a New York criminal justice agency (Poklemba, 1988), lists many of the counting units recorded at various stages of processing.

The units listed in Figure 11.1 can be grouped into two different categories: counts of events and counts of cases (Poklemba, 1988: III3). Counts of events—such as an arrest, indictment, or admission—are more straightforward because they are of short duration. Cases, however, may persist over longer times that are bounded by initiating or terminating events. Cases are directly linked to individual persons, though in complex ways. For example, an indictment event begins a court case that does not terminate until a court disposition event. Or a prison admission event begins an inmate case that ends at prison discharge event. Further complicating matters is the multitude of possible relationships between units of count. For example, a court case can include multiple defendants, each facing multiple counts that can produce multiple dispositions.

The best solution for many research purposes is to define some equivalent to a person—a defendant, for example—as the unit of analysis. This was the approach used by James Eisenstein and Herbert Jacob (1977:175–76) in their study of felony courts:

Indictments and cases are full of definitional ambiguities that vary from city to city. Some defendants are named in multiple indictments whereas others are not; many defendants are washed out of the process before being indicted but after receiving some punishment. Cases may involve a single defendant or many, and tend to be linked together if there are overlapping defendants or indictments. The concept of “defendants” suffers from none of these ambiguities. Using defendants as our unit of analysis permits us to discern the number of indictments each defendant faced, the number of court cases in which each was involved, and the ultimate [disposition] each [defendant] faced.

Defining individual people as units can resolve the conceptual problems that emerge from complex relationships between different units of count. Practical difficulties may remain in linking individuals to other units of count, however, or in tracing the movement of individuals from one institution to another.

**Sampling**

It is often necessary to select subsets of agency records for particular research purposes. Just as we would not need to interview every resident of New York City to learn how they feel about subway crime, it may not be necessary to examine all court cases to understand patterns of case disposition or sentences.

You will be glad to hear that once units of analysis are defined sampling agency records is relatively simple. In most cases, a target population and sample frame may be readily defined. If, for example, we want to study the disposition of felony arrests in New York, our target population might be all cases that reached final disposition in the year 2006. We could then get a list or computer file that contains identifying numbers for all 2006 felony cases and draw a sample using systematic or other sampling procedures described in Chapter 8. We have to be alert for potential biases in the sample frame,
Reliability and Validity

Understanding the details of how agency records are produced is the best guard against reliability and validity problems.

The key to evaluating the reliability and validity of agency records, as well as the general suitability of those data for a research project, is to understand as completely as possible how the data were originally collected. Doing so can help researchers identify potential new uses of data. They will also be better able to anticipate and detect potential reliability or validity problems in agency records.

Any researcher who considers using agency records will benefit from a careful reading of Jacob’s invaluable little guide Using Published Data: Errors and Remedies (1984). In addition to warning readers about general problems with reliability and validity, such as those we discussed in Chapter 5, Jacob cautions users of these data to be aware of other potential errors that can be revealed by scrutinizing source notes. Clerical errors, for example, are unavoidable in such large-scale reporting systems as the UCR. These errors may be detected and reported in correction notices appended to later reports.

Users of data series collected over time must be especially attentive to changes in data collection procedures or changes in the operational definitions of key indicators. As you might expect, such changes are more likely to occur in data series that extend over several years. David Cantor and James Lynch (2005) describe how changes in the NCVS, especially the redesign elements introduced in 1992, should be kept in mind for any long-term analysis of NCVS data. In earlier chapters, we described the NCVS redesign that was completed in 1994. If we plan to conduct research on victimization over time, we will have to consider how changes in sample size and design, increased use of telephone interviews, and questionnaire revisions might affect our findings.

Longitudinal researchers must therefore diligently search for modifications of procedures or definitions over time in order to avoid attributing some substantive meaning to a change in a particular measure. Furthermore, as the time interval under investigation increases, so does the potential for change in measurement. Ted Robert Gurr (1989:24) cites a good example:

In the first two decades of the 20th century many American police forces treated the fatalities of the auto age as homicides. The sharp increase in “homicide” rates that followed has led to some dubious conclusions. Careful study of the sources and their historical and institutional context is necessary to identify and screen out the potentially misleading effects of these factors on long-term trends.

This example suggests that cross-sectional researchers must be alert for a slightly different type of potential error. Roger Lane (1997) points out that the tendency to classify fatal accidents as homicides was greater in some cities than in others.

The central point here is that researchers who analyze criminal justice data produced by different cities or states or other jurisdictions must be alert to variations in the definitions and measurement of key variables. Even in cases in which definitions and measurement seem straightforward, they may run into problems. For example, Craig Perkins and Darrell Gilliard (1992:4)—statisticians at the Bureau of Justice Statistics—caution potential users of corrections data:

Care should be exercised when comparing groups of inmates on sentence length and time served. Differences may be the result of factors not described in the tables, including variations in the criminal histories of each group, variations in the offense composition of each group, and variations...
among participating jurisdictions in their sentencing and correctional practices.

Fortunately, most published reports on regular data series present basic information on definitions and collection procedures. Many BJS publications include copies of the questionnaires used in surveys and enumerations. Researchers should, however, view summary descriptions in printed reports as no more than a starting point in their search for information on how data were collected. Before analyzing published data in earnest, they should contact the issuing organization to obtain details, perhaps in the form of technical reports.

Sources of Reliability and Validity Problems

We conclude this section on agency records by discussing some general characteristics of record keeping by public agencies. Think carefully about each of the features we mention, considering how each might apply to specific types of criminal justice research. It will also be extremely useful for you to think of some examples in addition to those we mention, perhaps discussing them with your instructor or others in your class.

Social Production of Data Virtually all criminal justice record keeping is a social process. By this we mean that indicators of, say, arrests, juvenile probation violations, court convictions, or rule infractions by prison inmates reflect decisions made by criminal justice officials in addition to the actual behavior of juvenile or adult offenders. As Baumer and associates state: “Researchers must realize that performance measures are composites of offenders’ behavior, organizational capacity to detect behavior, and decisions about how to respond to offenders’ misbehavior” (1993:139; emphasis added). A small number of classic articles illustrate the social production of crime records by police (Black, 1970; Kitsuse and Cicourel, 1963; Seidman and Couzens, 1974). Richard McCleary (1992) describes the social production of data by parole officers. They may fail to record minor infractions to avoid paperwork or, alternatively, may keep careful records of such incidents in an effort to punish troublesome parolees by returning them to prison. More generally, Coleman and Moynihan (1996) describe the social production of many criminal justice measures.

Discretionary actions by criminal justice officials and others affect the production of virtually all agency records. Police neither learn about all crimes nor arrest all offenders who come to their attention. Similarly, prosecutors, probation officers, and corrections staff are selectively attentive to charges filed or to rule violations by probationers and inmates. At a more general level, the degree of attention state legislatures and criminal justice officials devote to various crime problems varies over time. Levels of tolerance of such behaviors as child abuse, drug use, acquaintance rape, and even alcohol consumption have changed over the years.

Agency Data Are Not Designed for Research In many cases, criminal justice officials collect data because the law requires them to do so. More generally, agencies tend to collect data for their own use, not for the use of researchers. Court disposition records are maintained in part because of legal mandates, and such records are designed for the use of judges, prosecutors, and other officials. Record-keeping procedures reflect internal needs and directives from higher authorities. As a consequence, researchers sometimes find it difficult to adapt agency records for a specific research purpose.

For example, Maxfield once wished to trace court dispositions for arrests made by individual police officers in Louisville, Kentucky. “No problem,” he was assured by a deputy prosecutor, “we keep all disposition records on the computer.” Following this electronic version of a paper trail to the county data-processing facility, Maxfield discovered that only the previ-
ous year’s cases were maintained on computer tape, the most common bulk storage medium in the 1970s. Such tapes were said to be expensive (about $17), and nobody had authorized the data processing staff to buy new tapes for each year’s files. Instead, voluminous computer printouts from the previous year were microfilmed and saved, whereas the “costly” computer tape was erased for the new year. Maxfield abandoned the project after realizing that data collection would involve viewing literally hundreds of thousands of microfilmed case files, instead of running a quick-and-easy computer search.

Keep in mind that our research needs may not be congruent with agency record-keeping practices. Courts or police departments commonly use idiosyncratic definitions or methods of classifying information that make such records difficult to use. Also recognize that our conceptual and operational definitions of key concepts, however thoughtful and precise, will seldom be identical to actual measures maintained by criminal justice agencies.

Even when agencies have advanced record-keeping and data management systems, researchers may still encounter problems. The Chicago Police Department, for example, has developed one of the most advanced systems for collecting and analyzing crime and calls for service data. Richard and Carolyn Block (1995) sought to take advantage of this in an analysis relating the density of taverns and liquor stores to police crime reports from taverns. The researchers wanted to learn whether areas with high concentrations of taverns and liquor stores generated a disproportionate number of crime reports on which police had indicated “tavern or liquor store.” Because the Chicago Police Department recorded addresses and type of location, and the city Department of Revenue recorded addresses for liquor license holders, the task seemed straightforward enough: Match the addresses from the two data sources, and analyze the correspondence between crime and liquor establishments.

It wasn’t so simple. The Blocks discovered inconsistent recording of addresses in crime reports. Sometimes, police approximated a location by recording the nearest intersection—for example, “Clark and Division.” Some large establishments spanned several street addresses, and police recorded one address when the liquor license data had a different address. Other times, police recorded the name of the tavern—for example, “Red Rooster on Wilson Avenue.” In sum, police recorded addresses to meet their needs—to locate a tavern or liquor store so that officers could find it—whereas the Department of Revenue recorded a precise address on a license application. Each address was accurate for each agency’s purpose, but addresses did not match in about 40 percent of the cases analyzed by the Blocks.

For another example of how definitional differences can be traced to different agency needs, see the box titled “How Many Parole Violators Were There Last Month?”

**Tracking People, Not Patterns** At the operational level, officials in criminal justice organizations are generally more interested in keeping track of individual cases than in examining patterns. Police patrol officers and investigators deal with individual calls for service, arrests, or case files. Prosecutors and judges are most attentive to court dockets and the clearing of individual cases; corrections officials maintain records on individual inmates. Although each organization produces summary reports on weekly, monthly, or annual activity, officials tend to be much more interested in individual cases. Michael Geerken (1994) makes this point clearly in his discussion of problems researchers are likely to encounter in analyzing police arrest records. Few rap sheet databases are regularly reviewed for accuracy; rather, they simply accumulate arrest records submitted by individual officers. Joel Best (2001) offers additional examples of how small errors in case-by-case record keeping can accumulate to produce compound errors in summary data.
With continued advances in computer and telecommunications technology, more criminal justice agencies are developing the capability to analyze data in addition to tracing individual cases. Crime analysis by police tracks spatial patterns of recent incidents; prosecutors are attentive to their scorecards; state correctional intake facilities consider prison capacity, security classification, and program availability in deciding where to send new admissions.

As problem-oriented approaches to policing become more widely adopted, many law enforcement agencies have improved their record-keeping and crime analysis practices. New York City’s reduction in reported crime has been attributed to the use of timely, accurate crime data by police managers to plan and evaluate specific anticrime tactics (Bratton, 1999; Kelling and Coles, 1996; Maple, 1999). This illustrates an important general principle about the accuracy of agency-produced data: When agency managers routinely use data to make decisions, they will be more attentive to data quality. The box describing Marie Mele’s research offers another example—when domestic violence detectives realized how a database would help them prepare for Compstat, they endorsed the effort to improve their record-keeping procedures.

However, record-keeping systems used in most cities today are still designed more to track individual cases for individual departments than to produce data for management or research purposes. Individual agencies maintain what are sometimes called “silo databases,” a colorful label that refers to stacks of data that are isolated from each other.

This does not mean that computerized criminal history, court disposition, or prison intake records are of little use to researchers. However, researchers must be aware that even

### How Many Parole Violators Were There Last Month?

by John J. Poklemba  
New York State Division of Criminal Justice Services

**Question:** How many parole violators were there last month? **Answer:** It depends. More accurately, it depends on which agency is asked. Each of the three answers below is right in its own way:

- New York State Commission of Correction: 611
- New York Department of Correctional Services: 670
- New York Division of Parole: 356

The State Commission of Correction (SCOC) maintains daily aggregate information on the local under-custody population. Data are gathered from local sheriffs, using a set of common definitions. The SCOC defines a parole violator as follows: an alleged parole violator being held as a result of allegedly having violated a condition of parole—for example, a new arrest. This makes sense for local jails; a special category is devoted to counting alleged parole violators with new arrests. However, New York City does not distinguish between parole violators with and without new arrests, so the SCOC figure includes violators from upstate New York only.

The Department of Correctional Services (DOCS) is less interested in why people are in jail; their concern centers on the backlog of inmates whom they will soon need to accommodate. Furthermore, as far as DOCS is concerned, the only true parole violator is a technical parole violator. This makes sense for DOCS, because a parole violator convicted of a new crime will enter DOCS as a new admission, who—from an administrative standpoint—will be treated differently than a parolee returned to prison for a technical violation.

The Division of Parole classifies parole violators into one of four categories: (1) those who...
have violated a condition of parole, (2) those who have absconded, (3) those who have been arrested for a new crime, and (4) those who have been convicted of a new crime. Once again, this makes sense, because the Division of Parole is responsible for monitoring parolee performance and wishes to distinguish different types of parole violations. The Division also classifies a parole violation as either alleged (yet to be confirmed by a parole board) or actual (the violation has been confirmed and entered into the parolee’s file). Further differences in the fluid status of parolees and their violations, together with differences between New York City and other areas, add to the confusion.

Taking the varying perspectives and roles of these three organizations into account, answers to the “How many” question can be made more specific:

**SCOC:** Last month, there were 611 alleged parole violators who were believed to have violated a condition of their parole by being arrested for a new offense and are being held in upstate New York jails.

**DOCS:** Last month, there were 670 actual parole violators who were judged to have violated a condition of their parole and are counted among the backlog of persons ready for admission to state correctional facilities.

**Parole Division:** Last month, 356 parolees from the Division’s aggregate population were actually removed from the Division’s case-load and were en route to DOCS.

One of the major reasons that agency counts do not match is that agency information systems have been developed to meet internal operational needs. A systemwide perspective is lacking. Questions that depend on data from more than one agency are often impossible to answer with confidence. Recognize also that the availability and quality of state data depend on data from local agencies.

As stated above, the best answer to the question is: It depends.

**Source:** Adapted from Poklemba (1988:11-13).

state-of-the-art systems may not be readily adaptable for research purposes. The Louisville system described earlier was fine for the needs of court staff who might have to check on some individual past case, but it was virtually unusable for Maxfield’s research, which required data about patterns of dispositions.

**Error Increases with Volume**  The potential for clerical errors increases as the number of clerical entries increases. This seemingly obvious point is nonetheless important to keep in mind when analyzing criminal justice records. Lawrence Sherman and Ellen Cohn (1989:34) describe the “mirror effect” of duplicate CFS records. Handling a large volume of CFS, phone operators in Minneapolis, or any large city for that matter, are not always able to distinguish duplicate reports of the same incident. An updated report about a CFS may be treated as a new incident. In either case, duplicate data result.

The relationship between volume of data entry and the potential for error can be especially troublesome for studies of relatively rare crimes or incidents. Although murder is rare compared with other crimes, information about individual homicides might be keyed into a computer by the same clerk who inputs data on parking violations. If a murder record is just one case among hundreds of parking tickets and petty thefts awaiting a clerk, there is no guarantee that the rare event will be treated any differently than the everyday ones.

While preparing a briefing for an Indianapolis commission on violence, Maxfield discovered that a single incident in which four people had been murdered in a rural area appeared twice in computerized FBI homicide records. This was traced to the fact that officers
from two agencies—sheriff’s deputies and state police—investigated the crime, and each agency filed a report with the FBI. But the thousands of murders entered into FBI computer files for that year obscured the fact that duplicate records had been keyed in for one multiple murder in a rural area of Indiana. Colin Loftin (1986) discusses different types of error in murder records. For all their advantages, computers have the capacity to magnify clerical errors with a vengeance.

In concluding our lengthy discussion of agency records, we do not mean to leave you with the impression that data produced by and for criminal justice organizations are fatally flawed. Thousands of studies making appropriate use of such data are produced each year. It is, however, essential that researchers understand potential sources of reliability and validity problems, as well as ways they can be overcome. Public agencies do not normally collect information for research purposes. The data they do collect often reflect discretionary decisions by numerous individuals. And, like any large-scale human activity, making observations on large numbers of people and processes inevitably produces some error.

Content Analysis

Content analysis involves the systematic study of messages.

The Office of Community Oriented Policing Services (COPS) was established by the 1994 Crime Bill to promote community policing by providing funds to local law enforcement agencies. In addition to being concerned about the effectiveness of these efforts, COPS staff wanted to know something about the public image of community policing as presented in local newspapers. Stephen Mastrofski and Richard Ritti (1999) conducted a content analysis of stories about community policing in newspapers serving 26 cities. The researchers found over 7500 stories from 1993 through 1997, with most focusing on a small number of themes: “community, resources, and producing real results for the community. Stories that offer a viewpoint on community policing are nearly always overwhelmingly positive” (1999:10–11).

This is an example of content analysis, the systematic study of messages and the meaning those messages convey. For the COPS office, the study by Mastrofski and Ritti was satisfying—many stories about community policing were published in urban newspapers, and most stories presented positive images.

Content analysis methods may be applied to virtually any form of communication. Among the possible artifacts for study are books, magazines, videos, songs, speeches, television programs, letters, laws, and constitutions, as well as any components or collections of these. Suppose you’re interested in violence on television. Maybe you suspect that the manufacturers of consumer products for men are more likely to sponsor violent TV shows than are other kinds of sponsors. Content analysis would be the best way of finding out if that’s true.

Briefly, here’s how you might go about conducting that study. First, you develop operational definitions of the two key variables in your inquiry: men’s products and violence. Our later discussion of coding will explore some of the ways you can do that. Ultimately, you need a plan that allows you to watch television, classify sponsors, and rate the degree of violence on particular shows. Next, you have to decide (1) what stations to watch, (2) for what days or period, and (3) at what hours. Then you stock up on snacks and start watching, classifying, and recording. Once you have completed your observations, you can analyze the data you collected and determine whether men’s product manufacturers sponsor more blood and gore than other sponsors.

Content analysis, then, is particularly well suited to the study of communications and to answering the classic question of communications research: Who says what, to whom, why, how, and with what effect? As a mode of observation, content analysis requires a considered handling of the what, and the analysis of data
collected in this mode, as in others, addresses the *why* and *with what effect*.

**Units of Analysis and Sampling in Content Analysis**

In the study of communications, as in the study of people, it is usually impossible for researchers to directly observe all that interests them. In your study of television violence and sponsorship, you are well advised not to try to watch everything that’s broadcast. It’s not possible, and your brain will probably short-circuit before you get close to discovering that for yourself. Usually, then, it’s appropriate to sample the units on which content analysis will be conducted.

As we mentioned with respect to agency records in general, determining appropriate units of analysis can be a complicated task. For example, we may wish to compare crime rates of different cities in terms of their size, geographic region, racial composition, and other differences. Even though the characteristics of these cities are partly a function of the behaviors and characteristics of their individual residents, the cities will ultimately be the units of analysis.

The complexity of this issue is often more apparent in content analysis than in other research methods. That is especially the case when the units of observation differ from the units of analysis. A few examples should clarify this distinction.

Suppose we want to find out whether criminal law or civil law makes the most distinctions between juveniles and adults. In this instance, individual laws are both the units of observation and the units of analysis. We might select a sample of a state’s criminal and civil laws and then categorize each law by whether it makes a distinction between juveniles and adults. In this fashion, we could determine whether criminal or civil law distinguishes more by age.

Now, let’s look at a trickier example: the study of television violence and sponsors. What is the unit of analysis for the research question “Are manufacturers of men’s products more likely to sponsor violent shows than are other sponsors?” Is it the TV show? The sponsor? The instance of violence? In the simplest study design, it is none of these.

Although you might structure your inquiry in various ways, the most straightforward design is based on the commercial as the unit of analysis. Here, you use two kinds of observational units: the commercial and the program that gets squeezed in between commercials. You want to observe both units. You classify the commercials by whether they advertise men’s products and the programs by their violence. The program classifications are transferred to the commercials that are near them. Figure 11.2 is an example of the kind of record you might keep.

Notice that in the research design illustrated in the figure, all the commercials that appear together are bracketed and get the same scores. Also, the number of violent instances recorded as following one commercial is the same as the number preceding the next commercial. This simple design allows us to classify each commercial by its sponsorship and the degree of violence associated with it. Thus, for example, the first Grunt aftershave commercial is coded as being a men’s product and as having 10 instances of violence associated with it. The Precious perfume commercial is coded as not being a men’s product and as having no violent instances associated with it.

In Figure 11.2, we have four men’s product commercials with, on average, 7.5 violent instances associated with each. The four commercials classified as definitely not men’s products have an average of 1.75, and the two that might or might not be considered men’s products have an average of 1.0. If this pattern of differences persists across a much larger number of observations, we will probably conclude that manufacturers of men’s products are more likely to sponsor TV violence than are other sponsors.

The point of this illustration is to demonstrate how units of analysis figure into data collection and analysis. You need to be clear about your unit of analysis before planning...
your sampling strategy, but in this case, you can’t sample commercials. Unless you have access to the stations’ broadcasting logs, you won’t know when the commercials are going to run. Moreover, you need to observe the programming as well as the commercials. As a result, you must set up a sampling design that includes everything you need to observe.

In designing the sample, you need to establish the universe to be sampled. In this case, what TV stations will you observe? What will be the period of the study—the number of days? And what hours of each day will you observe? Also, how many commercials do you want to observe and code for analysis? Watch television for a while and find out how many commercials occur each hour; then you can figure out how many hours of observation you’ll need.

Now you’re ready to design the sample selection. As a practical matter, you wouldn’t have to sample among the different stations if you had assistants; each of you could watch a different channel during the same time period. But let’s suppose you are working alone. Your final sampling frame, from which a sample will be selected and watched, might look something like this:

Jan. 7, Channel 2, 7–9 P.M.
Jan. 7, Channel 4, 7–9 P.M.
Jan. 7, Channel 9, 7–9 P.M.
Jan. 7, Channel 2, 9–11 P.M.

### Figure 11.2 Example of Recording Sheet for TV Violence

<table>
<thead>
<tr>
<th>Sponsor</th>
<th>Men’s Product?</th>
<th>Number of Instances of Violence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Grunt aftershave</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Reef-buster jet skis</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Roperoot cigars</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Grunt aftershave</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Snowflake toothpaste</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Godliness cleanser</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Micro-zap computers</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Snowflake toothpaste</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Micro-zap computers</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Precious perfume</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
Notice that three decisions have been made for you in the illustration. First, it is assumed that Channels 2, 4, and 9 are the ones appropriate to your study. Second, it is assumed that the 7–11 P.M. prime-time hours are the most relevant and that 2-hour periods will do the job. January 7 was picked out of the hat for a starting date. In practice, of course, these decisions should be based on your careful consideration of what is appropriate to your particular study.

Once you are clear about your units of analysis and the observations appropriate to those units, and have created a sampling frame like the one illustrated, sampling is simple and straightforward. The alternative procedures available to you are the same ones described in Chapter 8: random, systematic, stratified, and so on.

If you want to analyze crime stories in major metropolitan newspapers, you might begin with a list of cities and then draw a systematic sample from that list. For cities with more than one newspaper, you will have to develop procedures for sampling papers within each city. Or suppose you are interested in studying how popular fiction depicts male and female detectives. Assuming you do not wish to read all published detective stories, you might consider stratifying by author’s gender and drawing samples of books by male and female authors. As another example, how might you sample movie videos for an analysis of violent acts by and against police officers? By thinking through these examples, you will gain a better understanding of sampling in general, as well as of the ways sampling methods are used in content analysis.

### Coding in Content Analysis

Content analysis is essentially a coding operation, and, of course, coding represents the measurement process in content analysis. Communications—oral, written, or other—are coded or classified according to some conceptual framework. Thus, for example, newspaper editorials might be coded as liberal or conservative. Radio talk shows might be coded as bombastic or not. Novels might be coded as detective fiction or not. Political speeches might be coded as containing unsupported rhetoric about crime or not. Recall that terms such as these are subject to many interpretations, and the researcher must specify definitions clearly.

Coding in content analysis involves the logic of conceptualization and operationalization we considered in Chapter 5. In content analysis, as in other research methods, researchers must refine their conceptual framework and develop specific methods for observing in relation to that framework.

For all research methods, conceptualization and operationalization typically involve the interaction of theoretical concerns and empirical observations. If, for example, you believe some newspaper editorials support liberal crime policies and others support conservative ones, ask yourself why you think so. Read some editorials, asking which ones are liberal and which ones are conservative. Is the political orientation of a particular editorial most clearly indicated by its manifest content or by its overall tone? Is your decision based on the use of certain terms (for example, moral decay or need for rehabilitation) or on the support or opposition given to a particular issue, such as mandatory prison sentences versus treatment programs for drug users?

As in other decisions relating to measurement, the researcher faces a fundamental choice between depth and specificity of
understanding. The survey researcher must decide whether specific closed-ended questions or more general open-ended questions will better suit her or his needs. By the same token, the content analyst has a choice between searching for manifest or for latent content. Coding the **manifest content**—the visible, surface content—of a communication more closely approximates the use of closed-ended items in a survey questionnaire. An example of manifest content is whether or not a police officer is quoted in a newspaper story.

Alternatively, coding the **latent content** of the communication—its underlying meaning—is an option. In the most general sense, manifest and latent content can be distinguished by the degree of interpretation required in measurement. Whether police officers are presented in a favorable or unfavorable manner is an example of latent content.

Throughout the process of conceptualizing manifest- and latent-content coding procedures, remember that the operational definition of any variable is composed of the attributes included in it. Such attributes, moreover, should be mutually exclusive and exhaustive. A newspaper editorial, for example, should not be described as both liberal and conservative, although we should probably allow for some to be middle-of-the-road. It may be sufficient to code TV programs as being violent or not violent, but it is also possible that some programs could be antiviolence.

No coding scheme should be used in content analysis unless it has been carefully pretested. We must decide what manifest or latent contents of communications will be regarded as indicators of the different attributes that make up our research variables, write down these operational definitions, and use them in the actual coding of several units of observation. If we plan to use more than one coder in the final project, each of them should independently code the same set of observations so that we can determine the extent of agreement. In any event, we’ll want to take special note of any difficult cases—observations that were not easily classified using the operational definition. Finally, we should review the overall results of the pretest to ensure that they are appropriate to our analytic concerns. If, for example, all of the pretest newspaper editorials have been coded as liberal, we should certainly reconsider our definition of that attribute.

Before beginning to code newspapers, crime dramas on TV, or detective fiction, we need to make plans to assess the reliability of coding. Fortunately, reliability in content analysis can readily be tested, if not guaranteed, in two related ways. First, inter-rater reliability can be determined by having two different people code the same message and then computing the proportion of items coded the same. For example, if 20 attributes of newspaper stories about crime are being coded and two coders score 18 attributes identically, their reliability is 90 percent. The second way to assess coding reliability is the test–retest method, in which one person codes the same message twice. Of course, some time should elapse between the two coding operations. Test–retest procedures can be used when only one person is doing the coding; reliability can be computed in the same way as if the inter-rater method were being used.

**Illustrations of Content Analysis**

We now turn to examples of content analysis in action. The first illustration uses content analysis to measure and characterize newspaper crime stories. The second describes content analysis of a different sort of message—video games. The third demonstrates how extracting information from police records is a form of content analysis.

**Newspaper Stories About Crime** Steven Chermak (1998) conducted a content analysis of crime stories in newspapers from six different cities. Crime stories were the units of analysis; Chermak sampled all crime stories every fifth day in the first 6 months of 1990, for a total of 1557 stories.
Chermak’s principal hypothesis was that the content of articles determines how much space they are allotted and how prominently they are placed in the newspaper. Prominence was assessed by how many inches of coverage were devoted to each story, where stories were placed, and how big the headlines were. Information on the type of offense, number of crimes mentioned in the story, weapons usage, location of offense, and offender and victim characteristics was also collected. These are examples of manifest content—news story characteristics that require little or no interpretation and can be reliably coded.

Other features of crime news stories are more difficult to code, largely because they refer to latent content and require more interpretation by coders. For example, in connection with a larger research project on reactions to crime, Margaret Gordon and associates (Gordon, Reis, and Tyler, 1979) conducted a content analysis of crime stories in nine metropolitan newspapers serving three cities. They examined all stories in each newspaper issue from November 1977 through April 1978, coding information from more than 11,000 stories concerning violent crime. Gordon and associates coded extensive details about crime news coverage in each story, including latent content. For example, each story was classified with respect to bias or slant in its overall tone. Instructions to coders for this item read as follows: “The categories for this variable represent a continuum: objective—biasing word or two—somewhat sympathetic—biased, slanted. ‘Toward victim’ [is] construed to include ‘against suspect,’ and ‘toward suspect’ includes ‘against victim’” (1979:Appendix E).

As you might expect, it was much more difficult for coders to interpret the bias or slant of a story than it was for them to determine, for example, whether the story was a news article, feature, or editorial. Evaluation of police action also proved difficult to code because a single story could be critical of an individual police officer yet praise police administrators for handling the misbehavior of an individual officer.

Violence in Video Games It seems that whenever some new technology or musical idiom becomes popular, someone becomes interested in linking it to behavior. Examples include: television and violence; pornography and sexual assault; suggestive lyrics in popular music and sexual behavior. It is always difficult to establish causality in such cases, and we will say nothing more about that. But content analysis is the appropriate research tool for classifying content as violent or sexually explicit.

Kimberly Thompson and Kevin Haninger examined the contents of video games rated in categories “E” (suitable for everyone) and “T” (teens, ages 13 and up) by the Entertainment Software Rating Board (ESRB). Their first study (Thompson and Haninger, 2001) sampled 55 of over 600 E-rated games available at the time. An undergraduate college student “with considerable video gaming experience” (p. 592) was assigned to play all games for 90 minutes, or until the game reached a natural conclusion. The game player was videotaped, which formed the basis for content analysis. One researcher, also described as being an experienced player, and the game player reviewed the video, coding several dimensions of what was depicted. In this case, the video game was the unit of analysis.

Coders counted the number of violent incidents depicted while the game was being played, and timed the duration of each violent incident. Violence was defined as: “... acts in which the aggressor causes or attempts to cause physical injury or death to another character.” This is an example of latent content. The duration of violent acts was manifest content, though researchers had to distinguish short pauses between violent acts. Additional variables coded included the number of deaths; the presence of drugs, alcohol, or tobacco; profanity and sexual behavior; weapon use; and whether any music was included that itself was rated as explicit. Comparing the duration of violent acts and
the number of deaths to how long each game was played yielded two standardized measures: violent minutes as a percent of all minutes, and the number of deaths per minute. Results showed quite a lot of violence. Action games ranged from 3.3 percent (“Sonic Adventure”) to 91 percent (“Nuclear Strike”) violence as a portion of total time. “Paperboy” depicted no deaths, but “Rat Attack” average 8.4 deaths per minute. Games classified as “sports” rarely showed violence.

Later research used similar methods to examine violence in a larger number of games rated as suitable for teens (Haninger and Thompson, 2004). These games displayed a wider variety of behaviors in the general domains of violence, obscenity, substance use, and sexual behavior. Again, the authors did not attempt to link such content with behavior. Their content analysis centered on systematically classifying what sorts of things were depicted in video games, thus providing information independent of industry ratings. Of their findings, authors highlight that ESRC ratings did not mention several examples of violence in almost half of the games reviewed.

Classifying Gang-Related Homicides

When is a homicide gang-related? Are there different types of gang-related homicides? These two questions guided research by Rosenfeld and associates (Rosenfeld, Bray, and Egley, 1999) to understand how gang membership might facilitate homicide in different ways. To address these questions, the researchers conducted a content analysis of police case files for homicides in St. Louis over a 10-year period.

By now, you should recognize the importance of conceptualization in most criminal justice research. Rosenfeld and associates began by further specifying the ambiguous term gang-related. They distinguished gang-motivated and gang-affiliated homicides. Gang-motivated killings “resulted from gang behavior or relationships, such as an initiation ritual, the ‘throwing’ of gang signs, or a gang fight” (1999:500). Gang-affiliated homicides involved a gang member as victim or offender, but with no indication of specific gang activity; a gang member killing a nongang person during a robbery is an example. A third category, nongang youth homicide, included all other murders in which no evidence of gang activity was available, and the suspected offender was between ages 10 and 24.

Because St. Louis police did not apply the labels “gang-affiliated” or “gang-motivated,” it was necessary for researchers to code each case into one of the three categories using information from case files. This was a form of content analysis—systematically classifying the messages contained in homicide case files. Homicide case files are good examples of police records that are not maintained for research purposes. Recognizing this, Rosenfeld and associates coded the files in a two-stage process, building reliability checks into each stage.

First, one person coded each case as either gang-related or not gang-related. This might seem a step backwards, but it focused researchers’ measurement on the separate dimensions of homicide of interest to them by simplifying the coding process. It was relatively easy to determine whether any evidence of gang activity or membership was present; if not, the case was classified as a nongang youth killing and set aside. Cases that had some evidence of gang involvement were retained for the second coding stage. During this stage, a second researcher randomly selected a 10 percent sample of cases and coded them again, without knowing how the first coder had classified the sampled cases. You will recognize this as an example of inter-rater reliability.

The second coding stage involved the finer and more difficult classification of cases as either gang-motivated or gang-affiliated. Inter-rater reliability checks were again conducted, this time on a 25 percent sample of cases. More cases were selected because reliability was lower in this stage—the two coders exhibited less agreement on how to classify gang homicides. Cases in which independent coding produced
discrepancies were reviewed and discussed by the two coders until they agreed on how the homicide should be classified.

From these very different examples, we expect that you can think of many additional applications of content analysis in criminal justice research. You might wish to consult Ray Surette’s (2006) excellent book Media, Crime, and Justice: Images, Realities, and Policies to learn more about the scope of topics for which content analysis can be used. The General Accounting Office (1996) has an excellent guide to content analysis generally.

Secondary Analysis

Data collected by other researchers are often used to address new research questions.

Our final topic encompasses all sources of criminal justice data we have described in this and preceding chapters: content analysis, agency records, field observation, and surveys. We begin with an example of an unusually ambitious use of secondary data by a prolific criminal justice scholar.

For almost three decades, Wesley Skogan has examined the influence of crime on the lives of urban residents. In most cases, his research has relied on sample surveys to investigate questions about fear of crime (Skogan and Maxfield, 1981), community crime prevention (Skogan, 1988), and the relationships between urban residents and police (Skogan, 1990), among others. He has long recognized the importance of incivilities—symbols of social disorder—as indicators of neighborhood crime problems and as sources of fear for urban residents.

In 1990, Skogan published a comprehensive study of incivilities, drawing on his own research as well as studies by others (Skogan, 1990). However, instead of conducting new surveys to collect original data, Skogan’s findings were based on secondary analysis of 40 surveys conducted in six cities from 1977 through 1983. He aggregated responses from about 13,000 individuals and examined questions about the sources of disorder, its impact, and the scope of action by individuals and police.

Secondary analysis of data collected by other researchers has become an increasingly important tool. Like Skogan, numerous criminal justice researchers have reanalyzed data collected by others. Several factors contribute to this, including the high cost of collecting original data through surveys or other means. More important, however, is that data for secondary analysis are readily available, largely due to efforts by the National Institute of Justice (NIJ), the BJS, and the Interuniversity Consortium for Political and Social Research (ICPSR).

Suppose you are interested in the relationship between delinquency, drug use, and school performance among adolescents. The National Youth Survey (NYS), which includes responses from 1725 youths interviewed nine times from 1975 through 2004, might suit your needs nicely. NYS data were originally collected by Delbert Elliott and associates (for example, Elliott, Huizinga, and Ageton, 1985). However, like Janet Lauritsen and associates (Lauritsen, Sampson, and Laub, 1991), who used the NYS to examine links between delinquency and victimization, you may be able to reanalyze the survey data to address your own research questions.

Or perhaps you wish to learn whether there are differences in the sentencing decisions of black and white judges. Cassia Spohn (1990) addressed this question using data originally collected by Milton Heumann and Colin Lof tin (1979), who were interested in the effect of a new Michigan law on plea bargaining. Spohn was able to conduct a secondary analysis of the same data to answer a different research question. Let’s examine these examples more closely to see how they illustrate the uses and advantages of secondary analysis.

Original NYS data were collected by Elliott and associates (1985:91) for three related research purposes: (1) to estimate the prevalence and incidence of delinquency and drug use among U.S. adolescents, (2) to assess causal
AGENCY RECORDS AND CONTENT ANALYSIS

Data from agency records lie at the heart of research and policy concern about racial profiling. As we have pointed out in earlier chapters, data that show disproportionate numbers of minorities stopped for traffic violations have been repeatedly cited as evidence of discrimination. Some such claims have been based on inappropriate comparisons, such as comparing records of traffic stops on an interstate highway to the population of a state. One interesting side effect of concern about racial profiling has been more careful attention to the characteristics of data on traffic stops.

What We Don't Know . . .

Until 1999, New Jersey State Police records of traffic stops were incomplete, unreliable, and were the foundation of claims that simultaneously documented and denied patterns of racial disparities in traffic stops. State police were supposed to record information on “the racial characteristics of detained motorists” (Veniero and Zoubek, 1999:31), but such information was often missing. Investigations by the New Jersey Attorney General and the U.S. Department of Justice lamented the inconsistent records. In fact, the missing information on race of people stopped was cited by New Jersey State Police, claiming that they had no way of knowing whether black drivers were stopped, searched, and otherwise subject to greater scrutiny than were white drivers. As a result, a formal agreement between the U.S. Department of Justice and the state of New Jersey included broad specifications for improving data collection. Furthermore, data collection in New Jersey was subject to semiannual audits for several years. This is a good example of the principle that the quality of data gathered by public agencies is directly related to the extent those data are used by the agency.

New Jersey was not alone. A 1999 survey by the Bureau of Justice Statistics found that found that only 9 of 49 state police agencies routinely collected demographic information on the race of drivers for all traffic stops (Strom and Durose, 2000). Growing national interest in racial profiling promoted several states to begin collecting such data, so that by 2004, 22 states recorded demographic information on drivers stopped (Hickman, 2005). The increase in states collecting data was accompanied by attention to data quality. Guides published by the National Institute of Justice (Ramirez, McDevitt, and Farrell, 2000), and the Police Executive Research Forum (Fridell, 2004, 2005) explained how police agencies should collect, audit, and analyze stop data for evidence of racial disparities.

Supplementing Pennsylvania Data

Engel and associates drew on these resources to design new data collection procedures for the Pennsylvania State Police. This was a laudable effort to enhance the reliability of traffic stop data at the outset of their research. The research team first worked with command staff to gain approval. They then included line officers and union representatives in a collaborative effort to design new data collection forms and procedures. As we mentioned in Chapter 3, the confidentiality of troopers was protected by removing identifying information from traffic stop records. Researchers believed this was an important component of their efforts to enhance the accuracy of data, at the same time supporting their claims that data would not be used to monitor individual troopers.

This kind of attention to detail appeared to pay off. About 327,000 stop records were produced in the first year; missing information was found in only 4 percent. Of the total stops, 75 percent were for speeding, adding support to the authors’ decision to use speeding as the primary indicator for traffic violation (2004:24). Residence of driver turned out to be an important item on the data collection form. Almost all
drivers (96%) were stopped outside the municipality where they lived, and two-thirds were stopped outside their home county (p. 43). An even larger proportion of black drivers were stopped outside their county of residence (82%), and half of black drivers lived outside of Pennsylvania. These findings offer strong evidence against using resident population as a benchmark to assess disproportionality in traffic stops.

Citation Zones
Examining traffic stop data in North Carolina, Smith et al. (2003) discovered very uneven spatial distributions of stops. They found concentrations of stops on some highway segments, whereas nearby roads had few traffic stops. This led the researchers to conclude that “citation zones” played a major role in where cars were stopped. Another phrase for citation zone is speed trap, but Smith et al. argue that this term oversimplifies a complex process of deciding where to deploy patrol units. Analysis of stop records indicated that citation zones were most often found on interstate highways, and were in fact concentrated in a relatively small number of locations. This proved to be quite important, as data from drivers’ surveys (discussed in Chapter 9) revealed that African American drivers were more likely to report traveling on interstate highways, especially in unfamiliar locations. As a result, African American drivers were disproportionately exposed to citation zones; they were more likely to drive in areas where enforcement was concentrated.

Maxfield and Kelling (2005) report similar findings for New Jersey, combining data on traffic stops with data on traffic volume. About 46 percent of traffic stops over the course of 1 year took place in the southern segment of the turnpike. From earlier chapters you may recall that the southern segment was an outlier with respect to race of drivers stopped. Consulting traffic volume data from the New Jersey Department of Transportation showed the southern segment had the lowest volume of traffic, just about one-fourth the average daily volume of the northern segment in the metropolitan New York area. Combining agency records from these different sources suggested that the southern segment was clearly a citation zone.

Content Analysis of Problems
In Chapter 10 we mentioned Carsten Andresen’s unstructured interviews with troopers. Among the topics he covered was whether certain areas or situations were believed to be particular problems. Here are excerpts from the very detailed coding instructions Andresen used to classify trooper comments (2005:236):

**Commercial Attractions:** Problems in a specific commercial area in the patrol beat. For example, one trooper complained about an amusement park that attracted scores of people throughout the year and created traffic problems.

**Crime Hot Spots:** Reference to geographic areas that were the setting for either minor or major crimes. This category does not refer to Commercial Attractions, Low Income Areas, Rest Areas, or Seasonal Areas. Rather this category refers to troopers who said they have to deal with a town full of bars that are filled with drunks, underpasses where people loiter, parks where people do drugs or take prostitutes, and/or sections of back rural roads where drunks travel.

**Illegal Immigrants:** Reference to illegal aliens or migrant workers. This category is important because some troopers stated that illegal aliens are a difficult group because they do not carry driver’s licenses and their cars are illegally registered, which is problematic during a traffic stop, traffic accident, and/or criminal incident.

**Seasonal Areas:** Problems in seasonal areas during the summer. Specifically, this category captures troopers who complained about increased traffic to the beach, the behavior of people at campsites, or problems with break-ins in summer homes (which are reported at the start of the summer when the summer people first come and find out that someone broke into their house).
relationships between drug use and delinquency, and (3) to test a comprehensive theory of delinquency. The NYS was designed as a panel survey, in which a nationally representative sample of youths ages 11–17 in 1976 was interviewed once each year from 1976 through 1989. As we described in Chapter 4, this is an example of a longitudinal study, and it is especially well suited to disentangling the time ordering of such behaviors as drug use and delinquency.

Lauritsen and associates (1991) were interested in the time order of somewhat different behaviors—delinquency and victimization—that were not directly addressed by the original researchers. A longitudinal design was equally important in this secondary analysis because Lauritsen and associates sought to determine whether youths experienced violent victimization after committing delinquent acts, or vice versa. Given this research interest, they faced two choices: collect original data by conducting a new panel survey, or reanalyze existing data from a panel survey that included questions on victimization and self-reported delinquency. Because the NYS included questions appropriate for their research purpose, Lauritsen and colleagues were spared the need and (considerable) expense of conducting a new panel study.

The second example mentioned earlier differs in two ways. First, the research questions addressed by Spohn (1990) and by Heumann and Loftin (1979), who collected the original data, are quite different. Heumann and Loftin studied the impact of a new Michigan law that specified mandatory minimum prison sentences for defendants who used firearms in the course of committing a felony offense. Their primary interest was in reductions in plea bargaining by prosecutors in Michigan’s largest county following passage of the firearm statute. Spohn, however, used the same data to address the very different question of whether sentences imposed by black judges systematically differed from those imposed by white judges. Her research interest required data from a site with a sufficient number of black criminal court judges, a condition that was met by Detroit Recorder’s Court in Wayne County (Michigan), where Heumann and Loftin had conducted their original research.

You may have already guessed the second difference between our two examples: Research by Spohn used data that had been collected from court records, whereas Lauritsen and associates conducted secondary analysis of survey data. Spohn could have gathered original information from court records, in Detroit or some other city, but she was able to address her research question by conducting a new analysis of data that had already been collected from court records.

**Sources of Secondary Data**

As a college student, you probably would not be able to launch an eight-wave panel study of a national sample of adolescents, or even gather records from some 2600 felony cases in Wayne County. You do, however, have access to the same data used in those studies, together with data from thousands of other research projects, through the ICPSR at the University of Michigan.

For over 40 years, the ICPSR has served as a central repository of machine-readable data collected by social science researchers. Current holdings include data from thousands of studies conducted by researchers all over the world. Of particular interest to criminal justice researchers is the National Archive of Criminal Justice Data (NACJD), established by the BJS in cooperation with the ICPSR. Here, you will find the NYS, Heumann and Loftin’s sentencing data, and each of the 40 surveys analyzed by Skogan for the book we mentioned earlier. There’s more, including surveys on criminal justice topics by national polling firms, the NCVS from 1972 to the present, periodic censuses of juvenile detention and correctional facilities, a study of thefts from commercial trucks in New York City, and data from
Chapter 11  Agency Records, Content Analysis, and Secondary Data

Marvin Wolfgang’s classic study of a Philadelphia birth cohort. Data from the growing National Incident-Based Reporting System (NIBRS) are now available, with an expanding number of participating agencies dating from 1996. Data from a growing number of regular data series are available for online data analysis. The possibilities are almost endless and grow each year as new data are added to the archives.

One of the most useful websites for aggregate secondary data is maintained by the BJS. Summary tabulations for many published data series are presented as graphs or tables. In addition, it’s possible to download summary data in spreadsheet format to easily conduct additional analysis, to display graphs in different forms, and even to prepare electronic slides for presentations.

Other sites on the Internet offer a virtually unlimited source of secondary data. You can obtain documentation for most data archived by the ICPSR and the NACJD, as well as health statistics, census data, and other sources limited only by your imagination. See the companion website for more information.

Advantages and Disadvantages of Secondary Data
The advantages of secondary analysis are obvious and enormous: It is cheaper and faster than collecting original data, and, depending on who did the original study, you may benefit from the work of topflight professionals and esteemed academics.

Potential disadvantages must be kept in mind, however. The key problem involves the recurrent question of validity. When one researcher collects data for one particular purpose, you have no assurance that those data will be appropriate to your research interests. Typically, you’ll find that the original researcher collected data that “come close” to measuring what you are interested in, but you may wish key variables had been operationalized just a little differently. The question, then, is whether secondary data provide valid measures of the variables you want to analyze.

This closely resembles one of the key problems in the use of agency records. Perhaps a particular set of data do not provide a totally satisfactory measure of what interests you, but other sets of data are available. Even if no one set of data provides totally valid measures, you can build up a weight of evidence by analyzing all the possibilities. If each of the imperfect measures points to the same research conclusion, you will have developed considerable support for its accuracy. The use of replication lessens the problem.

In general, secondary data are least useful for evaluation studies. This is the case because evaluations are designed to answer specific questions about specific programs. It is always possible to reanalyze data from evaluation studies, but secondary data cannot be used to evaluate an entirely different program. Thus, for example, a number of researchers have reexamined data collected for a series of domestic violence experiments conducted by Lawrence Sherman and others in several cities (see Sherman, 1992, for a summary). In most cases, these secondary researchers (such as Maxwell, Garner, and Fagan, 2001) wished to verify or reassess findings from the original studies. But it is not possible to use those data to answer questions about domestic violence interventions other than arrest or to evaluate arrest policies in new cities where the experiments did not take place.

In this book, the discussion of secondary analysis has a special purpose. As we conclude our examination of modes of observation in criminal justice research, you should have developed a full appreciation for the range of possibilities available in finding the answers to questions about crime and criminal justice policy. No single method of getting information unlocks all puzzles, yet there is no limit to the ways you can find out about things. And, more powerfully, you can zero in on an issue
from several independent directions, gaining an even greater mastery of it.

**Main Points**

- Reports and records produced by formal organizations may be the most common source of data in criminal justice research.
- Many public organizations produce statistics and data for the public record, and these data are often useful for criminal justice researchers.
- All organizations keep nonpublic records for internal operational purposes, and these records are valuable sources of data for criminal justice researchers.
- Agency records have many potential research uses, but because they are produced for purposes other than research, they may be unsuitable for a specific study.
- Researchers must be especially attentive to possible reliability and validity problems when they use data from agency records.
- “Follow the paper trail” and “Expect the expected” are two general maxims for researchers to keep in mind when using agency records in their research.
- Content analysis is a research method appropriate for studying human communications.
- Because communication takes many forms, content analysis can study many other aspects of behavior.
- Units of communication, such as words, paragraphs, and books, are the usual units of analysis in content analysis.
- Coding is the process of transforming raw data—either manifest or latent content—into a standardized, quantitative form.
- Secondary analysis refers to the analysis of data collected earlier by another researcher for some purpose other than the topic of the current study.
- Archives of criminal justice and other social data are maintained by the ICPSR and the NACJD for use by other researchers.
- The advantages and disadvantages of using secondary data are similar to those for agency records—data previously collected by some researcher may not match our own needs.

**Key Terms**

Latent content    Secondary analysis
Manifest content

**Review Questions and Exercises**

1. Each year, the BJS publishes the *Sourcebook of Criminal Justice Statistics*, a compendium of data from many different sources. Consult the online edition of the *Sourcebook* (see below), select a table of interest to you, and describe how the data presented in that table were originally collected.

2. In New York City, police officers assigned to a specialized gang squad pay special attention to graffiti, or tagging. In doing so, they conduct a type of content analysis to study actions, threats, and other messages presented in this form of communication. Describe how you would plan a formal content analysis of graffiti. Be sure to distinguish manifest and latent content, units of analysis, and coding rules for your study.

**Online Sessions for Chapter 11**

Online Sessions present further examples of selected material covered in each chapter. Your instructor may choose to use these for in-class discussion (as I do with my students) or you can access them at home. To ensure that you always have access to live, correct links for the websites described here, we provide the necessary links on the companion website for *Research Methods for Criminal Justice and Criminology* at http://www.thomsonedu.com/criminaljustice/maxfield. Once at the companion website, select this specific chapter, click on “Chapter Resources,” then click on “Web Links.”

**Sourcebook**

The *Sourcebook for Criminal Justice Statistics* is available online and provides summary statistics in many areas of interest to criminal justice researchers. Note data are organized in broad
categories. There is no single better compendium of agency data for criminal justice.

Visit this site, find a table or section that interests you, and bring your information to class to discuss. Be prepared to discuss the source of the data, the units of analysis, the way the constructs were operationalized, and any limitations to the reliability and validity of the data included in the table or section. http://www.albany.edu/sourcebook/

**Sources of Secondary Data**

As we discussed in this chapter, virtually all data presented in publications like the sourcebook are summary tabulations from original sources. Such tabulations are extremely useful, and can form the basis for original research and analysis. A vast array of original data resources are available at the National Archive of Criminal Justice Data (NACJD) maintained by the Interuniversity Consortium for Political and Social Research at the University of Michigan. As we mentioned in an earlier chapter, the data collections are extensively documented. By examining file descriptions you can learn quite a lot about how data were conducted.

Thousands of data files are available in two general categories. First, and of most general interest, are data series and one-time studies collected by or for government agencies. This includes the NCVS, NIBRS, and all other series we have described in earlier chapters. Second, data from most research projects sponsored by the National Institute of Justice are also archived. This makes it possible to replicate or expand on a very large number of published research studies.

The URL below links to a subject-oriented browsing page. Data can be accessed in many other ways, but most people find this is a good place to begin. We suggest you learn about one data file of each type. http://www.icpsr.umich.edu/NACJD/archive.html

**Summary Data from BJS**

Visit the Bureau of Justice Statistics website. Select and print a graph from the “Key Facts at a Glance” section. Then search other areas of the BJS site to find data used to prepare the graph you have printed. Download the data into a spreadsheet program. Then conduct secondary analysis on those data—present a different graph, or prepare a graph displaying the data in a different format. http://www.ojp.usdoj.gov/bjs/glance.htm

**Crime and Justice Data from States**

Many states and larger cities present very comprehensive data and other information online. We have referred to New York City’s site a number of times. Visit other states to learn what sort of agency data is available. Washington state has an especially comprehensive site. The one for Texas borders on the macabre, with links to details about death row inmates and people executed. In addition, many state and local governments maintain online lists of people in local correctional facilities.


**International Crime Victimization Survey (ICVS)**

First conducted in 1989, the ICVS has now covered a large number of countries and individual cities. International comparisons are challenging, but it’s fascinating to browse victimization data for cities in Africa and the Pacific rim. From the home page of the ICVS, link to “statistics” for summary tabulations for industrialized counties, central and eastern Europe, and developing nations. http://ruljis.leidenuniv.nl/group/jfcr/www/icvs/

**Terrorism Database**

Getting reliable information and data about terrorist events is probably even more challenging than getting data about crime, or traffic violations. Nevertheless, the Memorial Institute for the Prevention of Terrorism (MIPT) maintains a website that archives reports of terrorist incidents, and offers different ways to view and analyze the data. For example, try the “Incident Analysis Wizard.” Measurement problems abound here, but it’s interesting to browse through this source of archived data. http://www.tkb.org/Home.jsp

**Additional Readings**

the U.S. Office of Management and Budget directed that all federal agencies develop guidelines to maximize the quality of information they collect and disseminate. This publication describes how the BJS complied with that directive. It provides an excellent overview of validity and reliability issues in series of data often used by criminal justice researchers.


Jacob, Herbert, *Using Published Data: Errors and Remedies* (Thousand Oaks, CA: Sage, 1984). We have often referred to this small book. It is an extremely valuable source of insight into the promise and pitfalls of using agency records. All social scientists should read this book carefully.

This final section of the book draws on concepts and ideas from earlier chapters to bring you closer to the actual process of criminal justice research. Having examined the role of theory, cause and effect, measurement, experiments, and different ways of collecting data, we are now ready to see how these pieces come together.

Criminal justice research can be conducted in many ways to answer many different types of questions. We have touched on various research purposes throughout the text, but the first chapter in this section examines a specific research purpose more closely. Because crime is an important and seemingly intractable social problem, applied research is attracting growing interest from researchers and public officials alike.

Chapter 12 describes evaluation research and problem analysis. As we will see, carefully specifying concepts and being attentive to measures are as important for applied research as they are for other research purposes.

Chapter 13 takes up the question of analysis. After we have designed a research project, specified measures, and collected data, our attention will turn to a search for patterns and relationships for description, explanation, or evaluation, depending on the research purpose. In Chapter 13, we take a preliminary look at descriptive and inferential statistics. Our goal is to establish a familiarity with the principles of basic statistical analysis.
Chapter 12

Evaluation Research and Problem Analysis

In this chapter, our attention centers on applied criminal justice research. Evaluation studies are conducted to learn whether (and why) programs have succeeded or failed. Problem analysis helps officials plan their actions and anticipate the possible effects of new programs.

Introduction 357
Topics Appropriate for Evaluation Research and Problem Analysis 358
The Policy Process 358
Linking the Process to Evaluation 360
Getting Started 362
Evaluability Assessment 363
Problem Formulation 364
Measurement 366
Designs for Program Evaluation 370
Randomized Evaluation Designs 370

Home Detention: Two Randomized Studies 373
Quasi-Experimental Designs 375
Other Types of Evaluation Studies 379

Problem Analysis and Scientific Realism 379
Problem-Oriented Policing 380
Auto Theft in Chula Vista 381
Other Applications of Problem Analysis 382
Space- and Time-Based Analysis 382
Scientific Realism and Applied Research 384

(continued)
Introduction

Evaluation research and problem analysis are increasingly important activities for researchers and public officials alike.

Evaluation research—sometimes called “program evaluation”—refers to a research purpose rather than a specific research method. Its special purpose is to evaluate the effects of policies such as mandatory arrest for domestic violence, innovations in probation, and new sentencing laws. Another type of evaluation study, problem analysis, helps public officials plan and select alternative actions. Virtually all types of designs, measures, and data collection techniques can be used in evaluation research and problem analysis.

Evaluation research in criminal justice is probably as old as criminal justice research generally. Whenever people have instituted a new program for a specific purpose, they have paid attention to its actual consequences, even if they have not always done so in a conscious, deliberate, or sophisticated fashion. Over time, the field of evaluation research has become an increasingly popular and active research specialty, which is reflected in the proliferation of associated textbooks, courses, and projects. As a consequence, you are likely to read increasing numbers of evaluation reports and, as a researcher, to be asked to conduct evaluations.

In part, the growth of evaluation research no doubt reflects increasing desire on the part of criminal justice researchers to actually make a difference in the world. At the same time, we cannot discount the influence of two additional factors: (1) increased federal requirements for program evaluations to accompany the implementation of new programs and (2) the availability of research funds to meet that requirement.

Many people know that the 1994 Crime Bill authorized billions of federal dollars to support community policing, drug courts, and other criminal justice initiatives. Fewer people realize that the Crime Bill also required that most of these new initiatives be evaluated. In the first five years after the Crime Act’s passage in 1994, the National Institute of Justice alone spent more than $210 million on research and evaluation projects to test the effects of the new legislation (National Institute of Justice, 2001).

By the same token, increased interest in program evaluation and problem analysis has followed heightened concern for the accountability of public officials and public policy. Criminal justice agencies are expected to justify the effectiveness and cost of their actions. If traditional approaches to probation supervision, for example, do not deter future lawbreaking, new approaches should be developed and their effectiveness assessed. Or, if using temporary detention facilities fabricated from recycled semitrailers is less costly than constructing new jails, public officials should consider whether the lower-cost alternative will meet their needs for pretrial detention and short-term incarceration.

Justice agencies have come to rely more on evidence-based policy, in which the actions...
of justice agencies are linked to evidence used for planning and evaluation. Traditional practices are being re-evaluated against evidence provided by social science research. The Problem-Oriented Guides series summarizes evidence concerning police responses to problems ranging from Acquaintance Rape of College Students (Sampson, 2002) to Underage Drinking (Johnson, 2004). Compstat and its variations base police actions on evidence about the location and circumstances of crime problems. Corrections policies are increasingly evaluated to sort out those that do in fact reduce reoffending (Cullen and Sundt, 2003). This trend represents an expansion of applied research that moves beyond collaborations between justice professionals and professional researchers.

Topics Appropriate for Evaluation Research and Problem Analysis

Problem analysis and evaluation are used to develop justice policy and determine its impact.

Evaluation research is appropriate whenever some policy intervention occurs or is planned. A policy intervention is an action taken for the purpose of producing some intended result. In its simplest sense, evaluation research is a process of determining whether the intended result was produced. Problem analysis focuses more on deciding what intervention should be pursued. Given alternative courses of action, which is likely to be least costly, most effective, or least difficult to implement? Our focus, of course, is on the analysis and evaluation of criminal justice policy and criminal justice agencies. However, it will be useful to first consider a simple general model of the policy-making process in order to understand various topics appropriate to evaluation and problem analysis.

The Policy Process

Figure 12.1 presents our model, adapted from Robert Lineberry’s (1977:42–43) classic summary of a policy system. A similar type of “input–output” model is described in a National Institute of Justice publication on evaluation guidelines (McDonald and Smith, 1989). Although we will examine each step in turn, recognize that the policy process, like the research process generally (see Chapter 4), is fluid and does not always start at the beginning and conclude at the end.

The policy process begins with some demand that normally appears as support for a new course of action or opposition to existing policy. Such demands can emerge from within a public organization or from outside sources. For example, newspaper stories alleging racial discrimination in drug sentencing can generate demand for revised sentencing policies, or a prosecutor may independently decide to review all sentence recommendations made by deputies who prosecute drug cases. Before any action can be taken, demands must find a place on the policy agenda. A prosecutor might ignore accusations of discrimination published in an alternative newspaper that is held in low esteem, or she or he may decide not to take actions on suggestions for changes that emerge from within the district attorney’s office.

The next step shown in Figure 12.1 actually encompasses several steps. Policy makers consider ultimate goals they wish to accomplish and different means of achieving those goals. Does our prosecutor seek absolute equality in sentences recommended for all white and African American drug defendants, or should there be ranges of permissible variation based on criminal history, severity of charges, and so on? Resources must be allocated from available inputs, including personnel, equipment, supplies, and even time. Who will review sentence recommendations? How much time will that take, and will additional staff be required? Because the word policy implies some standard course of action about how to respond to some recurring problem or issue, routine practices and decision rules must be formulated. Will sentence recommendations for each case be reviewed as they are prepared, or is it sufficient to review all cases on a weekly basis?
Policy outputs refer to what is actually produced, in much the same manner that a manufacturer of office supplies produces paper clips and staples. In our hypothetical example, the prosecutor’s policy produces the routine review of sentence recommendations in drug cases. Or, to consider a different example, a selective traffic enforcement program intended to reduce auto accidents on a particular roadway may produce a visible police presence, together with traffic citations for speeding.

In the final stage, we consider the impact of policy outputs. Does the prosecutor’s review process actually eliminate disparities in sentences? Are auto accidents reduced in the targeted enforcement area?

The distinction between policy outputs and their impacts is important for understanding applications of evaluation to different stages of the policy process. Unfortunately, this difference is often confusing to both public officials and researchers. Impacts are fundamentally related to policy goals; they refer to the basic question of what a policy action is trying to achieve. Outputs embody the means to achieve desired policy goals. A prosecutor seeks to achieve equality in sentence recommendations (impact), so a review process is produced as a means to achieve that goal (output). Or a police executive allocates officers, patrol cars, and overtime pay to produce traffic citations (outputs) in the expectation that citations will achieve the goal of reducing auto accidents (impact).

Now consider the left side of Figure 12.1. Our policy model can be expressed as a simple cause-and-effect process such as we considered in earlier chapters. Some cause has produced the variation in sentences for African American and white defendants, or some cause has produced a concentration of auto accidents. Policies are formulated to produce some effect or impact. In this sense, a policy can be viewed as a hypothesis in which an independent variable is expected to produce change in some...
dependent variable. Sentence review procedures are expected to produce a reduction in sentence disparities; targeted enforcement is expected to produce a reduction in auto accidents. Goal-directed public policies may therefore be viewed as if-then statements: If some policy action is taken, then we expect some result to be produced.

Linking the Process to Evaluation

By comparing this simple model with a general definition of program evaluation given in one of the most widely used texts on the subject (Rossi, Freeman, and Lipsey, 1999), the topics appropriate to applied research will become clearer. Peter Rossi and associates (1999:4) define program evaluation as

*the use of social science research procedures to systematically assess the effectiveness of social intervention programs.* More specifically, evaluation researchers (evaluators) use social research methods to study, appraise, and help improve social programs in all their aspects, including the diagnosis of the social problems they address, their conceptualization and design, their implementation and administration, their outcomes, and their efficiency. (emphasis in original)

We have been discussing systematic social science research procedures throughout this book. Now let’s substitute criminal justice for social programs and see how this definition and Figure 12.1 help us understand program evaluation applications.

Problem Analysis Activities listed under “Institutional Processes” in Figure 12.1 refer to conceptualization and design. For example, faced with a court order to maintain prison populations within established capacity, corrections officials might begin by conceiving and designing different ways to achieve this demand. Problem analysis is an example of a social science research procedure that can help corrections officials analyze alternative actions, choose among them, and formulate routine practices for implementing policy to comply with a court order.

One approach might be to increase rated capacity through new construction or conversion of existing facilities. Another might be to devise a program to immediately reduce the existing population. Still another might be to cut back on the admission of newly sentenced offenders. A more general goal that would certainly be considered is the need to protect public safety. Each goal implies different types of actions, together with different types and levels of resources, that would be considered within constraints implied by the need to protect public safety. If officials from other organizations—prosecutors, judges, or state legislators—were involved in conceptualization and design, then additional goals, constraints, and policies might be considered.

Increasing capacity by building more prisons would be the most costly approach in financial terms, but it might also be viewed as the most certain way to protect public safety. Early release of current inmates would be cheaper and faster than building new facilities, but this goal implies other decisions, such as how persons would be selected and whether they would be released to parole or to halfway houses. And each of these alternatives requires some organizational capacity to choose inmates for release, place them in halfway houses, or supervise compliance with parole. Refusing new admissions would be least costly. Political support must be considered for each possible approach. Each alternative—spending money on new construction, accepting responsibility for early release, or tacitly passing the problem on to jails that must house inmates refused admission to state facilities—requires different types of political influence or courage.

Many other topics in criminal justice research are appropriate for problem analysis. Police departments use such techniques to help determine the boundaries of patrol
beats. In most large cities, analysts examine the concentration of calls for service in terms of space and time and consider how street layout and obstacles might facilitate or impede patrol car mobility.

A growing number of law enforcement agencies are using computerized crime maps to detect emerging patterns in crime and develop appropriate responses. Producing computer-generated maps that display reported crimes within days after they have occurred is one of the most important policy planning tools for the New York City Police Department (Silverman, 1999). Other departments have taken advantage of funding and technical assistance made available by the 1994 Crime Bill to enhance mapping and other crime analysis capabilities. We’ll have more to say about mapping later in this chapter.

Court administrators often make changes in the way cases are scheduled for individual judges after estimating what effects such changes will have. Estimates are usually based on analysis of past case volume and predicted future volume. Many states now conduct “prison impact” studies to estimate shifts in prison populations that will result from legislative changes in sentence length.

**Program Evaluation** Problem analysis takes places in the policy-making stage. In contrast, program evaluation studies are conducted in later stages and seek answers to two types of questions: (1) Are policies being implemented as planned? and (2) Are policies achieving their intended goals? Evaluation, therefore, seeks to link the intended actions and goals of criminal justice policy to empirical evidence that policies are being carried out as planned and are having the desired effects. These two types of questions correspond to two related types of program evaluations: process evaluation and impact assessment. Returning to our example of policies to reduce prison population, we will consider first impact assessment and then process evaluation.

Let’s assume that corrections department policy analysts select an early-release program to reduce the population of one large institution. Inmates who have less than 120 days remaining on their sentence and who were committed for nonviolent offenses will be considered for early release. Further assume that of those inmates selected for early release, some will be assigned to parole officers, and some will serve their remaining sentence in halfway houses—working at jobs during the week but spending evenings and weekends in a community-based facility.

The program has two general goals: (1) to reduce prison population to the court-imposed ceiling and (2) to protect public safety. Whereas the first goal is fairly straightforward, the second is uncomfortably vague. What do we mean by “protecting public safety”? For now, let’s say we will conclude that the program is successful in this regard if, after 6 months, persons in the two early-release conditions have aggregate rates of arrest for new offenses equal to or less than a comparison group of inmates released after completing their sentences.

Our **impact assessment** would examine data on the prison population before and after the new program was implemented, together with arrest records for the two types of early releases and a comparison group. We might obtain something like the hypothetical results shown in Table 12.1.

Did the program meet its two goals? Your initial reaction might be that it did not, but Table 12.1 presents some interesting findings. The prison population certainly was reduced, but it did not reach the court-imposed cap of 1350. Those released to halfway houses had lower arrest rates than all others, but persons placed on early parole had higher arrest rates than those in the normal release condition. Averaging arrest rates for all three groups shows that the total figure is about the same as that for persons released early. Notice also that almost twice as many people were released to early parole as were placed in halfway houses.
The impact assessment results in Table 12.1 would have been easier to interpret if we had conducted a process evaluation. A process evaluation focuses on program outputs, as represented in Figure 12.1, seeking answers to the question of whether as the program was implemented as intended. If we had conducted a process evaluation of this early-release program, we might have discovered that something was amiss in the selection process. Two pieces of evidence in Table 12.1 suggest that one of the selection biases we considered in Chapter 7, “creaming,” might be at work in this program. Recall that creaming is the natural tendency of public officials to choose experimental subjects least likely to fail. In this case, selectivity is indicated by the failure of the early-release program to meet its target number, the relatively small number of persons placed in halfway houses, and the lower rearrest rates for these persons. A process evaluation would have monitored selection procedures and probably revealed evidence of excessive caution on the part of corrections officials in releasing offenders to halfway houses.

Ideally, impact assessments and process evaluations are conducted together. Our example illustrates the important general point that process evaluations make impact assessments more interpretable. In other cases, process evaluations may be conducted when an impact assessment is not possible. To better understand how process evaluations and impact assessments complement each other, let’s now look more closely at how evaluations are conducted.

### Getting Started

Learning policy goals is a key first step in doing evaluation research.

Several steps are involved in planning any type of research project. This is especially true in applied studies, for which even more planning may be required. In evaluating a prison early-release program, we need to think about design, measurement, sampling, data collection procedures, analysis, and so on. We also have to address such practical problems as obtaining access to people, information, and data needed in an evaluation.

In one sense, however, evaluation research differs slightly in the way research questions are developed and specified. Recall that we equated program evaluation with hypothesis testing; policies are equivalent to if-then statements postulating that some intervention will have some desired impact. Preliminary versions of research questions, therefore, will already have been formulated for many types of evaluations. Problem analysis usually considers a limited range of alternative choices; process evaluations focus on whether programs are carried out according to plans; and impact assessments evaluate whether specified goals are attained.

This is not to say that evaluation research is a straightforward business of using social science methods to answer specific questions that are clearly stated by criminal justice officials. It is often difficult to express policy goals in the form of if-then statements that are empirically testable. Another problem is the presence of conflicting goals. Many issues in criminal justice are complex, involving different organizations and people. And different organizations

<table>
<thead>
<tr>
<th>New Arrests After 6 Months</th>
<th>Number of Persons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal release</td>
<td>26%</td>
</tr>
<tr>
<td>Early release to halfway houses</td>
<td>17</td>
</tr>
<tr>
<td>Early parole</td>
<td>33</td>
</tr>
<tr>
<td>Subtotal early release</td>
<td>27</td>
</tr>
<tr>
<td>Total</td>
<td>26%</td>
</tr>
</tbody>
</table>

Note: Preprogram population = 1578; actual population after implementation = 1402; court-imposed population cap = 1350.
and people may have different goals that make it difficult to define specific evaluation questions. Perhaps most common and problematic are vague goals. Language describing criminal justice programs may optimistically state goals of “enhancing public safety by reducing recidivism” without clearly specifying (or knowing) what is meant by that objective.

In most cases, researchers have to help criminal justice officials formulate testable goals, something that is not always possible. Other obstacles may interfere with researchers’ access to important information. Because of these and similar problems, evaluation researchers must first address the question of whether to evaluate at all.

**Evaluability Assessment**

An evaluability assessment is described by Rossi and associates (1999:157) as sort of a “preevaluation,” in which a researcher determines whether conditions necessary for conducting an evaluation are present. One obvious condition is support for the study from organizations delivering program components that will be evaluated. The word *evaluation* may be threatening to public officials, who fear that their own job performance is being rated. Even if officials do not feel personally threatened by an impact assessment or other applied study, evaluation research can disrupt routine agency operations. Ensuring agency cooperation and support is therefore an important part of evaluability assessment. Even if no overt opposition exists, officials may be ambivalent about evaluation. This might be the case, for example, if an evaluation is required as a condition of launching some new program.

This and other steps in evaluability assessment may be accomplished by “scouting” a program and interviewing key personnel (Rossi, Freeman, and Lipsey, 1999:135). The focus in scouting and interviewing should be on obtaining preliminary answers to questions that eventually will have to be answered in more detail as part of an evaluation. What are general program goals and more specific objectives? How are these goals translated into program components? What kinds of records and data are readily available? Who will be the primary consumers of evaluation results? Do other persons or organizations have some direct or indirect stake in the program? Figure 12.2 presents a partial menu of questions that can guide information gathering for the evaluability assessment and later stages.

The answers to these and similar questions should be used to prepare a program description. Although “official” program descriptions may be available, evaluation researchers should always prepare their own description, one that reflects their own understanding of program

---

**Figure 12.2** Evaluation Questions

*Source: Adapted from Stecher and Davis (1987: 58–59).*
goals, elements, and operations. Official documents may present incomplete descriptions or ones intended for use by program staff, not by evaluators. Even more importantly, official program documents often do not contain usable statements about program goals. As we will see, formulating goal statements that are empirically testable is one of the most important components of evaluation research.

Douglas McDonald and Christine Smith (1989:1) describe slightly different types of questions to be addressed by criminal justice officials and evaluators in deciding whether to evaluate state-level drug control programs:

- How central is the project to the state’s strategy?
- How costly is it relative to others?
- Are the project’s objectives such that progress toward meeting them is difficult to estimate accurately with existing monitoring procedures?

Such questions are related to setting both program and evaluation priorities. On the one hand, if a project is not central to drug control strategies, or if existing information can help determine project effectiveness, then an evaluation should probably not be conducted. On the other hand, costly projects that are key elements in antidrug efforts should be evaluated so that resources can be devoted to new programs if existing approaches are found to be ineffective.

Although Rossi and associates (1999) describe it as a distinct type of research, an evaluability assessment does not need to be a major project in and of itself. Often, a few questions posed to a few people, together with a careful reading of program documents, will yield sufficient information to decide whether to proceed further. The questions presented in Figure 12.2 are a good guide for developing a scouting report. If the scouting report indicates that a particular program will not be readily evaluable, it is far better to make that determination before beginning a full-blown study.

**Problem Formulation**

We mentioned that evaluation research questions may be defined for you. This is true in a general sense, but formulating applied research problems that can be empirically evaluated is an important and often difficult step. Evaluation research is a matter of finding out whether something is or is not there, whether something did or did not happen. To conduct evaluation research, we must be able to operationalize, observe, and recognize the presence or absence of what is under study.

This process normally begins by identifying and specifying program goals. The difficulty of this task, according to Rossi and associates (1999:167), revolves around the fact that formal statements of goals are often abstract statements about ideal outcomes. Here are some examples of goal statements paraphrased from actual program descriptions:

- Equip individuals with life skills to succeed (a state-level shock incarceration program; MacKenzie, Layton, Shaw, and Gowdy, 1993).
- Provide a safe school environment conducive to learning (a school resource officer program; Johnson, 1999).
- Encourage participants to accept the philosophy and principles of drug-free living (an urban drug court; Finn and Newlyn, 1993).
- Provide a mechanism that engages local citizens and community resources in the problem-solving process (a probation–police community corrections program; Wooten and Hoelter, 1998).

Each statement expresses a general program objective that must be clarified before we can formulate research questions to be tested empirically. We can get some idea of what the first example means, but this goal statement raises several questions. The objective is for individuals to succeed, but succeed at what? What is meant by “life skills”—literacy, job training, time management, self-discipline? We might
also ask whether the program focuses on outputs (equipping people with skills) or on impacts (promoting success among people who are equipped with the skills). On the one hand, an evaluation of program outputs might assess individual learning of skills, without considering whether the skills enhance chances for success. On the other hand, an evaluation of program impacts might obtain measures of success such as stable employment or not being arrested within some specified time period.

In all fairness, these goal statements are taken somewhat out of context; source documents expand on program goals in more detail. They are, however, typical of stated goals, or initial responses we might get to the question, What are the goals of this program? Researchers, however, require more specific statements of program objectives.

Wesley Skogan (1985) cautions that official goal statements frequently “oversell” what a program might realistically be expected to accomplish. It’s natural for public officials to be positive or optimistic in stating goals, and overselling may be evident in goal statements. Another reason officials and researchers embrace overly optimistic goals is that they fail to develop a micromodel of the program production process (Weiss, 1995). That is, they do not adequately consider just how some specified intervention will work. Referring back to Figure 12.1, we can see that developing a micromodel can be an important tool for laying out program goals and understanding how institutional processes are structured to achieve those goals. Skogan (1985:38; emphasis in original) describes a micromodel as part of what is meant by a “theory-driven” evaluation. Researchers and program personnel should together consider just how each element of a program should affect its targets. If there is not a good reason why “X” should cause “Y” the evaluation is probably not going to find that it did! Micromodeling is another good reason for monitoring the actual implementation of programs.

A micromodel can also reveal another problem that sometimes emerges in applied studies: inconsistent goals.

For example, Michael Maxfield and Terry Baumer (1992) evaluated a pretrial home detention program in which persons awaiting trial for certain types of offenses were released from jail and placed on home detention with electronic monitoring. Five different criminal justice organizations played roles in implementation or had stakes in the program. The county sheriff’s department (1) faced pressure to reduce its jail population. Under encouragement from the county prosecutor (2), the pretrial release program was established. Criminal court judges (3) had the ultimate authority to release defendants to home detention, following recommendations by bail commissioners in a county criminal justice services agency (4). Finally, a community corrections department (5) was responsible for actually monitoring persons released to home detention.

Maxfield and Baumer (1992) interviewed persons in each of these organizations and discovered that different agencies had different goals. The sheriff’s department was eager to release as many people as possible to free up jail space for convicted offenders and pretrial defendants who faced more serious charges. Community corrections staff, charged with the task of monitoring pretrial clients, were more cautious and sought only persons who presented a lower risk of absconding or committing more offenses while on home detention. The county prosecutor viewed home detention as a way to exercise more control over some individuals who would otherwise be released under less restrictive conditions. Some judges refused to release people on home detention, whereas others followed prosecutors’ recommendations. Finally, bail commissioners viewed pretrial home detention as a form of jail resource management, adding to the menu...
of existing pretrial dispositions (jail, bail, or release on recognizance).

The different organizations involved in the pretrial release program comprised multiple stakeholders—persons and organizations with a direct interest in the program. Each stakeholder had different goals for and different views on how the program should actually operate—who should be considered for pretrial home detention, how they should be monitored, and what actions should be taken against those who violated various program rules. After laying out these goals and considering different measures of program performance, Maxfield and Baumer (1992:331) developed a micromodel of home detention indicating that electronic monitoring is suitable for only a small fraction of defendants awaiting trial.

Clearly specifying program goals, then, is a fundamental first step in conducting evaluation studies. If officials are not certain about what a program is expected to achieve, it is not possible to determine whether goals are reached. Or, if multiple stakeholders embrace different goals, evaluators must specify different ways to assess those goals. Maxfield (2001) describes a number of different approaches to specifying clear goals, a crucial first step in the evaluation process.

Measurement
After we identify program goals, our attention turns to measurement, considering first how to measure a program’s success in meeting goals. Rossi and associates (1999:83–84) state this in terms that should now be familiar:

For an evaluation question to be answerable, it must be possible to identify in advance some evidence or “observables” that can realistically be obtained and will be credible as the basis for an answer. This generally means developing questions (1) that involve measurable performance dimensions; (2) that are sufficiently unambiguous so that explicit, noncontroversial definitions can be given for each of their terms; and (3) for which the relevant standards or criteria are specified or obvious.

Obtaining evaluable statements of program goals is conceptually similar to the measurement process, in which program objectives represent conceptual definitions of what a program is trying to accomplish. Chapter 5 began with the following sentence: “This chapter describes the progression from having a vague idea about what we want to study to being able to recognize it and measure it in the real world.” As evaluation researchers, we must similarly recognize the need to state clear conceptual definitions and then describe specific operational definitions.

Specifying Outcomes If a criminal justice program is intended to accomplish something, we must be able to measure that something. If we want to reduce fear of crime, we need to be able to measure fear of crime. If we want to increase consistency in sentences for drug offenses, we need to be able to measure that. Notice, however, that, although outcome measures are derived from goals, they are not the same as goals. Program goals represent desired outcomes, whereas outcome measures are empirical indicators of whether those desired outcomes are achieved. Furthermore, if a program pursues multiple goals, then researchers may have to either devise multiple outcome measures or select a subset of possible measures to correspond with a subset of goals.

Keeping in mind our program-as-hypothesis simile, outcome measures correspond to dependent variables—the $Y$ in a simple $X \rightarrow Y$ causal hypothesis. Because we have already considered what’s involved in developing measures for dependent variables, we can describe how to formulate outcome measures. Pinning down program goals and objectives results in a conceptual definition. We then specify an operational definition by describing empirical indicators of program outcomes.
In our earlier example, Maxfield and Baumer (1992) translated the disparate interests of organizations involved in pretrial home detention into three more specific objectives: (1) ensure appearance at trial, (2) protect public safety, and (3) relieve jail crowding. These objectives led to corresponding outcome measures: (1) failure-to-appear rates for persons released to pretrial home detention, (2) arrests while on program, and (3) estimates of the number of jail beds made available, computed by multiplying the number of persons on pretrial home detention by the number of days each person served on the program.

Table 12.2 Pretrial Home Detention with Electronic Monitoring: Goals, Objectives, and Measures

<table>
<thead>
<tr>
<th>Actor/Organization</th>
<th>Goals</th>
<th>Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sheriff</td>
<td>Release jail inmates</td>
<td>Failure-to-appear courts</td>
</tr>
<tr>
<td>Prosecutor</td>
<td>Increase supervision of pretrial defendants</td>
<td>Arrests while on program</td>
</tr>
<tr>
<td>Judges</td>
<td>Protect public safety</td>
<td></td>
</tr>
<tr>
<td>Ball commission</td>
<td>Provide better jail resource management</td>
<td></td>
</tr>
<tr>
<td>Community corrections</td>
<td>Monitor defendant compliance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Return violators to jail</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>N defendants × days served</td>
</tr>
</tbody>
</table>

Source: Adapted from Maxfield and Baumer (1992).

In our earlier example, Maxfield and Baumer (1992) translated the disparate interests of organizations involved in pretrial home detention into three more specific objectives: (1) ensure appearance at trial, (2) protect public safety, and (3) relieve jail crowding. These objectives led to corresponding outcome measures: (1) failure-to-appear rates for persons released to pretrial home detention, (2) arrests while on program, and (3) estimates of the number of jail beds made available, computed by multiplying the number of persons on pretrial home detention by the number of days each person served on the program. Table 12.2 summarizes the goals, objectives, and measures defined by Maxfield and Baumer.

Let’s consider another example. Staff from the Department of the Youth Authority in California conducted an evaluation of a boot camp and intensive parole program known as LEAD—an acronym for qualities the program sought to instill in participants: leadership, esteem, ability, and discipline. The three major goals of LEAD were to (1) reduce recidivism, (2) ease crowded conditions in juvenile facilities, and (3) provide a cost-effective treatment option (Department of the Youth Authority, 1997:18).

Evaluators assessed the effectiveness of the LEAD program by comparing LEAD participants with other young offenders who were not in LEAD. Data were collected on lengths of stay in each institution, estimated program costs, and rates of recidivism. Measures of recidivism were obtained from official records of arrests and parole violations for 12-, 18-, and 24-month periods after release. Length of incarceration was used to measure the reduction of institutional crowding, as well as cost savings. It was assumed that LEAD participants would have shorter average lengths of incarceration than youths not participating but that these short-term cost savings would be reduced if LEAD participants were later reincarcerated more rapidly or at higher rates. Recidivism data routinely collected on all subjects were used to examine whether this was the case.

Measuring Program Contexts Measuring the dependent variables directly involved in an impact assessment is only a beginning. As Ray Pawson and Nick Tilley (1997:69) point out, it is usually necessary to measure the context within which the program is conducted. These variables may appear to be external to the experiment itself, yet they still affect it.

Consider, for example, an evaluation of a job-skills training program coupled with early prison release to a halfway house. The primary outcome measure might be participants’ success at gaining employment after completing the program. We will, of course, observe and calculate the subjects’ employment rates. We should also be attentive to what has happened to the employment/unemployment rates of the community and state where the program is located. A general slump in the job market should be taken into account in assessing what might otherwise seem to be a low employment
rate for subjects. Or, if all the experimental subjects get jobs following the program, that might result more from a general increase in available jobs than from the value of the program itself.

There is no magic formula or set of guidelines for selecting measures of program context, any more than there is for choosing control variables in some other type of research. Just as we read what other researchers have found with respect to some topic we are interested in—say, explanatory research—we should also learn about the production process for some criminal justice program before conducting an evaluation.

Theory also plays an important role. If we study a program to provide job training so that participants can better compete for employment, we need to understand how labor markets operate in general. This is part of a theory-driven evaluation: Understanding how a program should work in theory will better enable researchers to specify measures of program contexts that should be considered (Weiss, 1995).

**Measuring Program Delivery** In addition to making measurements relevant to the outcomes of a program, it is necessary to measure the program intervention—the experimental stimulus or independent variable. In some cases, this measurement will be handled by assigning subjects to experimental and control groups, if that’s the research design. Assigning a person to the experimental group is the same as scoring that person “yes” on the intervention, and assignment to the control group represents a score of “no.” In practice, however, it’s seldom that simple.

Let’s continue with the job-training example. Some inmates will participate in the program through early release; others will not. But imagine for a moment what job-training programs are actually like. Some subjects will participate fully; others might miss sessions or fool around when they are present. So we may need measures of the extent or quality of participation in the program. And if the program is effective, we should find that those who participated fully have higher employment rates than those who participated less.

An evaluation of a drug court operating in Washington, DC, provides an example of uncommonly thorough attention to program delivery (Carver, Boyer, and Hickey, 1996). First, program staff kept attendance records for training activities and classes that participants were required to attend. In addition, evaluators recorded how attentive participants were and how actively they participated in discussions. Assuming that degree of participation in program activities would affect success rates, evaluators expected lower recidivism among drug users who were more attentive and active.

Other factors may further confound the administration of the experimental stimulus. Suppose we are evaluating a new form of counseling designed to cure drug addiction. Several counselors administer it to subjects composing an experimental group. We can compare the recovery rate of the experimental group with that of a control group (a group that received some other type of counseling or none at all). It might be useful to include the names of the counselors who treat specific subjects in the experimental group, because some may be more effective than others. If that turns out to be the case, we must find out why the treatment works better for some counselors than for others. What we learn will further elaborate our understanding of the therapy itself.

Michael Dennis (1990) describes an excellent example of the importance of measuring interventions in this type of study. Intravenous drug users were randomly assigned to receive enhanced treatment (the experimental stimulus) or standard treatment from counselors. Recognizing that some counselors might be more skilled than others, Dennis also randomly assigned counselors to provide either enhanced or standard treatments. There was still some potential for variation in counseling within
the enhanced and standard treatment groups, so Dennis tape-recorded sample sessions between patients and counselors. Research staff who were blind to the intended level of counseling then rated each recorded session according to whether they felt it represented enhanced or standard counseling.

Obtaining measures of the experimental intervention is very important for many types of evaluation designs. Variation in the levels of treatment delivered by a program can be a major threat to the validity of even randomized evaluation studies. Put another way, uncontrolled variation in treatment is equivalent to unreliable measurement of the independent variable.

**Specifying Other Variables** It is usually necessary to measure the population of subjects involved in the program being evaluated. In particular, it is important to define those for whom the program is appropriate. In evaluation studies, such persons are referred to as the program’s “target population.” If we are evaluating a program that combines more intensive probation supervision with periodic urine testing for drug use, it’s probably appropriate for convicted persons who are chronic users of illegal drugs, but how should we define and measure chronic drug use more specifically? The job-skills training program mentioned previously is probably appropriate for inmates who have poor employment histories, but a more specific definition of employment history is needed.

This process of definition and measurement has two aspects. First, the program target population must be specified. This is usually done in a manner similar to the process of defining program goals. Drawing on questions like those in Figure 12.2, evaluators consult program officials to identify the intended targets or beneficiaries of a particular program. Because the hypothetical urine testing program is combined with probation, its target population will include persons who might receive suspended sentences with probation. However, offenders convicted of crimes that carry non-suspendible sentences will not be in the target population. Prosecutors and other participants may specify additional limits to the target population—employment or no previous record of probation violations, for example.

Most evaluation studies that use individual people as units of analysis also measure such background variables as age, gender, educational attainment, employment history, and prior criminal record. Such measures are made to determine whether experimental programs work best for males, those over age 25, high school graduates, persons with fewer prior arrests, and so forth.

Second, in providing for the measurement of these different kinds of variables, we need to choose whether to create new measures or use ones already collected in the course of normal program operation. If our study addresses something that’s not routinely measured, the choice is easy. More commonly, at least some of the measures we are interested in will be represented in agency records in some form or other. We then have to decide whether agency measures are adequate for our evaluation purposes.

Because we are talking about measurement here, our decision to use our own measures or those produced by agencies should, of course, be based on an assessment of measurement reliability and validity. If we are evaluating the program that combined intensive probation with urinalysis, we will have more confidence in the reliability and validity of basic demographic information recorded by court personnel than in court records of drug use. In this case, we might want to obtain self-report measures of drug use and crime commission from subjects themselves, rather than relying on official records.

By now, it should be abundantly clear that measurement must be taken very seriously in evaluation research. Evaluation researchers must carefully determine all the variables to be measured and obtain appropriate measures
for each. However, such decisions are typically not purely scientific ones. Evaluation researchers often must work out their measurement strategy with the people responsible for the program being evaluated.

**Designs for Program Evaluation**

*Designs used in basic research are readily adapted for use in evaluation research.*

Chapter 7 presented a good introduction to a variety of experimental and other designs that researchers use in studying criminal justice. Recall that randomly assigning research subjects to experimental or control groups controls for many threats to internal validity. Here, our attention turns specifically to the use of different designs in program evaluation.

**Randomized Evaluation Designs**

To illustrate the advantages of random assignment, consider this dialogue from Lawrence Sherman’s book *Policing Domestic Violence: Experiments and Dilemmas* (1992:67):

> When the Minneapolis domestic violence experiment was in its final planning stage, some police officers asked: “Why does it have to be a randomized experiment? Why can’t you just follow up the people we arrest anyway, and compare their future violence risks to the people we don’t arrest?”

Since this question reveals the heart of the logic of controlled experiments, I said, “I’m glad you asked. What kind of people do you arrest now?” “Assholes,” they replied. “People who commit aggravated POPO.”

“What is aggravated POPO?” I asked. “Pissing off a police officer,” they answered. “Contempt of cop. But we also arrest people who look like they’re going to be violent, or who have caused more serious injuries.”

> “What kind of people do you not arrest for misdemeanor domestic assault?” I continued.

> “People who act calm and polite, who lost their temper but managed to get control of themselves,” came the answer.

> “And which kinds of people do you think would have higher risks of repeat violence in the future?” I returned.

> “The ones we arrest,” they said, the light dawning.

> “But does that mean arrest caused them to become more violent?” I pressed.

> “Of course not—we arrested them because they were more trouble in the first place,” they agreed.

> “So just following up the ones you arrest anyway wouldn’t tell us anything about the effects of arrest, would it?” was my final question.

> “Guess not,” they agreed. And they went on to perform the experiment.

Sherman’s dialogue portrays the obvious problems of selection bias in routine police procedures for handling domestic violence. In fact, one of the most important benefits of randomization is to avoid the selectivity that is such a fundamental part of criminal justice decision making. Police selectively arrest people, prosecutors selectively file charges, judges and juries selectively convict defendants, and offenders are selectively punished. In a more general sense, randomization is the great equalizer: Through probability theory, we can assume that groups created by random assignment will be statistically equivalent.

At the same time, randomized designs are not suitable for evaluating all experimental criminal justice programs. Certain requirements of randomized studies mean that this design cannot be used in many situations. A review of those requirements illustrates many of the limits of randomized designs for applied studies.
Program and Agency Acceptance  Random assignment of people to receive some especially desirable or punitive treatment may not be possible for legal, ethical, and practical reasons. We discussed ethics and legal issues in Chapter 3. Sometimes, practical obstacles may also be traced to a misunderstanding of the meaning of random assignment. It is crucial that public officials understand why randomization is desirable and that they fully endorse the procedure.

Richard Berk and associates (2003) describe how researchers obtained cooperation for an evaluation of a new inmate classification system in the California Department of Corrections (CDC) by appealing to the needs of agency managers. Preliminary research suggested that the experimental classification system would increase inmate and staff safety at lower cost than classification procedures then in use. In addition:

Plans for the study were thoroughly reviewed by stakeholders, including CDC administrators, representatives of prison employee bargaining unions, . . . California State legislative offices, and a wide variety of other interested parties. There was widespread agreement that the study was worth doing. (p. 211)

At the same time, justice agencies have expanding needs for evaluations of smaller-scale programs. John Eck (2002) explains how designs that are less elaborate, less costly, and less disruptive of routine operations are more likely to be accepted by public agencies.

Minimize Exceptions to Random Assignment  In any real-world delivery of alternative programs or treatments to victims, offenders, or criminal justice agency staff, exceptions to random assignment are all but inevitable. In a series of experiments on police responses to domestic violence, officers responded to incidents in one of three ways, according to a random assignment procedure (Sherman, 1992). The experimental treatment was arrest; control treatments included simply separating parties to the dispute or attempting to advise and mediate. Although patrol officers and police administrators accepted the random procedure, exceptions were made as warranted in individual cases, subject to an officer’s discretionary judgment.

As the number of exceptions to random assignment increases, however, the statistical equivalence of experimental and control groups is threatened. When police (or others) make exceptions to random assignment, they are introducing bias into the selection of experimental and control groups. Randomized experiments are best suited for programs in which such exceptions can be minimized. The prison classification study by Berk and associates offers a good example. Random assignment was automatic—inmates having odd identification numbers at intake were assigned to the treatment group; those having even numbers were in the control group. This procedure produced treatment and control groups that were virtually identical in size: 9662 treatment and 9656 controls (2003:224–25).

Adequate Case Flow for Sample Size  In Chapter 8, we examined the relationship between sample size and accuracy in estimating population characteristics. As sample size increases (up to a point), estimates of population means and standard errors become more precise. By the same token, the number of subjects in groups created through random assignment is related to the researcher’s ability to detect significant differences in outcome measures between groups. If each group has only a small number of subjects, statistical tests can detect only very large program effects or differences in outcome measures between the two groups. This is a problem with statistical conclusion validity and sample size, as we discussed in Chapters 7 and 8.
Case flow represents the process through which subjects are accumulated in experimental and control groups. In Sherman’s domestic violence evaluations, cases flowed into experimental and control groups as domestic violence incidents were reported to police. Evaluations of other types of programs will generate cases through other processes—for example, offenders sentenced by a court or inmates entering a correctional facility. In the drug court evaluation by Denise Gottfredson and associates, it took 18 months to accumulate 235 cases for the treatment group (drug court). Berk and associates accumulated their 9662 treatment subjects in 6 months.

If relatively few cases flow through some process and thereby become eligible for random assignment, it will take a longer time to obtain sufficient numbers of cases. The longer it takes to accumulate cases, the longer it will take to conduct an experiment and the longer experimental conditions must be maintained. Imagine filling the gas tank of your car with a small cup: It would take a long time, it would test your patience, and you would probably tarnish the paint with spilled gasoline as the ordeal dragged on. In a similar fashion, an inadequate flow of cases into experimental groups risks contaminating the experiment through other problems.

Getting information about case flow in the planning stages of an evaluation is a good way to diagnose possible problems with numbers of subjects. For example, Sherman (1992:293–95) conducted what he calls a “pipeline” study in Milwaukee to determine whether there were enough suitable domestic violence cases for random assignment to three treatment conditions.

**Maintaining Treatment Integrity** Treatment integrity refers to whether an experimental intervention is delivered as intended. Sometimes called “treatment consistency,” treatment integrity is therefore roughly equivalent to measurement reliability. Experimental designs in applied studies often suffer from problems related to treatment inconsistencies. If, for example, serving time in jail is the experimental treatment in a program designed to test different approaches to sentencing drunk drivers, treatment integrity will be threatened if some defendants are sentenced to a weekend in jail, whereas others serve 30 days or longer.

Criminal justice programs can vary considerably in the amount of treatment applied to different subjects in experimental groups. For example, Gottfredson and associates (2003) acknowledge that the drug court treatment in Baltimore County was unevenly implemented. Only about half of those assigned to the experimental group received certified drug treatment. In contrast, the classification system tested by Berk and associates was a relatively simple treatment that was readily standardized. There is no danger of treatment dilution as was the case in the drug court experiment.

Midstream changes in experimental programs can also threaten treatment integrity. Rossi and associates (1999:297) point out that the possibility of midstream changes means that randomized designs are usually not appropriate for evaluating programs in early stages of development, when such changes are more likely. For example, assume we are evaluating an intensive supervision probation program with randomized experimental and control groups. Midway through the experiment, program staff decides to require weekly urinalysis for everyone in the experimental group (those assigned to intensive supervision). If we detect differences in outcome measures between the experimental and control groups (say, arrests within a year after release), we will not know how much of the difference is due to intensive supervision and how much might be due to the midstream change of adding urine tests.

**Summing Up the Limits of Randomized Designs** Randomized experiments therefore require that certain conditions be met. Staff responsible for program delivery must accept random assignment and further agree to
minimize exceptions to randomization. Case flow must be adequate to produce enough subjects in each group so that statistical tests will be able to detect significant differences in outcome measures. Finally, experimental interventions must be consistently applied to treatment groups and withheld from control groups.

These conditions, and the problems that may result if they are not met, can be summarized as two overriding concerns in field experiments: (1) equivalence between experimental and control groups before an intervention and (2) the ability to detect differences in outcome measures after an intervention is introduced. If there are too many exceptions to random assignment, experimental and control groups may not be equivalent. If there are too few cases, or inconsistencies in administering a treatment, or treatment spillovers to control subjects, outcome measures may be affected in such a way that researchers cannot detect the effects of an intervention.

Although these conditions are related to features of experimental design, more often they are practical issues in experiments in natural settings. David Weisburd and associates (Weisburd, Petrosino, and Mason, 1993) tie many of these issues together by noting something of a paradox. In efforts to gain more precise estimates of treatment effects, researchers commonly increase the size of experimental treatment and control groups. But having larger groups makes experiments more difficult to manage. This, in turn, introduces treatment inconsistencies and other errors that negate the sought-after advantage of a larger sample size.

Let’s now look at an example that illustrates both the strengths of random experiments and constraints on their use in criminal justice program evaluations.

**Home Detention: Two Randomized Studies**

Terry Baumer and Robert Mendelsohn conducted two random experiments to evaluate programs that combine home detention with electronic monitoring (ELMO). In earlier chapters, we examined how different features of these studies illustrated measurement principles; here, our focus is on the mechanics of random assignment and program delivery.

In their first study, Baumer and Mendelsohn evaluated a program that targeted adult offenders convicted of nonviolent misdemeanor and minor felony offenses (Baumer and Mendelsohn, 1990; also summarized in Baumer, Maxfield, and Mendelsohn, 1993). The goal of the program was to provide supervision of offenders that was more enhanced than traditional probation but less restrictive and less costly than incarceration. Several measures of outcomes and program delivery were examined, as we have described in earlier chapters.

Baumer and Mendelsohn selected a randomized posttest-only design, in which the target population was offenders sentenced to probation. Subjects were randomly assigned to an experimental group in which the treatment was electronically monitored home detention or to a control group sentenced to home detention without electronic monitoring. Figure 12.3 summarizes case flow into the evaluation experiment. After a guilty plea or trial conviction, probation office staff reviewed offenders’ backgrounds and criminal records for the purpose of recommending an appropriate sentence. The next step was a hearing, at which sentences are imposed by a criminal court judge.

Persons sentenced to probation were eligible for inclusion in the experiment. Their case files were forwarded to staff in the community corrections agency responsible for administering the home detention programs. On receiving an eligible case file, community corrections staff telephoned the evaluation researchers, who, having prepared a random list of case numbers, assigned subjects to either the treatment or control group. Subject to two constraints, this process produced 78 treatment subjects and 76 control subjects.

Thinking back on our consideration of ethics in Chapter 3, you should be able to think of
one constraint: informed consent. Researchers and program staff explained the evaluation project to subjects and obtained their consent to participate in the experiment. Those who declined to participate in the evaluation study could nevertheless be assigned to home detention as a condition of their probation. The second constraint was made necessary by the technology of electronic monitoring: Subjects could not be kept in the treatment group if they did not have a telephone that could be connected to the electronic monitoring equipment.

Notice that random assignment was made after sentencing. Baumer and Mendelsohn began their evaluation by randomizing subjects between stages 2 and 3 in Figure 12.3. This produced problems because judges occasionally overruled presentence investigation recommendations to probation, thus overriding random assignment. After detecting this problem, Baumer and Mendelsohn (1990:27–29) moved randomization “downstream,” so that judicial decisions could not contaminate the selection process.

Baumer and Mendelsohn (1990:26) obtained agreement from community corrections staff, prosecutors, and judges to use random assignment by getting all parties to accept an assumption of “no difference”:

That is, in the absence of convincing evidence to the contrary, they were willing to assume that there was no difference between the . . . methods of monitoring. This allowed the prosecutor to negotiate and judges to assign home detention as a condition of probation only, while permitting the community corrections agency to make the monitoring decision.

Convinced of the importance of random assignment, the community corrections agency “delegated” to researchers the responsibility for making the monitoring decision, a “decision” that was randomized.
In this example, the experimental condition—electronic monitoring—was readily distinguished from the control condition, home detention without electronic monitoring. There was no possibility of treatment spillover; control subjects could not unintentionally receive some level of electronic monitoring because they had neither the bracelet nor the home-base unit that embodied the treatment. Electronic monitoring could therefore be readily delivered to subjects in the experimental group and withheld from control subjects. This treatment was not necessarily consistent, however.

The second ELMO evaluation conducted by Baumer and Mendelsohn reveals how program delivery problems can undermine the strengths of random assignment (Baumer, Maxfield, and Mendelsohn, 1993). In their study of juvenile burglars, they used similar procedures for randomization, but eligible subjects were placed in one of four groups, as illustrated in this table:

<table>
<thead>
<tr>
<th>Electronic Monitoring?</th>
<th>Police Visits?</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
<td>C</td>
<td>E1</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>E2</td>
<td>E3</td>
</tr>
</tbody>
</table>

Juvenile burglars could be randomly assigned to three possible treatments: electronic monitoring only (E1), police visits to their home after school only (E2), or electronic monitoring and police visits (E3). Subjects in the control group (C) were sentenced to home detention only. As in the adult study, outcome measures included arrests after release.

Although there were no problems with random assignment, inconsistencies in the delivery of each of the two experimental treatments produced uninterpretable results (Maxfield and Baumer, 1991:5):

Observations of day-to-day program operations revealed that, compared with the adult program, the juvenile court and cooperating agencies paid less attention to delivering program elements and using information from . . . the electronic monitoring equipment. Staff were less well-trained in operating the electronic monitoring equipment, and police visits were inconsistent.

The box titled “Home Detention” in Chapter 1 elaborates on differences in the operation of these two programs and a third ELMO program for pretrial defendants. However, the lesson from these studies bears repeating here: _Randomization does not control for variation in treatment integrity and program delivery._

Randomized experiments can be powerful tools in criminal justice program evaluations. At the same time, it is often impossible to maintain the desired level of control over experimental conditions. This is especially true for complex interventions that may change while an evaluation is under way. Experimental conditions are also difficult to maintain when different organizations work together in delivering some service—a community-based drug treatment provider coupled with intensive probation, for example.

Largely because of such problems, evaluation researchers are increasingly turning to other types of designs that are less “fragile”—less subject to problems if rigorous experimental conditions cannot be maintained.

### Quasi-Experimental Designs

Quasi-experiments are distinguished from “true” experiments by the lack of random assignment of subjects to an experimental and a control group. Random assignment of subjects is often impossible in criminal justice evaluations. Rather than forgo evaluation altogether in such instances, it is usually possible to create and execute research designs that will permit evaluation of the program in question.

Quasi-experiments may also be “nested” into experimental designs as backups should one or more of the requisites for a true experiment break down. For example, Michael
Dennis (1990) describes how a time-series design was nested into a series of random experiments to evaluate enhanced drug abuse counseling. In the event that case flow is inadequate or random assignment to enhanced or standard counseling regimes breaks down, the nested time-series design will salvage a quasi-experiment.

We considered different classes of quasi-experimental designs—nonequivalent groups, cohorts, and time series—in Chapter 7, together with examples of each type. Each of these designs has been used extensively in criminal justice evaluation research.

**Ex Post Evaluations** Often, a researcher or public official may decide to conduct an evaluation sometime after an experimental program has gone into effect. These so-called ex post evaluations (Rossi, Freeman, and Lipsey, 1999:312) are not usually amenable to random assignment after the fact. For example, if a new job-skills training program is introduced in a state correctional facility, an ex post evaluation might compare rates of employment among released inmates with similar outcome measures for a matched comparison institution. Or an interrupted time-series design might examine records of alcohol-related accidents before and after a new law that allows administrative suspension of driver’s licenses to take effect.

**Full-Coverage Programs** Interventions such as new national or statewide laws are examples of full-coverage programs in which it is not possible to identify subjects who are not exposed to the intervention, let alone randomly assign persons to receive or not receive the treatment. Quasi-experimental designs may be the strongest possible approach for evaluating such programs. Statewide sentencing guidelines and mandatory minimum sentences are common examples of full-coverage interventions.

**Larger Treatment Units** Similarly, some experimental interventions may be designed to affect all persons in some larger unit—a neighborhood crime prevention program, for example. It is not possible to randomly assign some neighborhoods to receive the intervention while withholding it from others. Nor is it possible to control which individuals in a neighborhood are exposed to the intervention.

Different types of quasi-experimental designs can be used in such cases. For example, in a Kansas City program to reduce gun violence, police targeted extra patrols at gun crime “hot spots” (Sherman, Shaw, and Rogan, 1995). Some beats were assigned to receive the extra patrols; comparison beats—selected for their similar frequency of gun crimes—did not get the special patrols. Several outcome measures were compared for the two types of areas. After 29 weeks, gun seizures in the target area increased by more than 65 percent and gun crimes dropped by 49 percent. There were no significant changes in either gun crimes or gun seizures in the comparison beat. Drive-by shootings dropped from 7 to 1 in the target area and increased from 6 to 12 in the comparison area. Homicides declined in the target area but not in the comparison area. Citizen surveys showed less fear of crime and more positive feelings about the neighborhood in the target area than in the comparison area.

**Nonequivalent-Groups Designs** As we saw in Chapter 7, quasi-experimental designs lack the built-in controls for selection bias and other threats to internal validity. Nonequivalent groups designs, by definition, cannot be assumed to include treatment and comparison subjects who are statistically equivalent. For this reason, quasi-experimental program evaluations must be carefully designed and analyzed to rule out possible validity problems.

For evaluation designs that use nonequivalent groups, attention should be devoted to constructing experimental and comparison groups that are as similar as possible on important variables that might account for differences in outcome measures. Rossi and associates (1999) caution that procedures for constructing such
groups should be grounded in a theoretical understanding of what individual and group characteristics might confound evaluation results. In a study of recidivism by participants in shock incarceration programs, for example, we certainly want to ensure that equal numbers of men and women are included in groups assigned to shock incarceration and groups that received some other sentence. Alternatively, we can restrict our analysis of program effects to only men or only women.

David Farrington and associates (1993) conducted a nonequivalent-groups quasi-experiment to evaluate different approaches to preventing shoplifting. The experiment was carried out in nine electronics stores operated in England by the same retail chain. Stores were placed in one of four groups, created by matching the stores on three characteristics: physical size, sales volume, and type of location. Three experimental interventions to prevent shoplifting were tested: (1) physical redesign of store layout to reduce opportunities for shoplifting, (2) electronic tagging of items so that they would trigger sensors at the exit to each shop, and (3) uniformed guards posted at store entrances. In addition, one store in three of the four groups served as a comparison site, where no new measures to prevent shoplifting were introduced.

The outcome measure was based on a system whereby labels were placed on certain small items thought to be frequently stolen: audiotapes, videotapes, stereo headphones, and photographic film. Each item was fixed with a sticky label that sales clerks were to remove whenever an item was sold. The total number of labeled items before a store opened for the day was known. Comparing this total with the number of labels accumulated by clerks and the number of items remaining in each store at day’s end yielded a count of missing items. Pretest measures were obtained for each store over a 3-day period. Two posttest measures were made at most sites, 1 week and 4 weeks after introducing prevention measures. Table 12.3 summarizes the grouping of stores with experimental treatments and the results of the evaluation. The figures in this table represent the number of missing tags as a percentage of all items that left the store (sold or stolen).

Notice first the different mixes of stores and experimental treatments. The four groupings of stores reflect within-group similarities and between-group differences in matching criteria. The experimental and control conditions across stores in different groups were mixed to compare prevention effectiveness with store characteristics.

Although we have omitted tests of statistical significance from Table 12.3, some interesting patterns are evident in the results. Electronic tagging (stores 2 and 3) reduced shoplifting in both the 1- and 4-week posttests. Store redesign was effective only in the first posttest (store 1); a second posttest showed an increase in shoplifting, which suggests that thieves learned to adapt to the new layout (only one posttest was conducted in the other redesign site, store 8). Uniformed guards (stores 4 and 7) had no

<table>
<thead>
<tr>
<th>Table 12.3</th>
<th>Situational Crime Prevention of Shoplifting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A</td>
<td></td>
</tr>
<tr>
<td>Store 1 (R)</td>
<td>36.5</td>
</tr>
<tr>
<td>Store 2 (T)</td>
<td>30.8</td>
</tr>
<tr>
<td>Group B</td>
<td></td>
</tr>
<tr>
<td>Store 3 (T)</td>
<td>17.3</td>
</tr>
<tr>
<td>Store 4 (G)</td>
<td>10.7</td>
</tr>
<tr>
<td>Store 5 (C)</td>
<td>15.3</td>
</tr>
<tr>
<td>Group C</td>
<td></td>
</tr>
<tr>
<td>Store 6 (C)</td>
<td>15.4</td>
</tr>
<tr>
<td>Store 7 (G)</td>
<td>6.9</td>
</tr>
<tr>
<td>Group D</td>
<td></td>
</tr>
<tr>
<td>Store 8 (R)</td>
<td>24.4</td>
</tr>
<tr>
<td>Store 9 (C)</td>
<td>13.6</td>
</tr>
</tbody>
</table>

Note: Experimental treatments: R = store redesign; T = electronic tagging; G = security guard; C = comparison. Source: Adapted from Farrington and associates (1993:108).
consistent impact on shoplifting. Shoplifting in the three comparison sites (stores 5, 6, and 9) either increased or declined less than it did in matched experimental sites.

Several features of this evaluation warrant comment. First, randomization was not possible because the experimental treatment was delivered to a small number of large units—electronics stores. Second, Farrington and associates used matching criteria that could reasonably be expected to affect shoplifting rates. A related point is that all sites were franchise stores owned by the same large chain, which minimized differences in inventory mixes, management procedures, and the like. Third, the three experimental conditions were mixed and tested in different groups of matched stores. This produced a stronger test of the relative effectiveness of different prevention measures.

Finally, consider for a moment the degree of cooperation required between researchers and evaluation clients. On the one hand, Farrington and colleagues required some effort from store staff in collecting and counting tagged items, not to mention implementing different crime prevention measures. On the other hand, you should also see the potential benefits to retail stores from this quasi-experiment to determine what works in preventing shoplifting.

**Time-Series Designs** Interrupted time-series designs require attention to different issues because researchers cannot normally control how reliably the experimental treatment is actually implemented. Foremost among these issues are instrumentation, history, and construct validity. In many interrupted time-series designs, conclusions about whether an intervention produced change in some outcome measure rely on simple indicators that represent complex causal processes.

In their evaluation of legislation to provide for mandatory minimum sentences in Oregon, Nancy Merritt and associates (Merritt, Fain, and Turner, 2006) examined changes in sentences for different types of offenses. They found that sentences for offenses clearly covered by the law did increase in the first five years after its passage. However, they also found declines in the number of cases filed that were clearly included in the mandatory provisions. Meanwhile, more charges were filed for offenses covered by discretionary provisions. Of course, criminal case prosecution and sentencing are complex processes. The authors could not directly control for different circumstances surrounding cases processed before and after the law took effect. However, their time-series analysis does clearly show changes in case filings, suggesting that prosecutors exercised discretion to evade the mandatory provisions of Oregon’s legislation.

Understanding the causal process that produces measures used in time-series analysis is crucial for interpreting results. Such understanding can come in two related ways. First, we should have a sound conceptual grasp of the underlying causal forces at work in the process we are interested in. Second, we should understand how the indicators used in any time-series analysis are produced.

Patricia Mayhew and associates (Mayhew, Clarke, and Elliott, 1989) concluded that laws requiring motorcycle riders to wear helmets produced a reduction in motorcycle theft. This might seem puzzling until we consider the causal constructs involved in stealing motorcycles. Assuming that most motorcycle thefts are crimes of opportunity, Mayhew and associates argue that few impulsive thieves stroll about carrying helmets. Even thieves are sufficiently rational to recognize that a helmetless motorcycle rider will be unacceptably conspicuous—an insight that deters them from stealing motorcycles. Mayhew and colleagues considered displacement as an alternative explanation for the decline in motorcycle theft, but they found no evidence that declines in motorcycle theft were accompanied by increases in either stolen cars or bicycles. By systematically thinking through the causal process of motorcycle theft, Mayhew and associates were able to
conclude that helmet laws were unintentionally effective in reducing theft.

Other Types of Evaluation Studies

Earlier in this chapter, we noted how process evaluations are distinct from impact assessments. Whereas the latter seek answers to questions about program effects, process evaluations monitor program implementation, asking whether programs are being delivered as intended.

Process evaluations can be invaluable aids in interpreting results from an impact assessment. We described how Baumer and Mendelsohn were better able to understand outcome measures in their evaluation of ELMO for juvenile burglars because they had monitored program delivery. Similarly, process evaluations were key elements of CCTV evaluations reported by Gill and Spriggs (2005). They were able to describe whether cameras were placed and monitored as intended. In many cases camera placement was modified, something the authors suggest was related to the relative success of different CCTV installations. Without a process evaluation, information about program implementation cannot be linked to outcome measures.

Process evaluations can also be useful for criminal justice officials whose responsibility centers more on the performance of particular tasks than on the overall success of some program. For example, police patrol officers are collectively responsible for public safety in their beat, but their routine actions focus more on performing specific tasks such as responding to a call for service or, in community policing, diagnosing the concerns of neighborhood residents. Police supervisors are attentive to traffic tickets written, arrests made, and complaints against individual officers. Probation and parole officers are, of course, interested in the ultimate performance of their clients, but they are also task oriented in their use of records to keep track of client contacts, attendance at substance abuse sessions, or job performance. Process evaluations center on measures of task performance—on the assumption that tasks are linked to program outcomes. So process evaluations can be valuable in their own right as well as important for diagnosing measures of program effects.

Problem Analysis and Scientific Realism

Problem analysis, coupled with scientific realism, helps public officials use research to select and assess alternative courses of action.

Program evaluation differs from problem analysis with respect to the time dimension and where each activity takes place in the policy process. Problem analysis is used to help design alternative courses of action and choose among them.

In reality, there is not much of a difference between these two types of applied research. Similar types of research methods are used to address problem analysis questions (What would happen? What should we do?) as are brought to bear on program evaluation questions (What did happen? What have we done?). Consider, for example, a definition of a similar approach, policy analysis, from a prominent text: “Attempting to bring modern science and technology to bear on society’s problems, policy analysis searches for feasible courses of action, generating information and marshaling evidence of the benefits and other consequences that would follow their adoption and implementation” (Quade, 1989:4). Except for the form of the verb—“would follow”—this is not too different from the way we defined program evaluation.

Results from program evaluations are frequently considered in choosing among future courses of action. Problem and policy analysis depend just as much on clearly specifying goals and objectives as does program evaluation. And the achievement of goals and objectives worked out through problem analysis can be tested through program evaluation. Measurement is also a fundamental concern in both types of applied studies.
Problem-Oriented Policing
More than an alternative approach to law enforcement, the core of problem-oriented policing is applying problem analysis methods to public safety problems. Problem-oriented policing depends on identifying problems, planning and taking appropriate action, then assessing whether those actions achieved intended results.

This approach centers on problems, not individual incidents. For example, traditional policing responds to reports of auto thefts, writing up details about the incident to support an insurance claim, then moving on to the next incident. Let’s consider this incident-oriented policing. In contrast, problem-oriented policing would begin by analyzing a number of auto theft reports. Reports would be examined for similarities, such as where and when they occurred, types of autos stolen, whether stolen cars were eventually recovered, and, if so, in what condition. Such analysis would define a more general problem of auto theft. Subsequent steps would consider what kinds of actions might be taken to address the problem.

Problem solving is a fundamental tool in problem-oriented policing. As described by Ronald Clarke and John Eck, problem solving involves four analytic steps:

(1) carefully define specific problems . . . ;
(2) conduct in-depth analysis to understand their causes; (3) undertake broad searches for solutions to remove these causes and bring about lasting reductions in problems; (4) evaluate how successful these activities have been. (2005:Step 7-1)

You can easily see how problem solving merges the application of problem analysis and evaluation (assessment) of the effects of interventions.

Problem-oriented policing is an especially useful example of applied research because a large number of resources are available. We’ll briefly describe three categories. Most of the first two categories have been prepared with support from Community Oriented Policing Services (COPS) Office in the U.S. Department of Justice. These are available at the Center for Problem-Oriented Policing website: http://www.popcenter.org.

How-to-Do-It Guides Ronald Clarke and John Eck (2005) have prepared a general guide to crime analysis to support problem-oriented policing. Adapted from a document originally prepared for the Jill Dando Institute of Crime Science in London, this publication offers succinct guidance on how to conduct problem analysis and report results. The COPS office has also sponsored guides that provide more detail on different problem analysis tools: assessment and evaluation (Eck, 2003); understanding the process of repeat victimization (Weisel, 2005); conducting background research on problems (Clarke and Schultz, 2005); interviewing offenders (Decker, 2005); collaborating with private-sector interests to solve problems (Chamard, 2006).


Problem and Response Guides In an earlier chapter we mentioned that justice agencies frequently adopt programs that appear to have been successful in other jurisdictions. Although this can sometimes be desirable, a key principle of problem-oriented policing is to base local actions on an understanding of local problems. Instead of trying an off-the-shelf program, or so-called “best practice,”
appropriate interventions should be considered only after analyzing data.

This principle is evident in two series of guides that describe what is known about effective responses based on past experience. Problem guides describe how to analyze very specific types of problems (for example, “Financial Crimes Against the Elderly”) and what are known to be effective or ineffective responses. Response guides describe very general kinds of actions that might be undertaken to address different types of problems (for example, “Video Surveillance of Public Places”).

Case Studies and Other Research One of the hallmarks of applied research is to use research to change practice. The two groups of guides discussed so far were prepared for use by criminal justice professionals, but they were developed following many years of research. Many examples of research that contributed to changes in justice policy have been published in the series Crime Prevention Studies. We now turn to an example that illustrates the application of problem analysis, as well as other research principles presented in this and earlier chapters.

Auto Theft in Chula Vista

Chula Vista is a medium-sized city of just under 200,000 residents, bordered by the Pacific Ocean on the west, and sandwiched by San Diego on the north and south; the city is about seven miles north of the U.S.-Mexico border. Nanci Plouffe and Rana Sampson (2004) began their analysis of vehicle theft by comparing Chula Vista to other southern California cities. After noting that theft rates tended to be higher for cities closer to the border, they began to disaggregate the problem by searching for areas where vehicle thefts and break-ins were concentrated. Deborah Weisel (2003) refers to this as “parsing,” or breaking down a large-area measure to examine smaller areas.

Plouffe and Sampson first determined that 10 parking lots accounted for 25 percent of thefts and 20 percent of break-ins in the city. Furthermore, 6 of those 10 lots were also among the top 10 calls-for-service locations in Chula Vista. This meant that auto theft hot spots also tended to be hot spots for other kinds of incidents. Their analysis found other notable patterns:

• Recovery rates for stolen cars and trucks were lower in Chula Vista than in areas to the north;
• recovery rates in 4 of the 10 hot parking lots were especially low, under 40 percent;
• smaller pickup trucks and older Toyota Camrys had even lower recovery rates;
• high-risk lots were close to roads that led to the Mexico border.

Together these findings suggested many cars stolen from the high-risk areas were being driven into Mexico.

Plouffe and Sampson next moved beyond using existing data from police records. This is again consistent with the methods of problem analysis: use existing data to identify problems and their general features, then collect additional data to better understand the mechanisms of problems. For Plouffe and Sampson that meant conducting environmental surveys of high-risk parking lots; observing operations and interviewing officials at U.S.-Mexico border crossings; and interviewing a small number of individuals arrested for auto theft from target lots. They sought to understand why particular lots were targeted, and whether stolen cars could be easily driven into Mexico.

We described environmental surveys in Chapter 10. In conducting theirs, Plouffe and Sampson discovered that the highest-risk lot was a two-minute drive from vehicle entry points into Mexico. The lot served a midrange general shopping mall with typical open parking. Access was easy and thieves could expect that vehicles would be unguarded for some time. Information gathered from the border crossing confirmed that few cars entering
In-person interviews with auto thieves used a 93-item questionnaire, asking about target selection, techniques, and other routines. Thieves preferred older cars because they could be easily stolen—steering column locks wear out and can be broken with simple tools. They watched people entering stores, judged that their vehicle would be unguarded for a time, and drove the few minutes into Mexico. Cars were rarely stolen from parking garages because thieves would have to produce a ticket in order to exit.

With this and other information, Plouffe and Sampson discussed strategies with Chula Vista police and security staff at parking lots and shopping malls. More diligent screening at the border was rejected, largely because most vehicles had been driven into Mexico before the theft was even discovered. They recommended that high-risk shopping malls install gates at entrance and exit points for parking lots. Drivers would take a ticket upon entering and would have to produce it when leaving. This, it was argued, would substantially increase the effort required to steal vehicles from parking lots near the border.

**Other Applications of Problem Analysis**

Partly because it has proved helpful in law enforcement applications, problem solving is being adopted by other criminal justice agencies. For example, Veronica Coleman and associates (Coleman, Holton, Olson, Robinson, and Stewart, 1999) describe how local and federal prosecutors in several U.S. cities have formed planning teams to identify crime problems and develop appropriate interventions. Teams include U.S. attorneys, researchers, and other criminal justice professionals who pursue a form of problem analysis labeled Strategic Approaches to Community Safety Initiatives (SACSI). SACSI involves five steps, four of which should look familiar (Coleman et al., 1999:18):

1. Form an interagency working group.
2. Gather information and data about a local crime problem.
3. Design a strategic intervention to tackle the problem.
4. Implement the intervention.
5. Assess and modify the strategy as the data reveal effects.

We have only scratched the surface of problem analysis applications in criminal justice. This is an area of applied research that is growing daily. Other examples draw on methods of systems analysis, operations research, and economics for such purposes as cost–benefit studies, police patrol allocation, and decisions about hiring probation officers. Cost–benefit analysis, in particular, is used to assess the relative value and expense of alternative policies. Although the mathematical tools that form the base of problem analysis can be sophisticated, the underlying logic is relatively simple. For example, police departments traditionally used pin maps to represent the spatial and temporal concentrations of reported crime.

**Space- and Time-Based Analysis**

Pin maps are examples of “low-tech” problem analysis that are nonetheless conceptually identical to computer models of hot spots used in many departments to plan police deployment. Growing numbers of justice agencies, especially police and sheriff’s departments, have taken advantage of rapid advances in computing and telecommunications. Computerized mapping systems now permit police to monitor changes in crime patterns on a daily or hourly basis and to develop responses accordingly. Furthermore, simultaneous advances in computing power and declines in the cost of that power make it possible for even small agencies to use mapping tools (Harries, 1999). The ongoing technological advances in mapping...
have fueled the application of statistical models to geographic clusters of crime problems. Thomas Rich (1999) describes this as analytic mapping, whereby statistical tools supplement the “eyeballing” approach to locating concentrations of crime.

Crime maps usually represent at least four different things: (1) one or more crime types; (2) space or area; (3) some time period; and (4) some dimension of land use, usually streets. The most useful crime maps will show patterns that can help analysts and police decide what sort of action to take. That’s part of applied research. An example will illustrate some basic features of crime maps.

Figure 12.4 shows four crime maps prepared by Shuryo Fujita, a graduate student at the Rutgers University School of Criminal Justice, for a midsized city in the northeast United States. All four maps show completed auto theft, but for different areas and time periods. The map in Panel A shows auto thefts for the year 2005 in one of four police precincts in the city. About 1750 completed thefts are represented, about 33 percent of all thefts in the city. You will probably notice two things about
Panel A. First, car theft seems to be everywhere in this area, except for blank spots in the center, and to the right side of the map—a large park and a river, respectively. Second, because car theft seems to be everywhere, the map is not especially useful. Much of the district appears to be a hot spot. Panel B changes the time reference, showing the 30 car thefts that occurred in the first week of August 2005. You might think this is somewhat more useful, showing more theft in the southern part of the district. But whereas Panel A shows too much, Panel B shows smaller numbers that don’t seem to cluster very much.

Panel C shifts the geographic focus to one sector within the district, to the left of the park. This sector happens to have the highest volume of car theft, 464 completed thefts in 2005; it’s the hottest sector in the hottest precinct in the city. Again, car theft seems to be all over the sector. A closer look shows more dots on the longer north–south streets than on cross streets. This is clearer in Panel D, which shows a crime density map of the sector. Crime density is a numerical value showing how close some dots are to each other, and how distant those clusters are from outlying dots. These values are mapped, showing patterns much more clearly than simple dots. The darker areas of Panel D represent more dense concentrations of car theft. There seem to be two corridors of car theft, running north–south below the diagonal street that bisects the map. These corridors are sort of connected in the middle, showing a rough H-shape. This shape happens to correspond with some major thoroughfares in the area. You might be able to imagine cruising up, across, and down, looking for cars to steal. That’s useful information a crime analyst can provide for police managers. During the summer months of 2006, police in this city deployed special patrols on the streets within the H-shaped area depicted in Panel D.

Tools for mapping crime and other problems are similar to the tools of statistical analysis, a topic we consider in the final chapter. Maps and statistics are most useful when we seek to understand patterns in a large number of observations. Very small police departments that report very few incidents need neither statistical nor geographic analysis to understand crime problems. But departments serving cities like the one in Figure 12.4 can really benefit from space-based analytic tools like crime mapping and density analysis.

Computerized crime mapping has been used for many years in a small number of departments, and is spreading to many large and midsized cities. Software is more powerful, and web-based mapping programs have been used to make crime maps generally available. At the same time, more published guides are appearing that describe how to combine maps with other analysis programs and sources of data. Jerry Ratcliffe (2004) describes how to classify crime concentrations across space and time dimensions to produce a hot spot matrix.

Crime mapping and other types of problem analysis illustrate another advantage of incident-based data—the potential for use in the kind of problem analysis we have described. Most crime mapping and similar tools are developed and used by individual departments, reflecting the fact that crime analysis is based on locally generated data. With incident-based reporting, crime analysis can be conducted on larger units. For example, Donald Fagiani and Colleen McLaughlin (1999) describe how National Incident Based Reporting System (NIBRS) data can show state or regional patterns in drug arrests and offenses. Using NIBRS data for Virginia, the authors demonstrate differences in types of arrests and drugs for different areas of the state.

**Scientific Realism and Applied Research**

Traditional research and evaluation are based on the model of cause and effect we considered in Chapters 4 and 7. An independent variable (cause) produces some change in a dependent variable (effect). Experimental and
quasi-experimental designs seek to isolate this causal process from the possible effects of intervening variables. Designs thus try to control for the possible effects of intervening variables.

Problem analysis as we have described it represents a bridge between traditional research approaches and applied research that is the foundation of scientific realism. Ray Pawson and Nick Tilley (1997) propose that, instead of trying to explain cause in a traditional sense, evaluators should search for mechanisms acting in context to explain outcomes. As we have seen, experiments do this by producing pretest statistical equivalence between groups of subjects who receive an intervention and groups of subjects who do not. Quasi-experiments using nonequivalent groups seek to control intervening variables by, for example, holding possible intervening variables constant. So, for example, if we believe that employment status might be an intervening variable in the relationship between arrest and subsequent domestic violence, we will try to structure an evaluation to hold employment status constant between treatment and control groups.

Scientific realism treats employment status as the context in which an arrest mechanism operates on the outcome of repeat domestic violence. Rather than try to control for employment status, a scientific realist will study the mechanism in context and conclude, for example, that arrest is effective in reducing subsequent violence in situations in which an offender is employed but is not effective when the offender is unemployed. This finding will be no different from what Sherman (1992) concludes in his assessment of a series of randomized experiments.

What is different is that the scientific realist approach is rooted in the principle that similar interventions can naturally be expected to have different outcomes in different contexts. Most notably, this approach is more compatible with the realities of evaluation than is the experimental approach. Pawson and Tilley (1997:81) put it this way: “Ultimately, realist evaluation would be mechanism- and context-driven rather than program-led” (emphasis in original). This means that interventions should be designed not so much as comprehensive programs that apply equally in all situations. Instead, interventions should be developed for specific contexts, and evaluations of those interventions must consider context as a key factor in whether the intervention achieves the desired outcome.

Situational crime prevention (Clarke, 1997) is an example of the scientific realist approach that bridges problem analysis and evaluation because it focuses on what mechanisms operate for highly specific types of crime in specific situations. So, for example, rather than develop and evaluate large-scale programs intended to reduce auto theft generally, situational crime prevention seeks specific interventions that will be effective in reducing particular types of auto theft. Ronald Clarke and Patricia Harris (1992) distinguish several types of auto theft by their purposes: joyriding, temporary transportation, resale or stripping, or insurance fraud. Theft of certain models for joyriding may be reduced by modest increases in security, whereas theft of expensive cars for resale or export requires different approaches. And many types of auto theft can be reduced by placing attendants at the exits of parking garages, but car break-ins may not be affected by that intervention.

As we mentioned in Chapter 7, the realist approach resembles a case-study approach. Both are variable-oriented strategies for research—they depend on measures of many variables to understand and assess a small number of cases. Detailed data and information are gathered about specific interventions, often in very small areas. Whereas an experimental evaluation uses probability theory to control for intervening variables, the case-study approach depends on detailed knowledge to understand the context in which mechanisms operate.

In his discussion of applied research tools for problem solving, Eck (2002, 2003) makes the case even more strongly. Public officials,
he argues, are more interested in solving local problems than in identifying robust cause-and-effect relationships. Both problem solving and evaluation are concerned with answering the question “Did the problem decline?” But eliminating alternative explanations for a decline, which is the central concern of internal validity and the rationale for stronger evaluation designs, is important only if officials wish to use the same intervention elsewhere.

In what Eck terms “small-claim, small-area problem solving,” analysts develop appropri-
ate interventions for problems in context. This is the essence of the problem-solving process. Like Eck, we emphasize process—systematically studying a problem, developing appropriate interventions, and seeing if those interventions have the intended effect. This is quite different from what Eck terms “large-claim interventions”—such as Drug Abuse Resistance Education (D.A.R.E.) or corrections boot camps—that are developed to apply in a wide variety of settings. Because small-claim, small-scale interventions are tailored to highly specific settings,
they cannot easily be transferred intact to different settings. However, the process of diagnosing local problems, selecting appropriate interventions, and then assessing the effects of those interventions can be generally applied. Anthony Braga (2002) offers more examples of this reasoning. Gloria Laycock presents an even stronger case for scientific realism in applied criminal justice research generally (2002) and in making specific plans for crime prevention (Tilley and Laycock, 2002).

We might rightly be skeptical about the generalizability of small-scale evaluations, whether case studies or something else. But the scientific realists respond that generalizability is not really the point. Instead, they urge combining results from a large number of smaller-scale evaluations as an alternative to trying to generalize from a small number of large-scale evaluations. For example, in a series of studies, Terry Baumer and associates (Baumer, Maxfield, and Mendelsohn, 1993) found that home detention with electronic monitoring achieved desired outcomes in some situations, but not in others. Aggregating across the three populations studied, the same intervention delivered in the same city produced different results for convicted adults, convicted juveniles, and pretrial adults. Similarly, Paul Ekblom and Ken Pease (1995:625) describe how an intervention to reduce repeat victimization among domestic violence victims appeared to have no effect when postintervention counts of domestic violence incidents were examined. Breaking the simple measure of domestic violence incidents down into subsets of repeat victimizations and new victimizations showed that counts of repeat victims were substantially lower, but this was offset by an increase in first-time victimizations. This makes sense if we think through the theory of action implied by this program.

Randomized or quasi-experimental evaluations should be conducted when such designs are appropriate. But it is important to recognize the formidable requirements for deploying these designs. The scientific realist approach to evaluation is flexible and may be appropriate in many situations. A scientific realist evaluation or case study can be especially useful in smaller-scale evaluations in which interest centers on solving some particular problem in a specific context more than on finding generalizable scientific truths. In any case, a variety of approaches can satisfy the definition of program evaluation we discussed early this chapter, by systematically applying social science research procedures to an individual program or agency.

Our general advice in this regard is simple: Do the best you can. This requires two things: (1) understanding the strengths and limits of social science research procedures and (2) carefully diagnosing what is needed and what is possible in a particular application. Only by understanding possible methods and program constraints can we properly judge whether any kind of evaluation study is worth undertaking with an experimental, quasi-experimental, or nonexperimental design, or whether an evaluation should not be undertaken at all.

The Political Context of Applied Research

Public policy involves making choices, and that involves politics.

Applied researchers bridge the gap between the body of research knowledge about crime and the practical needs of criminal justice professionals—a process that has potential political, ideological, and ethical problems. In the final section of this chapter, we turn our attention to the context of applied research, describing some of the special problems that can emerge in such studies.

Some similarities are evident between this material and our discussion of ethics in Chapter 3. Although ethics and politics are often closely intertwined, the ethics of criminal justice research focuses on the methods used, whereas political issues are more concerned
with the substance and use of research findings. Ethical and political aspects of applied research also differ in that there are no formal codes of accepted political conduct comparable to the codes of ethical conduct we examined earlier. Although some ethical norms have political aspects—for example, not harming subjects relates to protection of civil liberties—no one has developed a set of political norms that can be agreed on by all criminal justice researchers.

**Evaluation and Stakeholders**

Most applied studies involve multiple stakeholders—people who have some direct or indirect interest in the program or evaluation results (Rossi, Freeman, and Lipsey, 1999:204–5). Some stakeholders may be enthusiastic supporters of an experimental program, others may oppose it, and still others may be neutral. Different stakeholder interests in programs can produce conflicting perspectives on evaluations of those programs.

For example, in their study of pretrial home detention, Maxfield and Baumer (1992) found support for the program in the prosecutor’s office and sheriff’s department. Each was pleased by the prospect of freeing up jail space. However, decision makers in the community corrections agency, responsible for delivering home detention, were less supportive, expressing concern over the increased workload and fear that too many persons released to pretrial home detention would be bad risks. And some community corrections staff were more worried about the evaluation than the program, feeling that their job performance was under scrutiny. The National Institute of Justice funded the evaluation, hoping that results would document a successful program that could be adopted in other jurisdictions. Community corrections decision makers had a different perspective on the evaluation—they wanted to know what did and did not work.

Emil Posavec and Raymond Carey (1997:36–40) describe such problems as dysfunctional attitudes toward program evaluation. Program supporters may have unrealistic expectations that evaluation results will document dramatic success. Conversely, they may worry that negative results will lead to program termination. Agency staff may feel that day-to-day experience in delivering a program imparts a qualitative understanding of its success that cannot be documented by a controlled experiment. Staff and other stakeholders may object that an evaluation consumes scarce resources better spent on actually delivering a program.

We have two bits of advice in dealing with such problems. First, identify program stakeholders, their perspectives on the program, and their likely perspectives on the evaluation. In addition to agency decision makers and staff, stakeholders include program beneficiaries and competitors. For example, store owners in a downtown shopping district might benefit from an experimental program to deploy additional police on foot patrol, whereas people who live in a nearby residential area might argue that additional police should be assigned to their neighborhood.

Second, educate stakeholders about why an evaluation should be conducted. This is best done by explaining that applied research is conducted to determine what works and what does not. Various offices in the Department of Justice have issued brief documents that describe how evaluation can benefit criminal justice agencies by rationalizing their actions (Eck, 2003; Kirchner, Przybylski, and Cardella, 1994; Maxfield, 2001). Such publications, together with examples of completed evaluations, can be valuable tools for winning the support of stakeholders.

For an excellent description of the political and logistical problems involved in a complex evaluation that affects multiple stakeholders, see the appendix to Sherman’s (1992) book on domestic violence experiments. He presents a detailed report on the process of planning one experiment in Milwaukee, from initial negotiations through project completion.
More generally, recognize that applied research is very much a cooperative venture. Accordingly, researchers and program staff are mutual stakeholders in designing and executing evaluations. Evaluators’ interest in a strong design that will meet scientific standards must be balanced against the main concern of program sponsors—obtaining information that is useful for developing public policy. Among other things, the existence of mutual stakeholders’ perspectives implies that applied researchers have obligations to program sponsors, their collaborative partners in evaluation studies.

The flip side of being cautious about getting caught in stakeholder conflict is the benefit of applied research in influencing public policy. Evaluation studies can provide support for continuing or expanding successful criminal justice programs, or evidence that ineffective programs should be modified or terminated. And problem analysis results can sometimes be used to influence actions by public officials. For an example, see the box titled “When Politics Accommodates Facts,” in which Tony Fabelo describes how problem analysis dissuaded Texas legislators from costly lawmaking.

**Politics and Objectivity**

Politics and ideology can color research in ways even more subtle than those described by Fabelo. You may, for example, consider yourself an open-minded and unbiased person who aspires to be an objective criminal justice researcher. However, you may have strong views about different sentencing policies, believing that probation and restitution are to be preferred over long prison sentences. Because there is no conclusive evidence to favor one approach over the other, your beliefs would be perfectly reasonable.

Now, assume that one of the requirements for the course you are taking is to write a proposal for an evaluation project on corrections policy. In all likelihood, you will prepare a proposal to study a probation program rather than, say, a program on the use of portable jails to provide increased detention capacity. That is natural, and certainly legitimate, but your own policy preferences will affect the topic you chose.

Or let’s say that you deplore racism and racial discrimination in any form. Nonetheless, you know from other courses you have taken that African Americans are disproportionately arrested and imprisoned. You also know that there is some evidence of higher crime-commission rates among African Americans than among whites. You will probably reject an explanation that suggests genetic or other biological factors as the causes, believing that one race is not genetically superior to another. But how might you feel about a research project that looks for relationships between genetics and crime? Because genetic differences between races do exist, does this mean that such research is racist? Probably not, as we have described it here, but you should know that in the past decade the U.S. National Institutes of Health canceled conferences to discuss research on the link between genetics and crime because of the racial implications of such research.

Ronald Clarke (1997:28) describes political objections to applied studies of situational crime prevention: “Conservative politicians regard it as an irrelevant response to the breakdown in morality that has fueled the postwar rise in crime. Those on the left criticize it for neglecting issues of social justice and for being too accepting of the definitions of crime of those in power.” By the same token, electronic monitoring is distrusted for being simultaneously too lenient by allowing offenders to do time at home and too close to a technological nightmare by enabling the government to spy on individuals. Evaluations of situational crime prevention or ELMO programs may be criticized for tacitly supporting either soft-on-crime or heavy-handed police-state ideologies (Lilly, 2006; Nellis, 2006).

It is difficult to claim that criminal justice research, either applied or basic, is value free. Our own beliefs and preferences affect the topics we choose to investigate. Political
preferences and ideology may also influence criminal justice research agendas by making funds available for some projects but not others. For example, in 2004, the National Institute of Justice awarded money for projects to study these topics: “Chinese Connection: Changing Patterns of Drug Trafficking in the Golden Triangle” and “Assessment of Risk Factors Associated With Sexual Violence in the Texas Prison System.” No funds were awarded,
however, for such projects as “The Scope of Institutionalized Racism in the War on Drugs,” or “Exploratory Research on Torture in Federal Detention Camps.” It is, of course, possible for researchers—consciously or unconsciously—to become instruments for achieving political or policy objectives in applied research.

It may sometimes seem difficult to maintain an acceptable level of objectivity about or distance from evaluation results in criminal justice research. This task can be further complicated if you have strong views one way or another about a particular program or policy. Researchers who evaluate, say, an experimental program to prevent offenders from repeating probably sincerely hope that the program will work. However, substantially less consensus exists about other criminal justice problems and policies. For example, how do you feel about a project to test the effects of restrictive handgun laws or mandatory jail sentences for abortion protesters? We conclude this chapter with one final example that we expect will make you think about some of the political issues involved in applied research.

In 1990, the elected prosecutor of Marion County, Indiana—in which Indianapolis is located—was sharply criticized in a series of newspaper stories that claimed to present evidence of racial disparity in drug sentences handed down in the county. Convicted minority offenders, it was asserted, received longer prison terms than white offenders. The prosecutor immediately responded, criticizing the data collected and methods used by the investigative reporter. He also contacted Maxfield and asked him to conduct an independent analysis of drug cases accepted for prosecution.

In the first place, the prosecutor claimed, he had had previous feuds with the author of the newspaper stories. Second, he categorically denied any discriminatory policies in making sentence requests in drug cases. Third, he said he knew that the data and methods reported in the newspaper stories were deficient even though the reporter would not reveal details about his sources and information. Finally, if any pattern of racial disparity existed, it was certainly inadvertent, and the prosecutor wanted to know about it so that the problem could be fixed. Maxfield accepted the project and was paid to produce a report.

How do you feel about this example? Did Maxfield sell out? How would you feel if Maxfield turned up clear evidence of disparity in sentences? Or no evidence of disparity? What about political party affiliation—would it make a difference if the prosecutor and Maxfield identified with the same party? With different parties?

**Main Points**

- Evaluation research and problem analysis are examples of applied research in criminal justice.
- Different types of evaluation activities correspond to different stages in the policy process—policy planning, process evaluation, and impact evaluation.
- An evaluability assessment may be undertaken as a scouting operation or a preevaluation to determine whether it is possible to evaluate a particular program.
- A careful formulation of the problem, including relevant measurements and criteria of success or failure, is essential in evaluation research.
- Organizations may not have clear statements or ideas about program goals. In such cases, researchers must work with agency staff to formulate mutually acceptable statements of goals before proceeding.
- Evaluation research may use experimental, quasi-experimental, or nonexperimental designs. As in studies with other research purposes, designs that offer the greatest control over experimental conditions are usually preferred.
- The use of randomized field experiments requires careful attention to random assignment, case flow, and treatment integrity.
- Randomized designs cannot be used for evaluations that begin after a new program has been implemented or for full-coverage programs in which it is not possible to withhold an experimental treatment from a control group.
• Process evaluations can be undertaken independently or in connection with an impact assessment. Process evaluations are all but essential for interpreting results from an impact assessment.
• Problem analysis is more of a planning technique. However, problem analysis draws on the same social science research methods used in program evaluation. Many variations on problem analysis are used in applied criminal justice research.
• The scientific realist approach to applied research focuses on mechanisms in context, rather than generalizable causal processes.
• Criminal justice agencies are increasingly using problem analysis tools for tactical and strategic planning. Crime mapping and other space-based procedures are especially useful applied techniques.
• Problem solving, evaluation, and scientific realism have many common elements.
• Evaluation research entails special logistical, ethical, and political problems because it is embedded in the day-to-day events of public policy and real life.

✪ Key Terms

Evaluation research  Problem solving
Evidence-based policy  Process evaluation
Impact assessment  Stakeholders
Problem analysis

✪ Review Questions and Exercises

1. In presentations to justice practitioners, Maxfield describes evaluation as answering two questions: “Did you get what you expected?” and “Compared to what?” Discuss how particular sections of this chapter relate to those two questions.
2. When programs do not achieve their expected results, it’s due to one of two things: The program was not a good idea to begin with, or it was a good idea but was not implemented properly. Discuss why it is necessary to conduct both a process and an impact evaluation to learn why a program failed.
3. What are the principal advantages and disadvantages of randomized designs for field experiments? Are such designs used in problem analysis? Explain your answer.

✪ Online Sessions for Chapter 12

Online Sessions present further examples of selected material covered in each chapter. Your instructor may choose to use these for in-class discussion (as I do with my students) or you can access them at home. To ensure that you always have access to live, correct links for the websites described here, we provide the necessary links on the companion website for Research Methods for Criminal Justice and Criminology at http://www.thomsonedu.com/criminaljustice/maxfield. Once at the companion website, select this specific chapter, click on “Chapter Resources,” then click on “Web Links.”

Promoting Evaluation  Many state and federal agencies support program evaluation. Some have internal staff evaluators; others work through grants and contracts with outsider researchers. Other organizations compile various resources for program evaluation. One good example is the Centers for Disease Control (CDC) Evaluation Working Group. Start by considering the evaluation framework that presents an overview of the approach to evaluation. Then review the listed steps in preparing to do an evaluation. You will find some differences between this approach and what we have described. Evaluation is a framework, not a precise list of steps and tools.

Similar kinds of resources are available from the Juvenile Justice Evaluation Center, maintained by the Justice Research and Statistics Association. Information is geared to public agencies and service providers who may be required to include an evaluation component as part of any new program. You will find examples of juvenile justice evaluations as well as links to a variety of useful information on how to design and conduct evaluations.

http://www.cdc.gov/eval/framework.htm
http://www.jrsa.org/jjec/resources/index.html

Crime Prevention as Applied Research

Crime prevention is much more prominent in the United Kingdom than in the United States. Because crime prevention depends extensively on problem analysis, many tools for planning and evaluation are available from government sites. The Crime Reduction Group in the Home Office has many “toolkits” available for different types of crimes. These are similar to the problem-oriented policing guides we have mentioned ear-
lier. The “Audits and Strategies Toolkit” is especially relevant to this chapter on applied research. It’s intended more for public officials and service providers than for researchers. http://www.crimereduction.gov.uk/aud00.htm

**More on Problem-Oriented Policing** In earlier chapters we have seen many of the resources and tools available from the Center for Problem-Oriented Policing. There is no better example of applied research. In addition to problem and tool guides, the site offers interesting interactive modules for problem analysis. http://popcenter.org/learning.htm

**Crime Mapping** The National Institute of Justice has supported crime mapping for many years through its Center for Mapping Analysis for Public Safety (MAPS). The site includes links to several publications on mapping applications and techniques. Other resources include lists of cities and states that have online mapping tools. You will learn much about the potential uses of crime mapping by exploring some of these sites. http://www.ojp.usdoj.gov/nij/maps/

**Crime Maps for the People in Chicago** One of the most advanced crime mapping sites available to the public is the Chicago Police Citizen ICAM (Information Collection for Automated Mapping) site. Based on data made available to the public by the Chicago Police Department, the site supports mapping by address, police beat, or public school. The last is especially interesting. A scroll-down menu lists all schools in Chicago, organized by level. You can then display reported crimes for a period ranging from 1 to 14 days.

Another URL for Chicago crime maps offers even more listing options, including zip code, ward (a local political district), and route. Let’s say you want to plan a route to walk from your Chicago apartment to work. At this site you can trace different routes down different streets and reported crimes along the routes will be displayed. Thus empowered, you can select the safest route.

- http://12.17.79.6/
- http://www.mcwetboy.net/maproom/2005/05/chicago_crime_maps.phtml

**Additional Readings**


Rossi, Peter H., Freeman, Howard E., and Lipsey, Mark W., *Evaluation: A Systematic Approach*, 6th ed. (Thousand Oaks, CA: Sage, 1999). Of the many available “handbooks” on evaluation methods, this is the most widely read. Although the book is uneven in its coverage of recent developments, the authors provide a good general foundation in evaluation methods.

Tilley, Nick (ed.), *Analysis for Crime Prevention* (Crime Prevention Studies, Vol. 13, Monsey, NY: Criminal Justice Press, 2002); *Evaluation for Crime Prevention* (Crime Prevention Studies, Vol. 14, Monsey, NY: Criminal Justice Press, 2002). These companion volumes present innovative thinking about how problem analysis and program evaluation can be used by public officials in preventing crime. Some of the articles will be controversial. All are interesting and mostly fun to read.
Chapter 13

Interpreting Data

We’ll examine a few simple statistics frequently used in criminal justice research. We’ll also cover the fundamental logic of multivariate analysis. You’ll come away from this chapter able to perform a number of simple, though powerful, analyses to describe data and reach research conclusions.

Introduction 396
Univariate Description 396
Distributions 396
Measures of Central Tendency 397
Measures of Dispersion 398
Comparing Measures of Dispersion and Central Tendency 401
Computing Rates 403
Detail Versus Manageability 404
Describing Two or More Variables 405
Bivariate Analysis 405
Multivariate Analysis 408

Measures of Association 413

Inferential Statistics 418
Univariate Inferences 419
Tests of Statistical Significance 420
Visualizing Statistical Significance 421
Chi Square 422

PUTTING IT ALL TOGETHER: STOPS, DRIVERS, SPEEDERS, AND CITATION ZONES 424

Cautions in Interpreting Statistical Significance 426
Visualizing Discernible Differences 428

Main Points 429
Introduction

Empirical research usually uses some type of statistical analysis.

Many people are intimidated by empirical research because they feel uncomfortable with mathematics and statistics. And, indeed, many research reports are filled with otherwise unspecified computations. The role of statistics in criminal justice research is very important, but it is equally important for that role to be seen in its proper perspective.

Empirical research is, first and foremost, a logical rather than a mathematical operation. Mathematics is not much more than a convenient and efficient language for accomplishing the logical operations inherent in good data analysis. Statistics is the applied branch of mathematics especially appropriate to a variety of research analyses.

We'll be looking at two types of statistics: descriptive and inferential. Descriptive statistics are used to summarize and otherwise describe data in manageable forms. Inferential statistics help researchers form conclusions from their observations; typically, that involves forming conclusions about a population from the study of a sample drawn from it.

Before considering any numbers, we want to assure you that the level of statistics used in this chapter has been proven safe for humans. Our introduction to this important phase of criminal justice research is just that—an introduction. Our intent is to familiarize future producers and consumers of empirical criminal justice research with fundamental concepts of quantitative analysis. Many published criminal justice studies use sophisticated statistical techniques that are best learned in specialized courses. But the underlying logic and fundamental techniques of statistics are not at all complicated. It's mostly counting and comparing.

We assume that you are taking a course in research methods for criminology and criminal justice because you are interested in the subjects of crime and criminal justice policy. We suggest that you approach this chapter by thinking about statistics as tools for describing and explaining crime and criminal justice policy. Learning how to use these tools will help you better understand this fascinating subject. And learning how to summarize and interpret data about a subject you find inherently interesting is the least painful and most rewarding way to becoming acquainted with statistics.

Univariate Description

The simplest statistics describe some type of average and dispersion for a single variable.

Descriptive statistics represent a method for presenting quantitative descriptions in a manageable form. Sometimes, we want to describe single variables; this procedure is known as univariate analysis. Other times, we want to describe the associations that connect one variable with another. Bivariate analysis refers to descriptions of two variables, and multivariate analysis examines relationships among three or more variables.

Univariate analysis examines the distribution of cases on only one variable at a time. We'll begin with the logic and formats for the analysis of univariate data.

Distributions

The most basic way to present univariate data is to report all individual cases—that is, to list the attribute for each case under study in terms of the variable in question. Suppose we are interested in the ages of criminal court judges; our data might come from a directory of judges prepared by a state bar association. The most direct manner of reporting the ages of judges is simply to list them: 63, 57, 49, 62, 80, 72, 55, and so forth. Such a report will provide readers with complete details of the data, but it is too cumbersome for most purposes. We could arrange our data in a somewhat more manageable form without losing any of the detail by reporting that 5 judges are 38 years old,
7 are 39, 18 are 40, and so forth. Such a format avoids duplicating data on this variable.

For an even more manageable format—with a certain loss of detail—we might report judges’ ages as marginals, which are frequency distributions of grouped data: 24 judges under 45 years of age, 51 between 45 and 50 years of age, and so forth. Our readers will have less data examine and interpret, but they will not be able to reproduce fully the original ages of all the judges. Thus, for example, readers will have no way of knowing how many judges are 41 years old.

The preceding example presented marginals in the form of raw numbers. An alternative form is the use of percentages. Thus, for example, we might report that x percent of the judges are younger than 45, y percent are between 45 and 50, and so forth. The following table shows an example:

<table>
<thead>
<tr>
<th>Ages of Criminal Court Judges (Hypothetical)</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 35</td>
<td>9%</td>
</tr>
<tr>
<td>36–45</td>
<td>21</td>
</tr>
<tr>
<td>46–55</td>
<td>45</td>
</tr>
<tr>
<td>56–65</td>
<td>19</td>
</tr>
<tr>
<td>65 and older</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>100% = 433</td>
</tr>
<tr>
<td>No data</td>
<td>18</td>
</tr>
</tbody>
</table>

In computing percentages, it is necessary to determine the base from which to compute—the number that represents 100 percent. In the most straightforward situation, the base is the total number of cases under study. A problem arises, however, whenever some cases have missing data. Let’s consider, for example, a survey in which respondents are asked to report their ages. If some respondents fail to answer that question, we have two alternatives. First, we might still base our percentages on the total number of respondents, reporting those who fail to give their ages as a percentage of the total. Second, we might use the number of persons who give an answer as the base from which to compute the percentages; this approach is illustrated in the accompanying table. We will still report the number who do not answer, but they will not figure in the percentages.

The choice of a base depends entirely on the purposes of the analysis. If we wish to compare the age distribution of a survey sample with comparable data on the population from which the sample was drawn, we will probably want to omit the “no answers” from the computation. Our best estimate of the age distribution of all respondents is the distribution for those who answered the question. Because “no answer” is not a meaningful age category, its presence among the base categories will only confuse the comparison of sample and population figures.

### Measures of Central Tendency

Beyond simply reporting marginals, researchers often present data in the form of summary averages, or measures of central tendency. Options in this regard include the mode (the most frequent attribute, either grouped or ungrouped), the arithmetic mean (the sum of values for all observations, divided by the number of observations), and the median (the middle attribute in the ranked distribution of observed attributes). Here’s how the three averages are calculated from a set of data.

Suppose we are conducting an experiment that involves teenagers as subjects. They range in age from 13 to 19, as indicated in this frequency distribution:

<table>
<thead>
<tr>
<th>Age</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
</tr>
</tbody>
</table>
Part Four  Application and Analysis

Now that we know the actual ages of the 31 subjects, how old are these subjects in general, or on average? Let’s look at three different ways we might answer that question.

The easiest average to calculate is the mode, the most frequent value. The distribution of our 31 subjects shows there are more 16-year-olds (8 of them) than any other age, so the modal age is 16, as indicated in Figure 13.1.

Figure 13.1 also demonstrates the calculation of the mean. There are three steps: (1) Multiply each age by the number of subjects who are that age, (2) total the results of all those multiplications, and (3) divide that total by the number of subjects. As indicated in Figure 13.1, the mean age in this illustration is 15.87.

The median represents the “middle” value; half are above it and half below. If we had the precise age of each subject (for instance, 17 years and 124 days), we could arrange all 31 subjects in order by age, and the median for the whole group would be the age of the middle subject.

We do not, however, know precise ages; our data constitute “grouped data” in this regard. Three people who are not precisely the same age have been grouped in the category “13 years old,” for example.

Figure 13.1 illustrates the logic of calculating a median for grouped data. Because there are 31 subjects altogether, the “middle” subject is number 16 when they are arranged by age—15 are younger and 17 are older. The bottom portion of Figure 13.1 shows that the middle person is one of the eight 16-year-olds. In the enlarged view of that group, we see that number 16 is the third from the left.

Measures of Dispersion

In the research literature, we find both means and medians presented. Whenever means are presented, we must be aware that they are susceptible to extreme values: A few very large or very small numbers can change the mean dramatically. Because of this, it is usually important to examine measures of dispersion about the mean.

The simplest measure of dispersion is the range: the distance separating the highest from the lowest value. Thus, besides reporting that our subjects have a mean age of 15.87, we might also indicate that their ages range from 13 to 19. A somewhat more sophisticated measure of dispersion is the standard deviation, which can be described as the average amount of variation about the mean. If the mean is the average value of all observations in a group, then the standard deviation represents the average amount each individual observation varies from the mean. Table 13.1 presents some hypothetical data on the ages of persons in juvenile and adult courts that will help illustrate the concepts of deviation and average deviation. Let’s first consider the top of Table 13.1.

The first column shows the age for each of 10 juvenile court defendants. The mean age for these 10 juveniles is 14. The second column shows how much each individual’s age deviates from the mean. Thus, the first juvenile is 2 years younger than the mean, the second is 1 year older, and the third is the same age as the mean.

You might first think that the average deviation is calculated in the same way as the mean—add up all individual deviations for each case and divide by the number of cases. We did that in Table 13.1, but notice that the total deviation is zero; therefore, the average deviation is zero. In fact, the sum of deviations from the mean will always be zero. This is because some individual deviations will be negative and some will be positive—and the positive and negative values will always cancel each other out.

For this reason (and other reasons too complex to describe here), the standard deviation measure of dispersion is based on the squared deviations from the mean. Squaring any number always produces a positive value, so when we add all the squared deviations together, we will not get zero for the total. Summing these squared deviations in the top of Table 13.1 produces a total of 20, and dividing by the number
Figure 13.1 Three “Averages”
of observations produces an “average” deviation of 2. This quantity—the sum of squared deviations from the mean divided by the number of cases—is called the “variance.” Taking the square root of the variance produces the standard deviation, which is 1.41 for juveniles in Table 13.1.

How should we interpret a standard deviation of 1.41, or any other such value, for that matter? By itself, any particular value for the standard deviation has no intuitive meaning. This measure of dispersion is most useful in a comparative sense. Comparing the relative values for the standard deviation and the

<table>
<thead>
<tr>
<th>Juvenile Court</th>
<th>Age</th>
<th>Deviation from Mean</th>
<th>Squared Deviation from Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>−2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>−1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>−2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>−1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>140</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Average</td>
<td>14</td>
<td>(0)</td>
<td>(2)</td>
</tr>
<tr>
<td>Standard deviation</td>
<td></td>
<td>1.41</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Adult Court</th>
<th>Age</th>
<th>Deviation from Mean</th>
<th>Squared Deviation from Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>−10</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>9</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>−5</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>−6</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>−3</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>15</td>
<td>225</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>−9</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>22</td>
<td>484</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>−7</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>−6</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>280</td>
<td>0</td>
<td>1,126</td>
</tr>
<tr>
<td>Average</td>
<td>28</td>
<td>(0)</td>
<td>(112.6)</td>
</tr>
<tr>
<td>Standard deviation</td>
<td></td>
<td>10.61</td>
<td></td>
</tr>
</tbody>
</table>
mean indicates how much variation there is in a group of cases, relative to the average. Similarly, comparing standard deviations for different groups of cases indicates relative amounts of dispersion within each group.

In our example of juvenile court cases, the standard deviation of 1.41 is rather low relative to the mean of 14. Now compare the data for juvenile court to the bottom half of Table 13.1, which presents ages for a hypothetical group of adult court defendants. The mean is higher, of course, because adults are older than juveniles. More important for illustrating the standard deviation, there is greater variation in the distribution of adult court defendants, as illustrated by the standard deviation and the columns that show raw deviations and squared deviations from the mean of 28. The standard deviation for adult cases (10.61) is much higher relative to the mean of 28 than are the relative values of the standard deviation and mean for juvenile cases. In this hypothetical example, the substantive reason for this is obvious: There is much greater age variation in adult court than in juvenile court because the range for ages of adults is potentially greater (18 to whatever) than the range for ages in juveniles (1 to 17). As a result, the standard deviation for adult defendants indicates greater variation than the same measure for juvenile defendants.

In addition to providing a summary measure of dispersion, the standard deviation plays a role in the calculation of other descriptive statistics, some of which we will touch on later in this chapter. The standard deviation is also a central component of many inferential statistics used to make generalizations from a sample of observations to the population from which the sample was drawn.

Comparing Measures of Dispersion and Central Tendency
Other measures of dispersion can help us interpret measures of central tendency. One useful indicator that expresses both dispersion and grouping of cases is the percentile, which indicates what percentage of cases fall at or below some value. For example, scores on achievement tests such as the SAT are usually reported in both percentiles and raw scores. Thus, a raw score of 630 might fall in the 80th percentile, indicating that 80 percent of persons who take the SAT achieve scores of 630 or less; alternatively, the 80th percentile means that 20 percent of scores were higher than 630. Percentiles may also be grouped into quartiles, which give the cases that fall in the first (lowest), second, third, and fourth (highest) quarters of a distribution.

Table 13.2 presents a distribution of prior arrests for some hypothetical population of, say, probationers to illustrate different measures of central tendency and dispersion. Notice that, although the number of prior arrests ranges from 0 to 55, cases cluster in the lower end of this distribution. Half the cases have 4 or fewer prior arrests, as indicated by three descriptive statistics in Table 13.2: median, 50th percentile, and 2nd quartile. Only one-fourth of the cases have 8 or more prior arrests. Notice also the different values for our three measures of central tendency. The mode for prior arrests is 2, and the mean or average number is 5.76. Whenever the mean is much higher than the mode, it indicates that the mean is distorted by a small number of persons with many prior arrests. The standard deviation of 6.64 further indicates that our small population has quite a bit of variability. Figure 13.2 presents a graphic representation of the dispersion of cases and the different values for the three measures of central tendency.

Distributions such as those shown in Table 13.2 and Figure 13.2 are known as “skewed distributions.” Although most cases cluster near the low end, a few are spread out over very high values for prior arrests. Many variables of interest to criminal justice researchers are skewed in similar ways, especially when examined for some general population. Most people have no prior arrests, but a small number of persons have many. Similarly, most
people suffer no victimization from serious crime in any given year, but a small number of persons are repeatedly victimized.

In an appropriately titled article, “Deviating from the Mean,” Michael Maltz (1994) cautions that criminologists, failing to recognize high levels of variation, sometimes report means for populations that exhibit a great deal of skewness. When reading reports of criminal justice research, researchers are advised to look closely at measures of both dispersion and central tendency. When the numerical value of the standard deviation is high and that for the mean is low, the mean is not a good measure of central tendency. Instead, use the median or mode, which are not affected by extreme values.

The preceding calculations are not appropriate for all variables. To understand this, we must examine two types of variables: continuous and discrete. Age and number of prior arrests are continuous ratio variables; they increase steadily in tiny fractions instead of jumping from category to category as does a discrete variable such as gender or marital status. If discrete variables are being analyzed—a nominal or ordinal variable, for example—then some of the techniques discussed previously are not applicable.

Strictly speaking, medians and means should be calculated for only interval and ratio data, respectively. If the variable in question is gender, for instance, raw numbers or percentage marginals are appropriate and useful measures. Calculating the mode is a legitimate, though not very revealing, tool of analysis, but reports of mean, median, or dispersion summaries would be inappropriate.

**Table 13.2 Hypothetical Data on Distribution of Prior Arrests**

<table>
<thead>
<tr>
<th>Number of Prior Arrests</th>
<th>Number of Cases</th>
<th>Percentage of Cases</th>
<th>Percentile/Quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>8.89</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>17.22</td>
<td>25th/1st</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
<td>12.78</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>11.11</td>
<td>50th/2nd</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>8.89</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>19</td>
<td>10.56</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>18</td>
<td>10.00</td>
<td>75th/3rd</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>6.11</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>14</td>
<td>7.78</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>2.78</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>3</td>
<td>1.67</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>2</td>
<td>1.11</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>1</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>180</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

Mode: 2
Median: 4
Mean: 5.76
Range: 0–55
Standard deviation: 6.64
Computing Rates

Rates are fundamental descriptive statistics in criminal justice research. In most cases, rates are used to standardize some measure for comparative purposes. For example, the following table shows UCR figures on total murders for 2001 in four states.

<table>
<thead>
<tr>
<th>State</th>
<th>Total Murders, 2004</th>
<th>2004 Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>California</td>
<td>2407</td>
<td>35,894,000</td>
</tr>
<tr>
<td>Florida</td>
<td>946</td>
<td>17,397,000</td>
</tr>
<tr>
<td>Louisiana</td>
<td>574</td>
<td>4,516,000</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>650</td>
<td>12,406,000</td>
</tr>
</tbody>
</table>


Obviously, California had far more murders than the other three states had, but these figures are difficult to interpret because of large differences in the states’ total populations. Computing rates enables us to standardize by population size and make more meaningful comparisons, as the next table shows.

<table>
<thead>
<tr>
<th>State</th>
<th>Murder Rates per 100,000 Population, 2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>California</td>
<td>6.7</td>
</tr>
<tr>
<td>Florida</td>
<td>5.4</td>
</tr>
<tr>
<td>Louisiana</td>
<td>12.7</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>5.2</td>
</tr>
</tbody>
</table>

We can see that Louisiana, even with the fewest murders in 2004 (among the states reported here), had the highest murder rate. Notice also that the murder rate is expressed as the number of murders per 100,000 population. This is a common convention in reporting rates of crime and other rare events. To get the rate of
murder per person, move the decimal point five places to the left in each of the figures in the second table. You can clearly see which version is easier to interpret.

The arithmetic of calculating rates could not be much easier. What is not so simple, and in any event requires careful consideration, is deciding on the two basic components of rates: numerator and denominator. The numerator represents the central concept we are interested in measuring, so selecting the numerator involves all the considerations of measurement we have discussed elsewhere. Murder rates, arrest rates, conviction rates, and incarceration rates are common examples in which the numerator is a relatively straightforward count.

Choosing the right denominator is more a question of measurement than one of statistics. In most cases, we should compute rates to standardize according to some population eligible to be included in the numerator. Sometimes, the choice is fairly obvious, as in our use of each state’s total population to compute murder rates. To compute rates of rape or sexual assault, we should probably use the population of adult women in the denominator, although we should also consider how to handle rapes with male victims. Because households are at risk of residential burglary, burglary rates should be computed using some count of households. Similarly, commercial burglaries should be based on some count of commercial establishments, and auto theft on some indicator of registered autos.

More difficult problems can arise in computing rates to express some characteristic of a mobile population. For example, residents of Miami are at risk of criminal victimization in that city, but so are tourists and other visitors to Miami. Because many nonresidents visit or pass through the city in any given year, a measure of Miami’s crime rate based only on the city’s resident population (such as the U.S. Census) will tend to overestimate the number of crimes standardized by the population at risk; many people at risk will not be counted in the denominator. Or what about estimating the crime rate on a subway system? The population at risk here is users, who, in New York City, amount to millions of persons per day.

Fortunately, data to support better measures of rates is becoming increasingly available. For example, a Census Bureau report estimates the daytime population of individual cities by combining the number of residents who work in the city with the number of nonresidents who work there. For Miami, that produces an estimated daytime population in the year 2000 of 497,536, substantially higher than the resident population of 362,470 (U.S. Bureau of the Census, 2005). The estimated weekday ridership for New York’s subways in 2005 was 4.7 million per day, or 1.449 billion annually (New York City Transit, 2006). Crime totals for 2005 for the entire transit system (including buses and ferries) show 3,383 “major felony crimes,” producing a rate per 100,000 of 0.23, or 2.3 major felonies for every million people at risk (City of New York, 2006). If crime data were available for the subway only, the rate would be even lower.

Rates are very useful descriptive statistics that may be easily computed. It is important, however, to be careful in selecting numerators and denominators. Recognize that this caution applies as much to questions of making measurements as it does to questions of computing descriptive statistics. The box titled “Murder on the Job,” presented later in this chapter, gives an example of confusion about the meaning of rates.

**Detail Versus Manageability**

In presenting univariate—and other—data, we are constrained by two often conflicting goals. On the one hand, we should attempt to provide readers with the fullest degree of detail possible regarding those data. On the other hand, the data should be presented in a manageable form. Because these two goals often run counter to each other, researchers must seek the best compromise between them. One useful solution is to report a given set of data
in more than one form. In the case of age, for example, we might report the marginals on ungrouped ages plus the mean age and standard deviation.

Our introductory discussion of univariate analysis shows how this seemingly simple matter can be rather complex. The lessons of this section will be important as we move now to a consideration of analysis involving more than a single variable.

**Describing Two or More Variables**

Descriptive statistics applied to two or more variables are tools to understand relationships among those variables.

Univariate analyses describe the units of analysis of a study and, if they are a sample drawn from some larger population, allow us to make descriptive inferences about the larger population. Bivariate and multivariate analyses are aimed primarily at explanation.

Often, it’s appropriate to describe subsets of cases, subjects, or respondents. Table 13.3, for example, presents hypothetical data on sentence length for offenders grouped by prior felony record. In some situations, the researcher presents subgroup comparisons purely for descriptive purposes. More often, the purpose of subgroup descriptions is comparative. In this case, comparing sentences for subgroups of convicted offenders implies some causal connection between prior felony record and sentence length. Similarly, if we compare sentence lengths for men and women, it implies that something about gender has a causal effect on sentence length.

**Bivariate Analysis**

In contrast to univariate analysis, subgroup comparisons constitute a kind of bivariate analysis that two variables are involved. In such situations, we are usually interested in relationships among the variables. Thus, univariate analysis and subgroup comparisons focus on describing the people (or other units of analysis) under study, whereas bivariate analysis focuses more on the variables themselves.

Notice, then, that Table 13.4 can be regarded as a subgroup comparison: It independently describes gun ownership among male and female respondents in the 2000 General Social Survey. It shows—comparatively and descriptively—that fewer females than males report owning a gun.

The same table viewed as an explanatory bivariate analysis tells a somewhat different story. It suggests that the variable “gender” has an effect on the variable “gun ownership.” The behavior is seen as a dependent variable that is partially determined by the independent variable, gender. Explanatory bivariate analyses, then, involve the “variable language” we introduced in Chapter 1. In a subtle shift of focus, we are no longer talking about male and female as different subgroups but about gender as a variable—a variable that has an influence on other variables.

Adding the logic of causal relationships among variables has an important implication.

---

**Table 13.3** Hypothetical Illustration of Subgroup Comparisons: Length of Prison Sentence by Felony Criminal History

<table>
<thead>
<tr>
<th>Felony Criminal History</th>
<th>Median Sentence Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>No arrests or convictions</td>
<td>6 months</td>
</tr>
<tr>
<td>Prior arrests only</td>
<td>11 months</td>
</tr>
<tr>
<td>Prior convictions</td>
<td>23 months</td>
</tr>
</tbody>
</table>

**Table 13.4** Gun Ownership Among Male and Female Respondents, 2000

<table>
<thead>
<tr>
<th>Own a Gun?</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>42%</td>
<td>25%</td>
</tr>
<tr>
<td>No</td>
<td>58</td>
<td>75</td>
</tr>
<tr>
<td>100%</td>
<td>(817)</td>
<td>(1040)</td>
</tr>
</tbody>
</table>

for the construction and reading of percentage tables. One of the chief bugaboos for novice data analysts is deciding on the appropriate “direction of percentaging” for any given table. In Table 13.4, for example, we divided the group of subjects into two subgroups—male and female—and then described the behavior of each subgroup. That is the correct way to construct this table.

Notice, however, that it would have been possible, though inappropriate, to construct the table differently. We could have first divided the subjects into different categories of gun ownership and then described each of those subgroups by the percentage of male and female subjects in each. This method would make no sense in terms of explanation, however; owning a gun does not make someone male or female.

Table 13.4 suggests that gender affects gun ownership. Had we used the other method of construction, the table would have suggested that gun ownership affects whether someone is male or female—which makes no sense.

Another, related problem complicates the lives of novice data analysts: How do you read a percentage table? There is a temptation to read Table 13.4 as “Among females, only 25 percent owned a gun, and 75 percent did not; therefore, being female makes you less likely to own a gun.” That is not the correct way to read the table, however. The conclusion that gender—as a variable—has an effect on gun ownership must hinge on a comparison between males and females. Specifically, we compare the 25 percent of females with the 42 percent of males and note that women are less likely than men to own a gun. The appropriate comparison of subgroups, then, is essential in reading an explanatory bivariate table.

**Percentaging a Table**  In constructing and presenting Table 13.4, we have used a convention called “percentage down.” This means that we can add the percentages down each column to total 100 percent. We read this form of table across a row. For the row labeled “Yes,” what percentage of the males own a gun? What percentage of the females?

The percentage-down convention is just that—a conventional practice; some researchers prefer to percentage across. They would organize Table 13.4 with “Male” and “Female” on the left side of the table, identifying the two rows, and “Yes” and “No” at the top, identifying the columns. The actual numbers in the table would be moved around accordingly, and each row of percentages would total 100 percent.

In that case, we would make our comparisons between males and females by reading down, within table columns, still asking what percentage of males and females owned guns. The logic and the conclusion would be the same in either case; only the form would be different.

In reading a table that someone else has constructed, therefore, it’s necessary to find out in which direction it has been percentaged. Usually, that will be apparent from the labeling of the table or the logic of the variables being analyzed. Sometimes, however, tables are not clearly labeled. In such cases, the reader should add the percentages in each column and each row. If each of the columns totals 100 percent, the table has been percentaged down. If the rows total 100 percent each, the table has been percentaged across. Follow these rules of thumb:

1. If the table is percentaged down, read across.
2. If the table is percentaged across, read down.

By the way, we constructed Table 13.4 from General Social Survey (GSS) data available on the Internet, as you may have guessed from the Internet address shown at the bottom of the table. GSS data from 1976 onward can be accessed directly through the online data analysis capability of the University of California at Berkeley. You can use a simple interactive program at the Berkeley site to construct your own percentage tables for any GSS variables. This will give you invaluable practice in constructing bivariate percentage tables.
Here's another example: Suppose we are interested in investigating newspaper editorial policies regarding the legalization of marijuana. We undertake a content analysis of editorials on this subject that have appeared during a given year in a sample of daily newspapers across the nation. Each editorial has been classified as favorable, neutral, or unfavorable with regard to the legalization of marijuana. Perhaps we wish to examine the relationship between editorial policies and the types of communities in which the newspapers are published, thinking that rural newspapers might be more conservative than urban ones are. Thus, each newspaper (and so, each editorial) is classified in terms of the population of the community in which it is published.

Table 13.5 presents some hypothetical data describing the editorial policies of rural and urban newspapers. Note that the unit of analysis in this example is the individual editorial. Table 13.5 tells us that there were 127 editorials about marijuana in our sample of newspapers published in communities with populations under 100,000. (Note: This choice of 100,000 is for simplicity of illustration and does not mean that rural refers to a community of less than 100,000 in any absolute sense.) Of these, 11 percent (14 editorials) were favorable toward the legalization of marijuana, 29 percent were neutral, and 60 percent were unfavorable. Of the 438 editorials that appeared in our sample of newspapers published in communities with more than 100,000 residents, 32 percent (140 editorials) were favorable toward legalizing marijuana, 40 percent were neutral, and 28 percent were unfavorable.

When we compare the editorial policies of rural and urban newspapers in our imaginary study, we find—as expected—that rural newspapers are less favorable toward the legalization of marijuana than are urban newspapers. That is determined by noting that a larger percentage (32 percent) of the urban editorials than the rural ones (11 percent) were favorable. We might note, as well, that more rural than urban editorials were unfavorable (60 percent versus 28 percent). Note, too, that this table assumes that the size of a community might affect its newspaper’s editorial policies on this issue, rather than that editorial policy might affect the size of communities.

**Constructing and Reading Tables** Before introducing multivariate analysis, let’s review the steps involved in the construction of explanatory bivariate tables:

1. Divide the cases into groups according to attributes of the independent variable.
2. Describe each of these subgroups in terms of attributes of the dependent variable.
3. Read the table by comparing the independent variable subgroups with one another in terms of a given attribute of the dependent variable.

In the example of editorial policies regarding the legalization of marijuana, the size of a community is the independent variable, and a newspaper’s editorial policy is the dependent variable. The table is constructed as follows:

1. Divide the editorials into subgroups according to the sizes of the communities in which the newspapers are published.
2. Describe each subgroup of editorials in terms of the percentages favorable, neutral, or unfavorable toward the legalization of marijuana.
3. Compare the two subgroups in terms of the percentages favorable toward the legalization of marijuana.

<table>
<thead>
<tr>
<th>Editorial Policy Toward Legalizing Marijuana</th>
<th>Community Size Under 100,000</th>
<th>Community Size Over 100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favorable</td>
<td>11%</td>
<td>32%</td>
</tr>
<tr>
<td>Neutral</td>
<td>29</td>
<td>40</td>
</tr>
<tr>
<td>Unfavorable</td>
<td>60</td>
<td>28</td>
</tr>
<tr>
<td>100% =</td>
<td>(127)</td>
<td>(438)</td>
</tr>
</tbody>
</table>
Bivariate analyses typically have an explanatory purpose. This hypothetical example has hinted at the nature of causation as it is used by social scientists. We hope that the rather simplified approach to causation in these examples will have commonsense acceptability at this point.

**Bivariate Table Formats** Tables such as those we’ve been examining are commonly called **contingency tables**: Values of the dependent variable are contingent on values of the independent variable. Although contingency tables are commonly used in criminal justice research, their format has never been standardized. As a result, a variety of formats will be found in the research literature. As long as a table is easy to read and interpret, there is probably no reason to strive for standardization; however, these guidelines should be followed in the presentation of most tabular data.

1. Provide a heading or a title that succinctly describes what is contained in the table.
2. Present the original content of the variables clearly—in the table itself if at all possible, or in the text with a paraphrase in the table. This information is especially critical when a variable is derived from responses to an attitudinal question because the meaning of the responses will depend largely on the wording of the question.
3. Clearly indicate the attributes of each variable. Complex categories need to be abbreviated, but the meaning should be clear in the table, and, of course, the full description should be reported in the text.
4. When percentages are reported in the table, identify the base on which they are computed. It is redundant to present all the raw numbers for each category, because these could be reconstructed from the percentages and the bases. Moreover, the presentation of both numbers and percentages often makes a table more difficult to read.
5. If any cases are omitted from the table because of missing data (“no answer,” for example), indicate their numbers in the table.

By following these guidelines and thinking carefully about the kinds of causal and descriptive relationships they want to examine, researchers address many policy and research questions in criminal justice. We want to emphasize, however, the importance of thinking through the logic of contingency tables. Descriptive statistics—contingency tables, measures of central tendency, or rates—are sometimes misrepresented or misinterpreted. See the box titled “Murder on the Job” for an example of this.

**Multivariate Analysis**

A great deal of criminal justice research uses multivariate techniques to examine relationships among several variables. Like much statistical analysis, the logic of **multivariate analysis** is straightforward, but the actual use of many multivariate statistical techniques can be complex. A full understanding requires a solid background in statistics and is beyond the scope of this book. In this section, we briefly discuss the construction of multivariate tables—those constructed from three or more variables—and the comparison of multiple subgroups.

Multivariate tables can be constructed by following essentially the same steps outlined previously for bivariate tables. Instead of one independent variable and one dependent variable, however, we will have more than one independent variable. And instead of explaining the dependent variable on the basis of a single independent variable, we’ll seek an explanation through the use of more than one independent variable. Let’s consider an example from research on victimization.

**Multivariate Tables: Lifestyle and Street Crime** If we consult any source of published statistics on victimization (see Chapter 11), we will find several tables that document a relationship between age and personal crime...
victimization—younger people are more often victims of assault and robbery, for example. An influential book by Michael Hindelang and associates (Hindelang, Gottfredson, and Garofalo, 1978) suggested a “lifestyle” explanation for this relationship. The lifestyle of many younger people—visiting bars and clubs for evening entertainment, for example—exposes them to street crime and potential predators more than does the less active lifestyle of older people. This is certainly a sensible hypothesis, and Hindelang and associates found general support for the lifestyle explanation in their analysis of data from early versions of the National Crime Victimization Survey. But the U.S. crime survey data did not include direct measures of lifestyle concepts.

Questionnaire items in the British Crime Survey (BCS) provided better measures of individual behaviors. Using these data, Ronald Clarke and associates (Clarke, Ekbom, Hough, and Mayhew, 1985) examined the link between exposure to risk and victimization, while holding age and gender constant. Specifically, Clarke and colleagues hypothesized that older persons are less often victims of street crime because they spend less time on the streets. The 1982 BCS asked respondents whether they had left their homes in the previous week (that is, the week before they were interviewed) for any evening leisure or social activities. Those who responded yes were asked which nights they had gone out and what they had done. Hypothesizing that some types of evening activities are more risky than others, Clarke and associates restricted their analysis to leisure pursuits away from respondents’ homes, such as visiting a pub, nightclub, or theater. Their dependent variable—street crime victimization—was also carefully defined to include only crimes against persons (actual and attempted assault, robbery, rape, and theft) that occurred away from victims’ homes or workplaces or the homes of friends. Furthermore, because the leisure behavior questions asked about evening activities, only street crime victimizations that took place between 6:00 P.M. and midnight were included.

Clarke and associates therefore proposed a very specific hypothesis that involves three carefully defined concepts and variables: Older persons are less often victims of street crime because they less often engage in behavior that exposes them to risk of street crime. Parts A–C of Table 13.6 present cross-tabulations for the three possible bivariate relationships among these variables: evening street crime victimization by age and by evening leisure pursuits, and evening leisure pursuits by age.

The relationships illustrated in these tables are consistent with the lifestyle hypothesis of personal crime victimization. First, victimization was more common for younger people (ages 16–30) and for those who pursued leisure activities outside their home three or more evenings in the previous week (parts A and B of Table 13.6). Second, as shown in part C, the attributes of young age and frequent exposure to risk were positively related: About 41 percent of the youngest group had gone out three or more nights, compared with 20 percent of those ages 31–60 and only 10 percent of those over age 60.

However, because we are interested in the effects of two independent variables—lifestyle and age—on victimization, we must construct a table that includes all three variables.

Several of the tables we have presented in this chapter are somewhat inefficient. When the dependent variable—street crime victimization—is dichotomous (two attributes), knowing one attribute permits us to easily reconstruct the other. Thus, if we know from Table 13.6A that 1 percent of respondents ages 31–60 were victims of street crime, then we know automatically that 99 percent were not victims. So reporting the percentages for both values of a dichotomy is unnecessary. On the basis of this recognition, Table 13.6D presents the relationship between victimization and two independent variables in a more efficient format. In Table 13.6D, the percentages of respondents who reported a street crime victimization are
shown in the cells at the intersections of the two independent variables. The numbers presented in parentheses below each percentage are the numbers of cases on which the percentages are based. Thus, for example, we know that 563 people ages 16–30 did not go out for evening leisure in the week before their interview and that 3.9 percent of them were victims of street crime in the previous year. We can calculate from this that 22 of those 563 people were victims and the other 541 people were not victims.

Let’s now interpret the results presented in this table:

- Within each age group, persons who pursue outside evening leisure activities three or more times per week are more often victimized. There is not much difference between those who go out once or twice per week and those who stayed home.
- Within each category for evening leisure activities, street crime victimization declines as age increases.
• Exposure to risk through evenings out is less strongly related to street crime victimization than is age.

• Age and exposure to risk have independent effects on street crime victimization. Within a given attribute of one independent variable, different attributes of the second are still related to victimization.

• Similarly, the two independent variables have a cumulative effect on victimization. Younger people who go out three or more times per week are most often victimized.

Returning to the lifestyle hypothesis, what can we conclude from Table 13.6D? First, this measure of exposure to risk is, in fact, related to street crime victimization. People who go out more frequently are more often exposed to risk and more often victims of street crime. As Clarke and associates point out, however, differences in exposure to risk do not account for lower rates of victimization among older persons: Within categories of exposure to risk, victimization still declines with age. So lifestyle is related to victimization, but this measure

The results are interesting. Although a greater percentage of women than men are murdered, a much larger number of men than women are murdered.

Now, recall the story’s headline, “High Murder Rate.” This implies that the number of women murdered on the job, divided by the total number of women at risk of murder on the job, is higher than the same computed rate for men. We need more information than the story provides to verify this claim, but there is a clue. Women are about 45 percent of the workforce, so there are about 1.2 men in the workforce for every woman (55 percent ÷ 45 percent). But about five times as many men as women are murdered on the job (849 ÷ 170).

This should tip you off that the headline is misleading. If the ratio of male-to-female murders is 5 to 1, but the ratio of male-to-female workers is 1.2 to 1, how could the murder rate

for women be higher? You could compute actual rates of murder on the job by finding a suitable denominator; in this case, the number of men and women in the workforce would be appropriate. Consulting the Census Bureau publication Statistical Abstract of the United States would provide this information and enable you to compute rates, as in our final table:

<table>
<thead>
<tr>
<th>Cause of Death at Work</th>
<th>Women (est.)</th>
<th>Men (est.)</th>
<th>Total (est.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Murder</td>
<td>170</td>
<td>849</td>
<td>1004</td>
</tr>
<tr>
<td>Other</td>
<td>256</td>
<td>4808</td>
<td>5079</td>
</tr>
<tr>
<td>Total</td>
<td>426</td>
<td>5657</td>
<td>6083</td>
</tr>
</tbody>
</table>

The New York Times got it wrong; there is a higher murder rate for men on the job. Women are less often killed on the job by any cause, including murder. But women who die on the job (in much smaller numbers than men) are more likely to die from murder than are men who die on the job. A murder rate expresses the number of people murdered divided by the population at risk.

Rates are often computed with inappropriate denominators. But it is less common to find the term rate used so inaccurately.

of behavior—exposure to risk of street crime—does not account for all age-related differences in victimization. Furthermore, what we might call lifestyle “intensity” plays a role here. Going out once or twice a week does not have as much impact on victimization as going out more often. The most intensely active night people ages 30 or younger are most often victims of street crime.

Multivariate contingency tables are powerful tools for examining relationships between a dependent variable and multiple independent variables measured at the nominal or categorical level. Contingency tables can, however, become cumbersome and difficult to interpret if independent variables have several categories or if more than two independent variables are included. In practice, criminal justice researchers often employ more sophisticated techniques for multivariate analysis of discrete or nominal variables. Although the logic of such analysis is not especially difficult, most people learn these techniques in advanced courses in statistics.

Multiple-Subgroup Comparisons: Social Disorganization and Boston Neighborhoods

Just as subgroup comparisons can constitute a type of bivariate analysis, comparing values on some dependent variable across multiple subgroups is a type of multivariate analysis. In a sense, Table 13.6D compared victimization for multiple subgroups defined by age and evening leisure activities.

Multiple-subgroup comparisons are most frequently used to compare values for dependent variables measured at the interval or ratio level. For example, we might wish to compare crime rates among urban areas that have been classified according to different attributes. Table 13.7 is adapted from research by Barbara Warner and Glenn Pierce (1993) that examined the relationship between measures of social disorganization and crime problems.

Drawing on earlier studies by Clifford Shaw and Henry McKay (1969) and by William Julius Wilson (1987), Warner and Pierce compared
calls for police service from 60 Boston neighborhoods to two measures of social disorganization—poverty and population mobility—each computed from census data for Boston neighborhoods. Their poverty measure expresses the percentage of neighborhood residents with incomes below the poverty line in 1980; their population mobility measure expresses the percentage of residents who lived in the same dwelling for less than 5 years. Classifying each neighborhood as high or low on each of these two measures produces the two subgroups shown in Table 13.7. Each cell indicates the mean rate of police calls to report assaults for neighborhoods in each subgroup.

As the table shows, assaults were more common in high-poverty neighborhoods, regardless of residential mobility. No surprises here; this finding is consistent with virtually all research on social disorganization, from Shaw and McKay onward. However, the relationship between mobility and assaults varied, depending on neighborhood poverty levels. In low-poverty areas (the first row of Table 13.7), the mean assault rate was greater for high-mobility neighborhoods, but the opposite was true for high-poverty neighborhoods (the second row). Warner and Pierce refer to Wilson’s influential book, *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy* (1987), to interpret this finding: Assaults were most common in poor, stable neighborhoods—those where poverty was persistent and mobility was limited. As Warner and Pierce (1993:507) state, these are “neighborhoods where people remain because they have no choice. This type of stability appears to have an impact on crime in a very different way. Rather than building cohesiveness, it may build resentment, frustration, and isolation.”

Notice also how it is necessary to compare the *combined* categories of poverty and mobility to reveal this finding. If we examined the mean number of assaults for each variable separately, shown in the row and column totals for Table 13.7, we would conclude that assaults are more common in high-poverty and high-mobility neighborhoods. Comparing multiple subgroups defined by the intersection of poverty and mobility is a form of multivariate analysis that enabled Warner and Pierce to sort out the subtle relationships between these two independent variables and neighborhood assault rates. Later in this chapter, we will return to this study to illustrate the use of a more powerful multivariate technique. But multiple subgroup comparisons can provide useful information about the relationship between more than one independent variable and dependent variables measured at the interval or ratio level.

**Measures of Association**

As we’ve suggested, descriptive statistics is a method for presenting quantitative descriptions in a manageable form. Sometimes, we want to describe single variables; other times, we want to describe the associations that connect one variable with another.

Bivariate contingency tables are one way to examine the association between two variables. But contingency tables can be quite complex, presenting several different response categories for row and column variables. Nevertheless, a contingency table represents the association between two variables as a data matrix. Table 13.8 presents such a matrix, showing hypothetical data for the joint frequency distribution of

<table>
<thead>
<tr>
<th>Poverty</th>
<th>Mobility</th>
<th>Assault Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Low</td>
<td>12.2 (22)</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>19.5 (21)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>15.8 (43)</td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td>43.8 (4)</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>25.0 (13)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>29.4 (17)</td>
</tr>
<tr>
<td>Total</td>
<td>Low</td>
<td>17.1 (26)</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>21.6 (34)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>19.6 (60)</td>
</tr>
</tbody>
</table>
education and support for gun control laws. It provides all the information needed to determine the nature and extent of the relationship between education and support for gun control. Notice, for example, that 23 people (1) had no education and (2) scored low on support for gun control; 77 people (1) had graduate degrees and (2) scored high on support for gun control.

However, this matrix presents more information than we can easily comprehend. Studying the table carefully shows that as education increases from “None” to “Graduate Degree” there is a general tendency for gun control support to increase. A variety of descriptive statistics can summarize this data matrix. Selecting the appropriate measure depends initially on the nature of the two variables.

We’ll turn now to some of the options available for summarizing the association between two variables. Each measure of association we’ll discuss is based on the same model—proportionate reduction of error (PRE).

To see how this model works, suppose we ask you to guess respondents’ attributes on a given variable—for example, whether they answered yes or no to a given questionnaire item. To assist you, let’s first assume you know the overall distribution of responses in the total sample—say, 60 percent yes and 40 percent no. You will make the fewest errors in this process if you always guess the modal (most frequent) response: yes.

Second, let’s assume you also know the empirical relationship between the first variable and some other variable—say, gender. Now, each time we ask you to guess whether a respondent said yes or no, we’ll tell you whether the respondent is a man or a woman. If the two variables are related, you should make fewer errors the second time. It is possible, therefore, to compute the PRE by knowing the relationship between the two variables: the stronger the relationship, the greater the reduction of error.

This basic PRE model is modified slightly to take account of different levels of measurement—nominal, ordinal, or interval. The following sections will consider each level of measurement and present one measure of association appropriate to each. However, the three measures discussed are only a partial representation of the many appropriate measures.

### Nominal Variables

If the two variables consist of nominal data (for example, gender, marital status, or race), lambda (λ) is one appropriate measure. Lambda is based on your ability to guess values on one of the variables: the PRE achieved through knowledge of values on the other variable. A simple example will illustrate the logic and method of lambda.

Table 13.9 presents hypothetical data relating gender to a common measure of fear of crime, based on the question “How safe do you feel, or would you feel, walking alone on your neighborhood streets at night?” Overall, we note that 1100 people answered “safe” on the fear question, and 900 answered “unsafe.” If you try to predict how people will respond to the fear-of-crime question, knowing only the overall distribution on that variable, you will always predict “safe” because that will result in...
fewer errors than always predicting “unsafe.” Nevertheless, this strategy will result in 900 errors out of 2000 predictions.

But suppose you have access to the data in Table 13.9 and are told each person’s gender before making your predictions about fear of crime. Your strategy will change: For every man, you will predict “safe,” and for every woman, you will predict “unsafe.” In this instance, you will make 300 errors—the 100 men who responded “unsafe” and the 200 women who responded “safe”—600 fewer errors than you would make without knowing the person’s gender.

Lambda, then, represents the reduction in errors as a proportion of the errors that would have been made on the basis of the overall distribution. In this hypothetical example, lambda equals .67—that is, 600 fewer errors divided by the 900 total errors based on fear of crime alone. In this fashion, lambda measures the statistical association between gender and fear of crime.

If gender and fear were statistically independent, we would find the same distribution of fear of crime for men and women. In this case, knowing gender will not affect the number of errors made in predicting fear, and the resulting lambda will be zero. If, however, all men responded “safe” and all women responded “unsafe,” then by knowing gender you can avoid all errors in predicting fear of crime. You will make 900 fewer errors (out of 900), so lambda will equal 1.0—a perfect statistical association.

Lambda is only one of several measures of association appropriate to the analysis of two nominal variables. Any statistics textbook will describe other appropriate measures.

**Table 13.9 Hypothetical Data Relating Gender to Fear of Crime**

<table>
<thead>
<tr>
<th>Fear</th>
<th>Men</th>
<th>Women</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Safe</td>
<td>900</td>
<td>200</td>
<td>1100</td>
</tr>
<tr>
<td>Unsafe</td>
<td>100</td>
<td>800</td>
<td>900</td>
</tr>
<tr>
<td>Total</td>
<td>1000</td>
<td>1000</td>
<td>2000</td>
</tr>
</tbody>
</table>

**Ordinal Variables** If the variables being related are ordinal (for example, occupational status or education), gamma (γ) is one appropriate measure of association. Like lambda, gamma is based on the ability to guess values on one variable by knowing values on another. Instead of exact values, however, gamma is based on the ordinal arrangement of values. For any given pair of cases, you guess that their ordinal ranking on one variable will correspond (positively or negatively) to their ordinal ranking on the other. For example, if you suspect that political conservatism is negatively related to support for gun control, and if person A is more conservative than person B, then you guess that A is less supportive of gun control than B. Gamma is the proportion of paired comparisons that fits this pattern.

Table 13.10 presents hypothetical data relating political ideology to support for gun control. The general nature of the relationship between these two variables is that, as conservatism increases, support for gun control decreases. There is a negative association between conservatism and support for gun control.

Gamma is computed from two quantities:
1. the number of pairs that have the same ranking on the two variables and
2. the number of pairs that have the opposite ranking on the two variables. The pairs that have the same ranking are computed as follows: The frequency of each cell in the table is multiplied by the sum of all cells appearing below and to the right of it—with all these products being summed. In Table 13.10, the number of pairs with the same ranking is 200(900 + 300 + 400 + 100) + 500(300 + 100) + 400(400 + 100) + 900(100), or 340,000 + 200,000 + 200,000 + 90,000 = 830,000.

The pairs that have the opposite rankings on the two variables are computed as follows: The frequency of each cell in the table is multiplied by the sum of all cells appearing below and to the left of it—with all these products being summed. In Table 13.10, the number of pairs with opposite rankings is 700(500 + 800
Gamma is computed from the numbers of same-ranked pairs and opposite-ranked pairs as follows:

$$\text{Gamma} = \frac{\text{same} - \text{opposite}}{\text{same} + \text{opposite}}$$

In our example, gamma equals \((830,000 - 3,430,000)\) divided by \((830,000 + 3,430,000)\), or \(-0.61\). The negative sign in this answer indicates the negative association suggested by the initial inspection of the table. Conservatism and support for gun control, in this hypothetical example, are negatively associated. The numerical figure for gamma indicates that 61 percent more of the pairs examined had opposite rankings than the same ranking.

Note that whereas values of lambda vary from 0 to 1, values of gamma vary from \(-1\) to \(+1\) and represent both the direction and the magnitude of the association. Because nominal variables have no ordinal structure, it makes no sense to speak of the direction of the relationship. A negative lambda indicates that we made more errors in predicting values on one variable, while knowing values on the second, than we made in ignorance of the second, and that’s not logically possible.

### Interval or Ratio Variables

If interval or ratio variables (for example, age, income, or number of arrests) are being analyzed, one appropriate measure of association is Pearson’s product-moment correlation \((r)\). The derivation and computation of this measure of association are beyond the scope of this book, so we will make only a few general comments here.

Like both gamma and lambda, \(r\) is based on guessing the value of one variable by knowing the other. For a continuous interval or ratio variable, however, it is unlikely that we can predict the precise value of the variable. At the same time, predicting only the ordinal arrangement of values on the two variables does not take advantage of the greater amount of information conveyed by an interval or ratio variable. In a sense, \(r\) reflects how closely we can guess the value of one variable through our knowledge of the value of the other.

To understand the logic of \(r\), let’s consider how we might hypothetically guess values that particular cases have on a given variable. With nominal variables, we have seen that the best guess is the modal value. For interval or ratio data, we can minimize our errors by always guessing the mean value of the variable. Although this practice produces few, if any, perfect guesses, it will minimize the extent of our errors.

In the computation of lambda, we noted the number of errors produced by always guessing the modal value. In the case of \(r\), errors are measured in terms of the sum of the squared differences between the actual value and the mean. This sum, called the “total variation,” is calculated in much the same way as the standard deviation, discussed earlier in this chapter.

To understand that concept, we must expand the scope of our examination. Let’s look at the logic of regression analysis, and we’ll return to correlation within that context.

**Regression Analysis**

The general formula for describing the association between two variables is \(Y = f(X)\). This formula is read, “\(Y\) is a function of \(X\),” which means that values of \(Y\) can be explained in terms of variations in the values of \(X\). Stated more strongly, we might say that \(X\) causes \(Y\), so the value of \(X\) determines the value of \(Y\). **Regression analysis** is a method of determining the specific function relating \(Y\) to \(X\). There are several forms of regression analysis, depending on the complexity of the rela-
Let’s begin with the simplest. The regression model can be seen most clearly in the case of a perfect linear association between two variables. Figure 13.3 is a scattergram presenting in graphic form the values of \( X \) and \( Y \) produced by a hypothetical study. It shows that for the four cases in our study the values of \( X \) and \( Y \) are identical in each instance. The case with a value of 1 on \( X \) also has a value of 1 on \( Y \), and so forth. The relationship between the two variables in this instance is described by the equation \( Y = a + bX \); this is the regression equation. Because all four points lie on a straight line, we can superimpose that line over the points; this is the regression line.

The linear regression model has important descriptive uses. The regression line offers a graphic picture of the association between \( X \) and \( Y \), and the regression equation is an efficient form for summarizing that association. The regression model has inferential value as well. To the extent that the regression equation correctly describes the general association between the two variables, it may be used to predict other sets of values. If, for example, we know that a new case has a value of 3.5 on \( X \), then we can predict the value of 3.5 on \( Y \) as well.

In practice, of course, studies are seldom limited to four cases, and the associations between variables are seldom as clear as the one presented in Figure 13.3. Figure 13.4 is a somewhat more realistic example that represents a hypothetical relationship between population and crime rate in small to medium-size cities. Each dot in the scattergram is a city, and each dot’s placement reflects that city’s population and its crime rate. As in our previous example, the values of \( Y \) (crime rates) generally correspond to those of \( X \) (populations), and as values of \( X \) increase, so do values of \( Y \). However, the association is not nearly as clear as it was in Figure 13.3.

It is not possible in Figure 13.4 to superimpose a straight line that will pass through all the points in the scattergram. But we can draw an approximate line showing the best possible linear representation of the several points.

If you’ve ever studied geometry, you know that any straight line on a graph can be represented by an equation of the form \( Y = a + bX \), where \( X \) and \( Y \) are values of the two variables. In this equation, \( a \) equals the value of \( Y \) when \( X \) is zero, and \( b \) represents the slope of the line. If we know the values of \( a \) and \( b \), we can estimate \( Y \) for every value of \( X \).

Regression analysis is a technique for establishing the regression equation representing the geometric line that comes closest to the distribution of points. This equation is valuable both descriptively and inferentially. First, the regression equation provides a mathematical description of the relationship between the variables. Second, it allows us to infer values of \( Y \) when we have values of \( X \). Referring to Figure 13.4, we can estimate crime rates of cities if we know their populations.

To improve our guessing, we construct a regression line, stated in the form of a regression equation that permits the estimation of values on one variable from values on the other. The general format for this equation is \( Y' = a + b(X) \), where \( a \) and \( b \) are computed values, \( X \) is a given value on one variable, and \( Y' \) is the estimated value on the other. The values of \( a \) and \( b \) are computed to minimize the differences between the actual values of \( Y \) and the corresponding estimates (\( Y' \)) based on the
known value of $X$. The sum of squared differences between actual and estimated values of $Y$ is called the “unexplained variation” because it represents errors that exist even when estimates are based on known values of $X$.

The “explained variation” is the difference between the total variation and the unexplained variation. Dividing the explained variation by the total variation produces a measure of the proportionate reduction of error corresponding to the similar quantity in the computation of lambda. In the present case, this quantity is the correlation squared: $r^2$. Thus, if $r = .7$, then $r^2 = .49$, which means that about half the variation has been explained.

In practice, we compute $r$ rather than $r^2$ because the product-moment correlation can take either a positive or a negative sign, depending on the direction of the relationship between the two variables. Computing $r^2$ and taking a square root will always produce a positive quantity, which means we will lose information about the direction of the relationship between $X$ and $Y$. Consult any standard statistics textbook for the method of computing $r$. In practice, this and other measures of association are easily calculated with a variety of computer programs, including popular spreadsheet and database applications.

**Inferential Statistics**

*When we generalize from samples to larger populations, we use inferential statistics to test the significance of an observed relationship.*

Many criminal justice research projects examine data collected from a sample drawn from a larger population. A sample of people may be interviewed in a survey; a sample of court records may be coded and analyzed; a sample of newspapers may be examined through
content analysis. Researchers seldom, if ever, study samples merely to describe the samples per se; in most instances, their ultimate purpose is to make assertions about the larger population from which the sample has been selected. Frequently, then, we will want to interpret our univariate and multivariate sample findings as the basis for inferences about some population.

This section examines the statistical measures used for making such inferences and their logical bases. We’ll begin with univariate data and then move to bivariate.

Univariate Inferences
The opening sections of this chapter dealt with methods of presenting univariate data. Each summary measure was intended to describe the sample studied. Now we will use those measures to make broader assertions about the population. This section will focus on two univariate measures: percentages and means.

If 50 percent of a sample of people say they received traffic tickets during the past year, then 50 percent is also our best estimate of the proportion of people who received traffic tickets in the total population from which the sample was drawn. Our estimate assumes a simple random sample, of course. It is rather unlikely, however, that precisely 50 percent of the population got tickets during the year. If a rigorous sampling design for random selection has been followed, we will be able to estimate the expected range of error when the sample finding is applied to the population.

The section in Chapter 8 on sampling theory covered the procedures for making such estimates, so they will only be reviewed here. The quantity

\[ s = \sqrt{\frac{p \times q}{n}} \]

where \( p \) is a percentage, \( q \) equals 1 - \( p \), and \( n \) is the sample size, is called the “standard error.” As noted in Chapter 8, this quantity is very important in the estimation of sampling error. We may be 68 percent confident that the population figure falls within plus or minus 1 standard error of the sample figure, 95 percent confident that it falls within plus or minus 2 standard errors, and 99.9 percent confident that it falls within plus or minus 3 standard errors.

Any statement of sampling error, then, must contain two essential components: the confidence level (for example, 95 percent) and the confidence interval (for example, 2.5 percent). If 50 percent of a sample of 1600 people say they have received traffic tickets during the year, we might say we are 95 percent confident that the population figure is between 47.5 and 52.5 percent.

Recognize in this example that we have moved beyond simply describing the sample into the realm of making estimates (inferences) about the larger population. In doing that, we must be wary of three assumptions.

First, the sample must be drawn from the population about which inferences are being made. A sample taken from a telephone directory, for example, cannot legitimately be the basis for statistical inferences about the population of a city.

Second, the inferential statistics assume simple random sampling, which is virtually never the case in actual sample surveys. The statistics assume sampling with replacement, which is almost never done, but that is probably not a serious problem. Although systematic sampling is used more frequently than random sampling is, that, too, probably presents no serious problem if done correctly. Stratified sampling, because it improves representativeness, clearly presents no problem. Cluster sampling does present a problem, however, because the estimates of sampling error may be too small. Clearly, street-corner sampling does not warrant the use of inferential statistics. This standard error sampling technique also assumes a 100 percent completion rate. This problem increases in seriousness as the completion rate decreases.
Third, inferential statistics apply to sampling error only; they do not take account of nonsampling errors. Thus, although we might correctly state that between 47.5 and 52.5 percent of the population (95 percent confidence) will report getting a traffic ticket during the previous year, we cannot so confidently guess the percentage that had actually received them. Because nonsampling errors are probably larger than sampling errors in a respectable sample design, we need to be especially cautious in generalizing from our sample findings to the population.

Tests of Statistical Significance

There is no scientific answer to the question of whether a given association between two variables is significant, strong, important, interesting, or worth reporting. Perhaps the ultimate test of significance rests with our ability to persuade readers (present and future) of the association’s significance. At the same time, a body of inferential statistics—known as parametric tests of significance—can assist in this regard. As the name suggests, parametric statistics make certain assumptions about the parameters that describe the population from which the sample is selected.

Although tests of statistical significance are widely reported in criminal justice literature, the logic underlying them is subtle and often misunderstood. Tests of significance are based on the same sampling logic that has been discussed elsewhere in this book. To understand that logic, let’s return to the concept of sampling error with regard to univariate data.

Recall that a sample statistic normally provides the best single estimate of the corresponding population parameter, but the statistic and the parameter are seldom identical. Thus, we report the probability that the parameter falls within a certain range (confidence interval). The degree of uncertainty within that range is due to normal sampling error. The corollary of such a statement is, of course, that it is improbable that the parameter will fall outside the specified range only as a result of sampling error. Thus, if we estimate that a parameter (99.9 percent confidence) lies between 45 and 55 percent, we say by implication that it is extremely improbable that the parameter is actually, say, 70 percent if our only error of estimation is due to normal sampling.

The fundamental logic of tests of statistical significance, then, is this: Faced with any discrepancy between the assumed independence of variables in a population and the observed distribution of sample elements, we may explain that discrepancy in either of two ways: (1) We attribute it to an unrepresentative sample, or (2) we reject the assumption of independence. The logic and statistics associated with probability sampling methods offer guidance about the varying probabilities of different degrees of unrepresentativeness (expressed as sampling error). Most simply put, there is a high probability of a small degree of unrepresentativeness and a low probability of a large degree of unrepresentativeness.

The statistical significance of a relationship observed in a set of sample data, then, is always expressed in terms of probabilities. Significant at the .05 level (p \leq .05) simply means that the probability of a relationship as strong as the observed one being attributable to sampling error alone is no more than 5 in 100. Put somewhat differently, if two variables are independent of each other in the population, and if 100 probability samples were selected from that population, then no more than 5 of those samples should provide a relationship as strong as the one that has been observed.

There is, then, a corollary to confidence intervals in tests of significance, which represent the probability of the measured associations being due only to sampling error. This is called the level of significance. Like confidence intervals, levels of significance are derived from a logical model in which several samples are drawn from a given population. In the present case, we assume that no association exists between the variables in the population, and then we ask what proportion of the samples drawn from that population would produce associa-
tions at least as great as those measured in the empirical data. Three levels of significance are frequently used in research reports: .05, .01, and .001. These mean, respectively, that the chances of obtaining the measured association as a result of sampling error are no more than 5 in 100, 1 in 100, and 1 in 1000.

Researchers who use tests of significance normally follow one of two patterns. Some specify in advance the level of significance they will regard as sufficient. If any measured association is statistically significant at that level, they will regard it as representing a genuine association between the two variables. In other words, they are willing to discount the possibility of its resulting from sampling error only.

Other researchers prefer to report the specific level of significance for each association, disregarding the conventions of .05, .01, and .001. Rather than reporting that a given association is significant at, say, the .05 level, they might report significance at the .023 level, indicating that the chances of its having resulted from sampling error as no more than 23 in 1000.

**Visualizing Statistical Significance**

In a Bureau of Justice Statistics publication describing the National Crime Victimization Survey (NCVS) for a nontechnical audience, Michael Maltz and Marianne Zawitz (1998) present a very informative graphical display to show statistical significance.

Recall that the NCVS is a national sample designed to estimate nationwide rates of victimization. Maltz and Zawitz use visual displays of estimates and their confidence intervals to demonstrate the relative precision of victimization rates disclosed by the survey.

Figure 13.5, reproduced from Maltz and Zawitz, presents an example of this approach. The figure shows annual rates of change in all violent victimizations from 1973 through 1996. Notice the vertical line in Figure 13.5, representing no change in violent victimization rates for each year. Estimates of annual rates of change are shown in horizontal bars arrayed along the vertical line. The horizontal bars for each year present parameter estimates for annual change, signified by a dot or square, bracketed by the confidence intervals for each parameter estimate at three confidence levels: 68 percent (1 standard error), 90 percent (1.6 standard errors), and 95 percent (2 standard errors).

Consider first the topmost bar, showing estimates for change from 1973 to 1974. The small dot signifying the point estimate (an increase of 1.24 percent) is just to the right of the no-change vertical line. But notice also that the confidence intervals of 1, 1.6, and 2 standard errors cross over the no-change line; the estimate of a 1.24 percent increase is less than 2 standard errors above zero. This means that, using the .05 (2 standard errors) criterion, the estimated increase is not statistically significant. To further emphasize this point, notice the key in Figure 13.5, which shows different probabilities that a change occurred each year; the small dot representing the point estimate for 1973–74 change indicates the probability that a change actually occurred is less than 90 percent.

Now consider the most recent estimate shown in this figure—change from 1995 to 1996. The point estimate of −9.9 percent is well below the no-change line, and the confidence intervals are well to the left of this line. Bracketing the point estimate by 2 standard errors produces an interval estimate of between −15.7 percent (point estimate minus 2 standard errors) and −4.05 percent (point estimate plus 2 standard errors). This means we can be 95 percent certain that the violent victimization rate declined from 1995 to 1996 by somewhere between −4.05 and −15.7 percent. Because the point estimate of −9.9 percent is more than 2 standard errors from zero, we can confidently say there was a statistically significant decline in violent victimization.

Displaying point and interval estimates in this way accurately represents the concepts of statistical inference and statistical significance. Sample-based estimates of victimization, or
any other variable, are just that—estimates of population values, bounded by estimates of standard error. Statistical significance, in this example, means that our estimates of change are above or below zero, according to some specified criterion for significance.

Studying Figure 13.5 will enhance your understanding of statistical inference. We suggest that you consider the point and interval estimates for each year’s change. Pay particular attention to the confidence intervals and their position relative to the no-change line. Then classify the statistical significance (at the .05 level) for change each year into one of three categories: (1) no change, (2) significant increase, and (3) significant decrease. You’ll find our tabulation in the exercises at the end of this chapter.

**Chi Square**

Chi square ($\chi^2$) is a different type of significance test that is widely used in criminal justice research. It is based on the null hypothesis: the assumption that there is no relationship between two variables in a population. Given the observed distribution of values on two variables in a contingency table, we compute the joint distribution that would be expected if there were no relationship between the two variables. The result of this operation is a set of expected frequencies for all the cells in the contingency table. We then compare this ex-
Chapter 13  Interpreting Data

Expected distribution with the empirical distribution—cases actually found in the data—and determine the probability that the difference between expected and empirical distributions could have resulted from sampling error alone. Stated simply, chi square compares what you get (empirical) with what you expect given a null hypothesis of no relationship. Example will illustrate this procedure.

Let’s assume we are interested in the possible relationship between gender and whether people avoid areas near their home because of crime, which we will refer to as “avoidance behavior.” To test this relationship, we select a sample of 100 people at random. Our sample is made up of 40 men and 60 women; 70 percent of our sample report avoidance behavior, and the remaining 30 percent do not.

If there is no relationship between gender and avoidance behavior, then 70 percent of the men in the sample should report avoiding areas near their home, and 30 percent should report no avoidance behavior. Moreover, women should describe avoidance behavior in the same proportion. Table 13.11 (part I) shows that, based on this model, 28 men and 42 women say they avoid areas at night, with 12 men and 18 women reporting no avoidance.

Part II of Table 13.11 presents the observed avoidance behavior for the hypothetical sample of 100 people. Note that 20 of the men say they avoid areas at night, and the remaining 20 say they do not. Among the women in the sample, 50 avoid areas and 10 do not. Comparing the expected and observed frequencies (parts I and II), we note that somewhat fewer men report avoidance behavior than expected, whereas somewhat more women than expected avoid areas near their home at night.

Chi square is computed as follows: For each cell in the tables, we (1) subtract the expected frequency for that cell from the observed frequency, (2) square this quantity, and (3) divide the squared difference by the expected frequency. This procedure is carried out for each cell in the tables, and the results are added. Part III of Table 13.11 presents the cell-by-cell computations. The final sum is the value of chi square—12.70 in this example.

This value is the overall discrepancy between the observed distribution in the sample and the distribution we would expect if the
PUTTING IT ALL TOGETHER

As you probably realize by now, studies of racial profiling and traffic enforcement generated an enormous amount of data. That means some sort of summary analysis is crucial. A good general principle to keep in mind for data analysis is that as the number of data points or observations we make increases, some type of statistical analysis become essential. It need not be a sophisticated analysis, though the research reports we have cited include some advanced techniques.

Our final installment of the running example examines two simple tables that illustrate some basic elements of data analysis. We'll offer explanations of what the tables represent and our interpretation of results. It would be extremely value for you to consult the major studies we have mentioned to review other forms of data analysis that are presented in each.

One Population and Two Samples

The table below shows more detailed results from the New Jersey research conducted by Lange and associates:

New Jersey Turnpike, % Each Race Category (2 Standard Errors)

<table>
<thead>
<tr>
<th></th>
<th>Stops (Police Data)</th>
<th>Drivers (Tollbooth Survey)</th>
<th>Speeders (Observational Survey)</th>
</tr>
</thead>
<tbody>
<tr>
<td>South</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>52%</td>
<td>65.9% (3.8)</td>
<td>58.3% (3.0)</td>
</tr>
<tr>
<td>Black</td>
<td>29%</td>
<td>15.1% (2.6)</td>
<td>26.0% (2.6)</td>
</tr>
<tr>
<td>Central</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>57%</td>
<td>63.9% (5.0)</td>
<td>60.2% (2.4)</td>
</tr>
<tr>
<td>Black</td>
<td>23%</td>
<td>12.5% (2.4)</td>
<td>25.6% (2.1)</td>
</tr>
<tr>
<td>North</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>60%</td>
<td>58.7% (4.2)</td>
<td>68.4% (1.2)</td>
</tr>
<tr>
<td>Black</td>
<td>16%</td>
<td>13.1% (2.6)</td>
<td>18.2% (1.0)</td>
</tr>
</tbody>
</table>

Source: Adapted from Lange, Johnson, and Voas (2005: 209).

Each of three segments of the New Jersey Turnpike is presented separately because, as we will see, where you drive makes a difference. If you want some better idea about what areas are covered by these segments, you can see maps at the New Jersey Turnpike Authority website [http://www.state.nj.us/turnpike/nj-vcenter-maps.htm]. Each row represents the percentage of white or black drivers in each of the three segments. Three columns show moving vehicle stops by state police, results from the tollbooth survey of drivers, and results from the observational survey of speeders. So, for example, the first column of the first row indicates that 52 percent of drivers stopped for a moving violation in the southern segment were white; the second row shows black drivers were 29 percent of those stopped in the southern part.

The columns for drivers and speeders also show what percent of drivers are white and black. But the second entry for each category (in parentheses) shows 2 standard errors for each percentage. Recall these were samples. Samples have standard errors that reflect dispersion around each percentage estimate. We can be 95 percent confident that the percentage for a population lies within 2 standard errors above and below the point estimate. So, for the last column in the first row, we can be 95 percent confident that whites were between 55.3 percent and 61.3 percent of the speeders recorded on the southern segment of the turnpike. Also recall that a statistically significant difference exists between two statistics if their interval estimates do not overlap. Compare the interval estimates for the percentage of drivers who are black and the percentage of speeders who are black:

<table>
<thead>
<tr>
<th></th>
<th>Drivers</th>
<th>Speeders</th>
</tr>
</thead>
<tbody>
<tr>
<td>South</td>
<td>12.5–17.7%</td>
<td>23.4–28.6%</td>
</tr>
<tr>
<td>Central</td>
<td>10.1–14.9%</td>
<td>23.5–27.7%</td>
</tr>
<tr>
<td>North</td>
<td>10.5–15.7%</td>
<td>17.2–19.2%</td>
</tr>
</tbody>
</table>

Make the comparisons within segments, or within columns. Which intervals overlap? Within
each segment, the percentage of speeders who are black is significantly higher than the percentage of drivers who are black. Across the three segments, there are no significant differences in the percentage of drivers who are black. But significantly fewer speeders are black in the northern segment compared to the central or southern segments. If you graph these interval estimates, as Maltz and Zawitz did in Figure 13.5, you should see these patterns clearly.

**Estimating Risk of Traffic Stops**

The next table simply combines indicators from different sources to compare the relative likelihood of traffic stops in different segments of the turnpike. Row A shows the large difference in traffic volume between segments. The number of traffic stops also varies, but in the opposite direction. After first estimating the number of stops per day (Row C), Maxfield and Kelling standardized estimated daily stops by estimated traffic volume. In this way, row D shows the daily stop rate per 100,000 vehicles, something like a crime rate. The most useful items are in row E, comparing the risk of being stopped in the central and southern segments to the risk of being stopped in the north. You can see that drivers in the south are over six times more likely to be stopped compared to those in the north.

<table>
<thead>
<tr>
<th></th>
<th>North</th>
<th>Central</th>
<th>South</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Average daily volume</td>
<td>262,130</td>
<td>126,020</td>
<td>73,830</td>
</tr>
<tr>
<td>B. Moving stops</td>
<td>10,591</td>
<td>11,292</td>
<td>18,794</td>
</tr>
<tr>
<td>C. Est. daily stops</td>
<td>29.0</td>
<td>30.9</td>
<td>51.5</td>
</tr>
<tr>
<td>D. Moving stops/100,000 vehicles</td>
<td>11.1</td>
<td>24.5</td>
<td>69.7</td>
</tr>
<tr>
<td>E. Relative risk of stop</td>
<td>1.0</td>
<td>2.2</td>
<td>6.3</td>
</tr>
</tbody>
</table>

A. N.J. Department of Transportation counts for June 2002 at midpoint of segment
B. New Jersey State Police stop data, 1 May 2002–30 April 2003
C. Row B ÷ 365
D. Row C ÷ row A × 100,000
E. Row D / row D for North

*Source: Adapted from Maxfield and Kelling (2005:21).*

Now compare the data we have presented here to the summary findings by Maxfield and Kelling presented in Chapter 12’s installment. You should see how these data support those conclusions.

Much more sophisticated analysis is presented in reports for Pennsylvania and North Carolina. The Pennsylvania report includes maps for a large number of counties in the state. One finding common to these reports and research in New Jersey is that patterns of road use and traffic enforcement vary enormously. For that reason, among others, it is very misleading, and bordering on irresponsible, to make summary statements about road use and traffic enforcement for a large geographical area.

We urge you to consult one or more of these reports as one way of continuing to develop your research skills.
two variables were unrelated. Of course, the mere discovery of a discrepancy does not prove that the two variables are related, because normal sampling error might produce discrepancies even when there is no relationship in the total population. The magnitude of the value of chi square, however, permits us to estimate the probability of that having happened.

To determine the statistical significance of the observed relationship, we must use a standard set of chi-square values. That will require the computation of the degrees of freedom. For chi square, the degrees of freedom are computed as follows: The number of rows in the table of observed frequencies, minus one, is multiplied by the number of columns, minus one. This may be written as \((r - 1)(c - 1)\). In the present example, we have two rows and two columns (discounting the totals), so there is 1 degree of freedom.

Turning to a table of chi-square values (see the inside back cover), we find that, for 1 degree of freedom and random sampling from a population in which there is no relationship between two variables, 10 percent of the time we should expect a chi square of at least 2.7. Thus, if we select 100 samples from such a population, we should expect about 10 of those samples to produce chi squares equal to or greater than 2.7. Moreover, we should expect chi-square values of at least 6.6 in only 1 percent of the samples and chi-square values of 10.8 in only 0.1 percent of the samples. The higher the chi-square value, the less probable it is that the value can be attributed to sampling error alone.

In our example, the computed value of chi square is 12.70. If there is no relationship between gender and avoidance behavior, and a large number of samples were selected and studied, we can expect a chi square of this magnitude in fewer than 0.1 percent of those samples. Thus, the probability of obtaining a chi square of this magnitude is less than 0.001 if random sampling has been used and if there is no relationship in the population. We report this finding by saying that the relationship is statistically significant at the .001 level. Because it is so improbable that the observed relationship could have resulted from sampling error alone, we are likely to reject the null hypothesis and assume that a relationship does, in fact, exist between the two variables.

Many measures of association can be tested for statistical significance in a similar manner. Standard tables of values permit us to determine whether a given association is statistically significant and at what level.

Cautions in Interpreting Statistical Significance

Tests of significance provide an objective yardstick against which to estimate the significance of associations between variables. They assist us in ruling out associations that may not represent genuine relationships in the population under study. However, the researcher who uses or reads reports of significance tests should be aware of certain cautions in their interpretation.

First, we have been discussing tests of statistical significance; there are no objective tests of substantive significance. Thus, we may be legitimately convinced that a given association is not due to sampling error but still assert, without fear of contradiction, that two variables are only slightly related to each other. Recall that sampling error is an inverse function of sample size: the larger the sample, the smaller the expected error. Thus, a correlation of, say, .1 might well be significant (at a given level) if discovered in a large sample, whereas the same correlation between the same two variables would not be significant if found in a smaller sample. Of course, that makes perfect sense if one understands the basic logic of tests of significance: In the larger sample, there is less chance that the correlation between the same two variables would not be significant if found in a smaller sample. Of course, that makes perfect sense if one understands the basic logic of tests of significance: In the larger sample, there is less chance that the correlation is simply the product of sampling error.

Consider, for example, Table 13.12, in which 20 cases are distributed in the same proportions across row and column categories as in Table 13.11. In each table, 83 percent of women report avoidance behavior (10 out of 12 in
Table 13.12 Hypothetical Illustration of Chi-Square Sensitivity to Sample Size

<table>
<thead>
<tr>
<th>I. Expected Cell Frequencies</th>
<th>Men</th>
<th>Women</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avoid areas*</td>
<td>5.6</td>
<td>8.4</td>
<td>14</td>
</tr>
<tr>
<td>Do not avoid areas</td>
<td>2.4</td>
<td>3.6</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
<td>12</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>II. Observed Cell Frequencies</th>
<th>Men</th>
<th>Women</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avoid areas</td>
<td>4</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>Do not avoid areas</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
<td>12</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>III. (Observed − Expected)² ÷ Expected</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avoid areas</td>
<td>0.46</td>
<td>0.30</td>
</tr>
<tr>
<td>Do not avoid areas</td>
<td>1.07</td>
<td>0.71</td>
</tr>
</tbody>
</table>

*"Is there any area around here—that is, within a city block—that you avoid at night because of crime?"

Table 13.12, and 50 out of 60 in Table 13.11). But with one-fifth the number of cases in Table 13.12, the computed value of chi square is only one-fifth that obtained in Table 13.11. Consulting the distribution of chi-square values (inside back cover), we see that the probability of obtaining a chi square of 2.54 with 1 degree of freedom lies between .1 and .2. Thus, if there is no relationship between these two variables, we can expect to obtain a chi square of this size in 10–20 percent of samples drawn. Most researchers would not reject the null hypothesis of no relationship in this case.

We mentioned the point illustrated by this example in Chapter 12 because sample size is especially important for applied studies that use experimental designs. With small sample sizes, even moderately large differences might result from sampling error. Because randomized experiments are costly and time consuming, they are often conducted with relatively small numbers of subjects in experimental and control groups. With small numbers of subjects, only large differences in outcome variables will be statistically significant.

The distinction between statistical and substantive significance is perhaps best illustrated by those cases in which there is absolute certainty that observed differences cannot be a result of sampling error. That is the case when we observe an entire population. Suppose we are able to learn the age and gender of every murder victim in the United States for 1996. For argument’s sake, let’s assume that the average age of male murder victims is 25, as compared with, say, 26 for female victims. Because we have the ages of all murder victims, there is no question of sampling error. We know with certainty that the female victims are older than their male counterparts. At the same time, we can say that the difference is of no substantive significance. We conclude, in fact, that they are essentially the same age.

Second, lest you be misled by this hypothetical example, statistical significance should not be calculated on relationships observed in data collected from whole populations. Remember, tests of statistical significance measure the likelihood of relationships between variables being only a product of sampling error, which,
of course, assumes that data come from a sample. If there’s no sampling, there’s no sampling error.

Third, tests of significance are based on the same sampling assumptions we used to compute confidence intervals. To the extent that these assumptions are not met by the actual sampling design, the tests of significance are not strictly legitimate.

In practice, tests of statistical significance are frequently used inappropriately. If you were to review any given issue of an academic journal in criminal justice, we’d be willing to bet you would find one or more of these technically improper uses:

- Tests of significance computed for data representing entire populations
- Tests based on samples that do not meet the required assumptions of probability sampling
- Tests applied to measures of association that have been computed in violation of the assumptions made by those measures (for example, Pearson product-moment correlations computed from ordinal data)
- Interpretation of statistical significance as a measure of association (a “relationship” of \( p \leq .001 \) is “stronger” than one of \( p \leq .05 \))

We do not mean to suggest a “purist” approach by these comments. We encourage you to use any statistical technique—any measure of association or any test of significance—on any set of data if it will help you understand your data. In doing so, however, you should recognize what measures of association and statistical significance can and cannot tell you, as well as the assumptions required for various measures. Any individual statistic or measure tells only part of the story, and you should try to learn as much of the story as you can.

### Visualizing Discernible Differences

In this spirit, Michael Maltz proposes a compromise in the use of statistical significance tests on populations and on samples that violate the assumptions of probability sampling. Acting on a suggestion by criminologist Alfred Blumstein, Maltz (1994:440) advises researchers to use the phrase “statistically discernible difference” rather than “statistically significant difference” when using significance tests on inappropriate samples or populations. A **statistically discernible difference** is one that would be considered statistically significant if found in a random sample. In another example, Lange et al. (2005) use the phrase “statistically reliable difference.” Being careful with our language in this way alerts readers that, although significance tests are not technically appropriate, they can be used as a gauge to show the magnitude of deviation from the null hypothesis.

Researchers can also use the Maltz and Zawitz technique for presenting bounded point estimates to assess relationships between two or more variables. For example, Maxfield and associates (Maxfield, Weiler, and Widom, 2000) compared self-reports of arrests to actual records of arrests for Cathy Spatz Widom’s (1992) sample of child abuse/neglect victims and matched controls. Their research question centered on the relationship between two measures of offending: official records of arrests and self-reported arrests. Self-report measures, gathered through interviews with almost 1200 subjects, were compared with arrest records obtained from local, state, and federal law enforcement agencies.

Recall from earlier chapters that Widom’s subjects were selected purposively: Abuse/neglect victims were identified through court records, and controls were individually matched to victims. Because this sample can in no way be considered random, tests of statistical significance are not technically appropriate.

Figure 13.6 presents point and interval estimates of self-reported arrests for different subgroups of subjects who did not have an official arrest record. The first point estimate shows that, of the 628 subjects with no official arrest record, 21 percent told interviewers that they had been arrested. This estimate is bracketed by 95 percent confidence intervals, showing
the point estimate plus or minus 2 standard errors.

Now consider the separate estimates for abuse/neglect victims and their controls. A larger percentage of abuse/neglect victims (28 percent) self-reported arrests compared with control subjects (14 percent). Furthermore, when these point estimates are bounded by 2 standard errors, the “tails” of the two estimates do not overlap. This means that the estimates of self-reported arrests for abuse/neglect victims are more than 2 standard errors higher than those for control subjects. The last pair of subgroupings compares self-reported arrests by black and white subjects, regardless of whether they were abuse/neglect victims or controls. In this case, the tails of bounded estimates for the two groups overlap, indicating that self-reported arrests for black and white subjects are less than 2 standard errors apart.

Because the analysis shown in Figure 13.6 is not based on a random sample, it violates the assumptions of statistical significance tests. However, it’s still possible to assess the relative size of differences in self-reported arrests by comparing bounded point estimates. Following the suggestion of Blumstein and Maltz, we can make the following observation: There is a statistically discernible difference in self-reported arrests between abuse/neglect victims and control subjects, but not between black and white subjects.

**Main Points**

- Descriptive statistics are used to summarize data under study.
- A frequency distribution shows the number of cases that have each of the attributes of a given variable.
- Measures of central tendency reduce data to an easily manageable form, but they do not convey the detail of the original data.
- Measures of dispersion give a summary indication of the distribution of cases around an average value.
- Rates are descriptive statistics that standardize some measure for comparative purposes.
- Bivariate analysis and subgroup comparisons examine some type of relationship between two variables.
The rules of thumb in interpreting bivariate percentage tables are (1) “percentage down” and “read across” in making the subgroup comparisons or (2) “percentage across” and “read down” in making subgroup comparisons.

Multivariate analysis is a method of analyzing the simultaneous relationships among several variables and may be used to more fully understand the relationship between two variables.

Many measures of association are based on a proportionate reduction of error (PRE) model, which measures improvement in predictions about one variable, given information about a second variable.

Lambda (λ) and gamma (γ) are PRE-based measures of association for nominal and ordinal variables, respectively.

Pearson’s product-moment correlation (r) is a measure of association used in the analysis of two interval or ratio variables.

Regression equations are computed on the basis of a regression line—the geometric line that represents, with the least amount of discrepancy, the actual location of points in a scattergram.

The equation for a regression line predicts the values of a dependent variable on the basis of values of one or more independent variables.

Inferential statistics are used to estimate the generalizability of findings arrived at in the analysis of a sample to the larger population from which the sample has been selected.

Inferences about some characteristic of a population, such as the percentage that favors gun control laws, must contain an indication of a confidence interval (the range within which the value is expected to be—for example, between 45 and 55 percent favor gun control) and an indication of the confidence level (the likelihood that the value does fall within that range—for example, 95 percent confidence).

Tests of statistical significance estimate the likelihood that an association as large as the observed one could result from normal sampling error if no such association exists between the variables in the larger population.

Statistical significance must not be confused with substantive significance, which means that an observed association is strong, important, or meaningful.

Tests of statistical significance, strictly speaking, make assumptions about data and methods that are almost never satisfied completely by real social research. Claiming a “statistically discernible relationship” is more appropriate when assumptions are not satisfied.

**Key Terms**

<table>
<thead>
<tr>
<th>Average Proportionate</th>
<th>Bivariate analysis</th>
<th>Contingency table</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Review Questions and Exercises**

1. Using the data in the accompanying table, construct and interpret tables showing:
   a. The bivariate relationship between age and attitude toward capital punishment
   b. The bivariate relationship between political orientation and attitude toward capital punishment
   c. The multivariate relationship linking age, political orientation, and attitude toward capital punishment

<table>
<thead>
<tr>
<th>Age</th>
<th>Political Orientation</th>
<th>Attitude Toward Capital Punishment</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young</td>
<td>Conservative</td>
<td>Favor</td>
<td>90</td>
</tr>
<tr>
<td>Young</td>
<td>Conservative</td>
<td>Oppose</td>
<td>10</td>
</tr>
<tr>
<td>Young</td>
<td>Liberal</td>
<td>Favor</td>
<td>60</td>
</tr>
<tr>
<td>Young</td>
<td>Liberal</td>
<td>Oppose</td>
<td>40</td>
</tr>
<tr>
<td>Old</td>
<td>Conservative</td>
<td>Favor</td>
<td>60</td>
</tr>
<tr>
<td>Old</td>
<td>Conservative</td>
<td>Oppose</td>
<td>10</td>
</tr>
<tr>
<td>Old</td>
<td>Liberal</td>
<td>Favor</td>
<td>15</td>
</tr>
<tr>
<td>Old</td>
<td>Liberal</td>
<td>Oppose</td>
<td>15</td>
</tr>
</tbody>
</table>

2. Distinguish between measures of association and tests of statistical significance.

3. Here are our answers to the question about statistical significance relating to Figure 13.5:
Fifteen years show no significant change in violent victimization; significant increases are shown for 3 years; violent victimization decline significantly in 5 years. The increase from 1976 to 1977 and the decrease from 1979 to 1980 are close; notice the edge of the 95 percent confidence interval borders the no-change line. We recommend that you read the Maltz and Zawitz publication. It’s listed in the bibliography and can be downloaded from the Bureau of Justice Statistics website at http://www.ojp.usdoj.gov/bjs/abstract/dvctue.htm.

Online Sessions for Chapter 13

Online Sessions present further examples of selected material covered in each chapter. Your instructor may choose to use these for in-class discussion (as I do with my students) or you can access them at home. To ensure that you always have access to live, correct links for the websites described here, we provide the necessary links on the companion website for Research Methods for Criminal Justice and Criminology at http://www.thomsonedu.com/criminaljustice/maxfield. Once at the companion website, select this specific chapter, click on “Chapter Resources,” then click on “Web Links.”

Statistics Tutorials and Practice

Two sites offer general statistics information and access to interactive analysis. The first is especially interesting. We referred to STeve’s Attempt to Teach Statistics in Chapter 8 on sampling. This URL will take you to an overview of his site. As a statistical consultant to health-care professionals, Steve Simon has taken special efforts to explain how statistics are used and how they can be interpreted. Although the public health examples may seem unfamiliar, the overall guidance here is excellent. http://www.childrensmercy.org/stats/overview.asp http://home.clara.net/sisa/index.htm

Professor PollKatz Explains It All for You

We also mentioned the PollKatz site in an earlier chapter. Return to this site and review the section on “Rudimentary Statistics,” for a good overview of sampling error and confidence intervals, together with the role they play in statistical inference. http://www.pollkatz.homestead.com/

Summary Data Analysis with Spreadsheets

Return to the BJS web page on “Key Crime and Justice Facts at a Glance.” Scroll down to find graphs for many domains—trends in crime, arrests, convictions; information about people under correctional supervision. Clicking on a graph produces a large version of it; then clicking on the larger graph brings you to tables of data. Using this page, or other BJS pages that link you to spreadsheets of data, offers many opportunities to practice computing simple statistics. Use a spreadsheet program to alternately compute column and row percentages for sample tables. Then write out statements interpreting the percentages. This kind of practice is readily available through the BJS website, and extremely helpful in learning how to interpret descriptive and bivariate statistics. http://www.ojp.usdoj.gov/bjs/glance.htm

Online Data Analysis

Sometimes, organizing data files and reading them into statistical analysis programs can be especially difficult and time consuming. This is particularly true for large, complex data files like the NCVS or NIBRS. An alternative is the online data analysis system maintained by the National Archive of Criminal Justice Data. Here you can access and do statistical analysis on a large variety of data files. Select a victim survey and one source using data from police records. Or examine data on capital punishment. Each year the number of data files available for online analysis increases. With this resource you can analyze NIBRS data from your favorite WiFi-equipped coffee shop. http://www.icpsr.umich.edu/NACJD/das.html

Additional Readings

Babcock, Earl, Halley, Fred, and Zaino, Jeanne, Adventures in Social Research (Newbury Park, CA: Pine Forge Press, 2000). This book introduces you to data analysis through SPSS, a widely used computer program for statistical analysis. Several of the basic techniques described in this chapter are illustrated and discussed further.

Finkelstein, Michael O., and Levin, Bruce, Statistics for Lawyers (New York: Springer-Verlag, 1990). Law school trains people to be analytic, but few lawyers know much about statistics. This book provides a straightforward explanation of many basic statistical concepts. Examples are drawn from actual cases to illustrate how to
calculate and interpret statistics. In addition, readers gain insight into how to reason with statistics.

Miller, Jane E. *The Chicago Guide to Writing about Numbers* (Chicago: University of Chicago Press, 2004). This is an excellent guide to thinking about and interpreting results of data analysis. The book is helpful as an introduction for beginning researchers and as a reference for more experienced analysts.


Weisburd, David, and Britt, Chester, *Statistics in Criminal Justice*, 2nd ed. (Belmont, CA: Wadsworth, 2002). This text presents an excellent basic introduction to statistics using examples from criminal justice.
Glossary

Aggregate  Groups of units—people, prisons, courtrooms, or stolen autos, for example. Although criminal justice professionals are usually most concerned with individual units, social science searches for patterns that are reflected in aggregations of units. For example, a probation officer focuses on probation clients as individuals, whereas a social scientist focuses on groups of probation clients, or aggregates. See Chapter 1.

Anonymity  The identity of a research subject is not known, and it is therefore impossible to link data about a subject to an individual’s name. Anonymity is one tool for addressing the ethical issue of privacy. Compare to confidentiality. See Chapter 3.

Attributes  Characteristics of persons or things. See variables and Chapter 1.

Average  An ambiguous term that generally suggests typical or normal. The mean, median, and mode are specific mathematical averages. See Chapter 13.

Binomial variable  A variable that has only two attributes is binomial. Gender is an example; it has the attributes male and female. See Chapter 8.

Bivariate analysis  The analysis of two variables simultaneously for the purpose of determining the empirical relationship between them. The construction of a simple percentage table and the computation of a simple correlation coefficient are examples of bivariate analyses. See Chapter 13.

Case-oriented research  A research strategy in which many cases are examined to understand a comparatively small number of variables. Examples include experiments (Chapter 7) and surveys (Chapter 9).

Case study  A research strategy in which the researcher’s attention centers on an in-depth examination of one or a few cases on many dimensions. Case studies can be exploratory, descriptive, or explanatory. Case studies can also be used in evaluation research. See Chapters 7 and 12.

Classical experiment  A research design well suited to inferring cause, the classical experiment involves three major pairs of components: (1) independent and dependent variables, (2) pretesting and posttesting, and (3) experimental and control groups, with subjects randomly assigned to one group or the other. See Chapter 7.

Closed-ended questions  Survey questions in which the respondent is asked to select an answer from a list provided by the researcher. Closed-ended questions are especially popular in surveys because they provide a greater uniformity of responses and are more easily analyzed than closed-ended questions. See Chapter 9.

Cluster sample  A multistage sample in which natural groups (clusters) are sampled initially, with the members of each selected group being subsampled afterward. For example, you might select a sample of municipal police departments from a directory, get lists of the police officers at all the selected departments, then draw samples of officers from each. See Chapter 8.

Cohort study  A study in which some specific group is studied over time, although data may be collected from different members in each set of observations. See Chapter 4.

Computer-assisted interviewing  Survey research by computer, in which questionnaires are presented on computer screens instead of paper. In computer-assisted personal interviewing, an interviewer reads items from the computer screen and keys in responses. In computer-assisted self-interviewing, respondents read (silently) items on the screen of a laptop computer and key in their answers. Another variation is audio-assisted self-interviewing, whereby respondents hear questions through headphones, then key in their answers. Both types of self-interviewing are especially useful for sensitive questions like self-reports. See Chapter 9.

Concept  The words or symbols in language that we use to represent mental images. “Crime,” for example, is a concept that represents our mental images of violence and other acts that are prohibited and punished by government. We use conceptual definitions to specify the meaning of concepts. Compare to conception. See Chapter 5.

Conception  The mental images we have that represent our thoughts about things we routinely encounter. We use the word speeding (a concept) to represent our mental image (conception) of traveling above the posted speed limit. See Chapter 5.

Conceptual definition  Defining concepts by using other concepts. Concepts are abstract—the words and symbols that are used to represent mental images of things and ideas. This means that a conceptual definition uses words and symbols to define concepts. In practice, conceptual definitions represent explicit statements of what a researcher means by a concept.
A conceptual definition of “prior record” might be “recorded evidence of one or more convictions for a criminal offense.” See also operational definition and Chapter 5.

**Conceptualization** The mental process whereby fuzzy and imprecise notions (concepts) are made more specific and precise. So you want to study fear of crime? What do you mean by fear of crime? Are there different kinds of fear? What are they? See Chapters 5 and 6.

**Confidence interval** The range of values within which a population parameter is estimated to lie. A survey, for instance, may show that 40 percent of a sample favors a ban on handguns. Although the best estimate of the support that exists among all people is also 40 percent, we do not expect it to be exactly that. We might, therefore, compute a confidence interval (for example, from 35 to 45 percent) within which the actual percentage of the population probably lies. Note that it is necessary to specify a confidence level in connection with every confidence interval. See Chapters 8 and 13.

**Confidence level** The estimated probability that a population parameter lies within a given confidence interval. Thus, we might be 95 percent confident that between 35 and 45 percent of all residents of California favor an absolute ban on handguns. See Chapters 8 and 13.

**Confidentiality** Researchers know the identity of a research subject but promise not to reveal any information that can be attributed to an individual subject. Anonymity is similar, but sometimes researchers need to know subjects’ names in order to link information from different sources. Assuring confidentiality is one way of meeting our ethical obligation to not harm subjects. See Chapter 3.

**Construct validity** (1) The degree to which a measure relates to other variables as expected within a system of theoretical relationships. See Chapter 5. (2) How well an observed cause-and-effect relationship represents the underlying causal process a researcher is interested in. See Chapters 4 and 7. Also see validity threats.

**Content validity** The degree to which a measure covers the range of meanings included within the concept. See Chapter 5.

**Contingency table** A format for presenting the relationship among variables in the form of percentage distributions. See Chapter 13.

**Control group** In experimentation, a group of subjects to whom no experimental stimulus is administered and who should resemble the experimental group in all other respects. The comparison of the control group and the experimental group at the end of the experiment indicates the effect of the experimental stimulus. See Chapter 7.

**Criterion-related validity** The degree to which a measure relates to some external criterion. For example, the validity of self-report surveys of drug use can be shown by comparing survey responses to laboratory tests for drug use. See Chapter 5.

**Cross-sectional study** A study based on observations that represent a single point in time. Compare to longitudinal study. See Chapter 4.

**Deductive reasoning** The logical model in which specific expectations of hypotheses are developed on the basis of general principles. Starting from the general principle that all deans are meanies, you might anticipate that this one won’t let you change courses. That anticipation would be the result of deduction. See also inductive reasoning and Chapter 1.

**Dependent variable** The variable assumed to depend on or be caused by another variable (called the independent variable). If you find that sentence length is partly a function of the number of prior arrests, then sentence length is being treated as a dependent variable. See Chapters 1 and 7.

**Descriptive statistics** Statistical computations that describe either the characteristics of a sample or the relationship among variables in a sample. Descriptive statistics summarize a set of sample observations, whereas inferential statistics move beyond the description of specific observations to make inferences about the larger population from which the sample observations were drawn. See Chapter 13.

**Dimension** A specifiable aspect or characteristic of a concept. See Chapter 5.

**Dispersion** The distribution of values around some central value, such as an average. The range is a simple measure of dispersion. Thus, we may report that the mean age of a group is 37.9 and the range is from 12 to 89. See Chapter 13.

**Disproportionate stratified sampling** Deliberately drawing a sample that over- or underrepresents some characteristic of a population. We may do this to ensure that we obtain a sufficient number of uncommon cases in our sample. For example, believing violent crime to be more common in large cities, we might oversample urban residents to obtain a specific number of crime victims. See Chapter 8.

**Ecological fallacy** Erroneously drawing conclusions about individuals based solely on the observation of groups. See Chapter 4.
Empirical  From experience. Social science is said to be empirical when knowledge is based on what we experience. See Chapter 1.

Environmental survey  Structured observations undertaken in the field and recorded on specially designed forms. Note that interview surveys record a respondent’s answers to questions, while environmental surveys record what an observer sees in the field. For example, a community organization may conduct periodic environmental surveys to monitor neighborhood parks—whether facilities are in good condition, how much litter is present, and what kinds of people use the park. See Chapter 10.

Equal probability of selection method (EPSEM)  A sample design in which each member of a population has the same chance of being selected into the sample. See Chapter 8.

Evaluation research  An example of applied research, evaluation involves assessing the effects of some program or policy action, usually in connection with the goals of that action. Determining whether a sex-offender treatment program attained its goal of reducing recidivism by participants would be an example. Compare to problem analysis. See Chapter 12.

Evidence-based policy  Using data and other sources of information to formulate and evaluate justice policy. This usually means planning justice actions based on evidence of need, such as deploying police patrols to crime hot spots. It also includes assessing the results of justice policy, such as measuring any change in recidivism among a group of offenders processed through drug court. See Chapter 12.

Experimental group  In experimentation, a group of subjects who are exposed to an experimental stimulus. Subjects in the experimental group are normally compared to subjects in a control group to test the effects of the experimental stimulus. See Chapter 7.

External validity  Whether a relationship observed in a specific population, at a specific time, in a specific place would also be observed in other populations, at other times, in other places. External validity is concerned with generalizability from a relationship observed in one setting to the same relationship in other settings. Replication enhances external validity. See Chapters 4 and 7.

Face validity  The quality of an indicator that makes it seem a reasonable measure of some variable. That sentence length prescribed by law is some indication of crime seriousness seems to make sense without a lot of explanation; it has face validity. See Chapter 5.

Focus groups  Small groups (of 12 to 15) engaged in a guided discussion of some topic. Participants selected are from a homogeneous population. Although focus groups cannot be used to make statistical estimates about a population, members are nevertheless selected to represent a target population. Focus groups are most useful in two situations: (1) when precise generalization to a larger population is not necessary, and (2) when focus group participants and the larger population they are intended to represent are relatively homogeneous. See Chapter 9.

Frequency distribution  A description of the number of times the various attributes of a variable are observed in a sample. The report that 53 percent of a sample were men and 47 percent were women is a simple example of a frequency distribution. Another example is the report that 15 of the cities studied had populations of less than 10,000, 23 had populations between 10,000 and 25,000, and so forth. See Chapter 13.

Generalizability  That quality of a research finding that justifies the inference that it represents something more than the specific observations on which it was based. Sometimes, this involves the generalization of findings from a sample to a population. Other times, it is a matter of concepts: If you are able to discover why people commit burglaries, can you generalize that discovery to other crimes as well? See Chapter 7.

Grounded theory  A type of inductive theory that is based on (grounded in) field observation. The researcher makes observations in natural settings, then formulates a tentative theory that explains those observations. See Chapter 2.

Hypothesis  An expectation about the nature of things derived from a theory. It is a statement of something that ought to be observed in the real world if the theory is correct. See deduction and Chapters 2 and 7.

Hypothesis testing  The determination of whether the expectations that a hypothesis represents are indeed found in the real world. See Chapters 2 and 7.

Idiographic  A mode of causal reasoning that seeks detailed understanding of all factors that contribute to a particular phenomenon. Police detectives trying to solve a particular case use the idiographic mode of explanation. Compare to nomothetic. See Chapters 1 and 4.

Impact assessment  A type of applied research that seeks to answer the question: Did a public program have the intended effect on the problem it was meant to address? If, for example, a new burglary prevention program has the goal of reducing burglary in a particular neighborhood, an impact assessment would try to determine whether burglary was, in fact,
reduced as a result of the new program. Compare to process evaluation. See Chapter 12.

**Incident-based measure** Refers to crime measures that express characteristics of individual crime incidents. The FBI Supplementary Homicide Reports are well-known examples, reporting details on each homicide incident. Compare to summary-based measure. See Chapter 6.

**Independent variable** An independent variable is presumed to cause or determine a dependent variable. If we discover that police cynicism partly a function of years of experience, then experience is the independent variable and cynicism is the dependent variable. Note that any given variable might be treated as independent in one part of an analysis and dependent in another part of the analysis. Cynicism might become an independent variable in the explanation of job satisfaction. See Chapters 1 and 7.

**Inductive reasoning** The logical model in which general principles are developed from specific observations. Having noted that teenagers and crime victims are less supportive of police than older people and nonvictims are, you might conclude that people with more direct police contact are less supportive of police and explain why. That would be an example of induction. See also deductive reasoning and Chapter 1.

**Inferential statistics** The body of statistical computations relevant to making inferences from findings based on sample observations to some larger population. See also descriptive statistics and Chapter 13.

**Internal validity** Whether observed associations between two (or more) variables are, in fact, causal associations or are due to the effects of some other variable. The internal validity of causal statements may be threatened by an inability to control experimental conditions. See also validity threats and Chapters 4 and 7.

**Intersubjective agreement** That quality of science (and other inquiries) whereby two different researchers studying the same problem arrive at the same conclusion. Ultimately, this is the practical criterion for what is called objectivity. We agree that something is “objectively true” if independent observers with different subjective orientations conclude that it is “true.” See Chapter 2.

**Interval measure** A level of measurement that describes a variable whose attributes are rank ordered and have equal distances between adjacent attributes. The Fahrenheit temperature scale is an example of this because the distance between 17 and 18 is the same as that between 89 and 90. See also nominal measure, ordinal measure, and ratio measure and Chapter 5.

**Latent content** As used in connection with content analysis, this term describes the underlying meaning of communications as distinguished from their manifest content. See Chapter 11.

**Level of significance** In the context of tests of statistical significance, the degree of likelihood that an observed, empirical relationship could be attributable to sampling error. A relationship is significant at the .05 level if the likelihood of its being only a function of sampling error is no greater than 5 out of 100. See Chapter 13.

**Longitudinal study** A study design that involves the collection of data at different points in time, as contrasted to a cross-sectional study. See also trend study, cohort study, and panel study and Chapter 4.

**Manifest content** In connection with content analysis, the concrete terms contained in a communication, as distinguished from latent content. See Chapter 11.

**Mean** An average, computed by summing the values of several observations and dividing by the number of observations. If you now have a grade-point average of 4.0 based on 10 courses, and you get an F in this course, then your new grade-point (mean) average will be 3.6. See Chapter 13.

**Median** Another average, representing the value of the “middle” case in a rank-ordered set of observations. If the ages of five people are 16, 17, 20, 54, and 88, then the median is 20. (The mean is 39.) See Chapter 13.

**Mode** Still another average, representing the most frequently observed value or attribute. If a sample contains 1000 residents of California, 275 from New Jersey, and 33 from Minnesota, then California is the modal category for residence. See Chapter 13.

**Multivariate analysis** The analysis of the simultaneous relationships among several variables. Examining simultaneously the effects of age, gender, and city of residence on robbery victimization is an example of multivariate analysis. See Chapter 13.

**Nominal measure** A level of measurement that describes a variable whose different attributes are only different, as distinguished from ordinal, interval, and ratio measures. Gender is an example of a nominal measure. See Chapter 5.

**Nomothetic** A mode of causal reasoning that tries to explain a number of similar phenomena or situations. Police crime analysts trying to explain patterns of auto thefts, or burglaries, or some other offense use nomothetic reasoning. Compare to idiographic. See Chapters 1 and 4.
Nonprobability sample  A sample selected in some fashion other than those suggested by probability theory. Examples are purposive, quota, and snowball samples. See Chapter 8.

Nonsampling error  Imperfections of data quality that are a result of factors other than sampling error. Examples are misunderstandings of questions by respondents, erroneous recordings by interviewers and coders, and data entry errors. See Chapter 13.

Null hypothesis  In connection with hypothesis testing and tests of statistical significance, the hypothesis that suggests there is no relationship between the variables under study. You may conclude that the two variables are related after having statistically rejected the null hypothesis. See Chapter 13.

Objectivity  Doesn’t exist. See also intersubjective agreement and Chapter 2.

Open-ended questions  Questions for which the respondent is asked to provide his or her own answers. See Chapter 9.

Operational definition  Specifying what operations should be performed to measure a concept. The operational definition of “prior record” might be “Consult the county (or state or FBI) criminal history records information system. Count the number of times a person has been convicted of committing a crime.” See Chapter 5.

Operationalization  One step beyond conceptualization. Operationalization is the process of developing operational definitions that describe how actual measurements will be made. See Chapters 4 and 5.

Ordinal measure  A level of measurement that describes a variable whose attributes may be rank ordered along some dimension. An example is socioeconomic status as composed of the attributes high, medium, and low. See also nominal measure, interval measure, and ratio measure and Chapter 5.

Panel study  A type of longitudinal study in which data are collected from the same subjects (the panel) at several points in time. See Chapter 4.

Paradigm  A fundamental perspective or model that organizes our view of the world. Thomas Kuhn (1970) coined this term in the philosophy of science. Paradigms affect how we select and define problems for research, together with the methods we use in conducting research. Compare to theory. See Chapter 2.

Population  All people, things, or other elements we wish to represent. Researchers often study only a subset or sample of a population, then generalize from the people, things, or other elements actually observed to the larger population of all people, things, or elements. See Chapter 8.

Population parameter  The summary description of a particular variable in the population. For example if the mean age of all professors at your college is 43.7, then 43.7 is the population parameter for professors’ mean age. Compare to sample statistic and sampling distribution. See Chapter 8.

Probabilistic  A type of causal reasoning that certain factors make outcomes more or less likely to happen. Having an arrest as a juvenile makes it more likely that someone will be arrested as an adult. See Chapter 4.

Probability sample  The general term for a sample selected in accord with probability theory, typically involving some random selection mechanism. Specific types of probability samples include area probability sample, equal probability of selection method (EPSEM), simple random sample, and systematic sample. See Chapter 8.

Problem analysis  Using social science research methods to assess the scope and nature of a problem, to plan and select actions to address the problem. For example, examining patterns of auto theft to decide what preventive and enforcement strategies should be pursued is an example of problem analysis. Compare to evaluation research. See Chapter 12.

Problem solving  An example of applied research that combines elements of evaluation and policy analysis. The most widely known approach to problem solving in policing is the SARA model, which stands for scanning, analysis, response, and assessment. See Chapter 12.

Process evaluation  A type of applied research that seeks to answer the question: Was a public program implemented as intended? For example, a burglary prevention program might seek to reduce burglaries by having crime prevention officers meet with all residents of some target neighborhood. A process evaluation would determine whether meetings with neighborhood residents were taking place as planned. Compare to impact assessment. See Chapter 12.

Proportionate reduction of error (PRE)  A logical model for assessing the strength of a relationship by asking how much knowing values on one variable might reduce our errors in guessing values on the other. For example, if we know how much education people have, we can improve our ability to guess how much they earn, thus indicating that there is a relationship between the two variables. See Chapter 13.

Prospective  A type of longitudinal study that follows subjects forward in time. “How many people who were sexually abused as children are convicted
of a sexual offense as an adult?" is an example of a prospective question. Compare to retrospective. See Chapter 4.

**Purposive sample** A type of nonprobability sample in which you select the units to be observed on the basis of your own judgment about which ones will be best suited to your research purpose. For example, if you were interested in studying community crime prevention groups affiliated with public schools and groups affiliated with religious organizations, you would probably want to select a purposive sample of school- and church-affiliated groups. Most television networks use purposive samples of voting precincts to project winning candidates on election night; precincts that always vote for winners are sampled. See Chapter 8.

**Quasi-experiment** A research design that includes most, but not all, elements of an experimental design. Quasi- means "sort of," and a quasi-experiment is sort of an experiment. Two general classes of quasi-experiments are nonequivalent groups and time-series designs. Compare to classical experiment and see Chapter 7.

**Questionnaire** A document that contains questions and other types of items designed to solicit information appropriate to analysis. Questionnaires are used primarily in survey research and also in field research. See Chapter 9.

**Quota sample** A type of nonprobability sample in which units are selected into the sample on the basis of prespecified characteristics, so that the total sample will have the same distribution of characteristics as are assumed to exist in the population being studied. See Chapter 8.

**Randomization** A technique for randomly assigning experimental subjects to experimental groups and control groups. See Chapter 7.

**Range** A measure of dispersion, the distance that separates the highest and lowest values of a variable in some set of observations. In your class, for example, the range of ages might be from 17 to 37. See Chapter 13.

**Ratio measure** A level of measurement that describes a variable whose attributes have all the qualities of nominal, ordinal, and interval measures and in addition are based on a "true zero" point. Length of prison sentence is an example of a ratio measure. See Chapter 5.

**Reductionism** A fault of some researchers: a strict limitation (reduction) of the kinds of concepts to be considered relevant to the phenomenon under study. See Chapter 4.

**Regression analysis** A method of data analysis in which the relationships among variables are represented in the form of an equation called a regression equation. See Chapter 13.

**Reification** The process of regarding as real things that are not real. This is usually a problem in measurement. See Chapter 5.

**Reliability** That quality-of-measurement standard whereby the same data would have been collected each time in repeated observations of the same phenomenon. We would expect that the question "Did you see a police officer in your neighborhood today?" would have higher reliability than the question "About how many times in the past 6 months have you seen a police officer in your neighborhood?" This is not to be confused with validity. See Chapter 5.

**Replication** Repeating a research study to test the findings of an earlier study, often under slightly different conditions or for a different group of subjects. Replication results either support earlier findings or cause us to question the accuracy of an earlier study. See Chapter 1.

**Representativeness** That quality of a sample having the same distribution of characteristics as the population from which it was selected. Representativeness is enhanced by probability sampling and provides for generalizability and the use of inferential statistics. See Chapter 8.

**Retrospective** A type of longitudinal study that looks backward, asking subjects to recall events that happened earlier in their lives or tracing official records of someone’s previous actions. "How many current sex offenders were sexually abused as children?" is a retrospective question. Compare to prospective. See Chapter 4.

**Sample element** That unit about which information is collected and that provides the basis of analysis. Typically, in survey research, elements are people. Other kinds of units can be the elements for criminal justice research—correctional facilities, gangs, police beats, or court cases, for example. See Chapter 8.

**Sample statistic** The summary description of a particular variable in a sample. For example, if the mean age of a sample of 100 professors on your campus is 41.1, then 41.1 is the sample statistic for professors’ age. We usually use sample statistics to estimate population parameters. Compare to sampling distribution. See Chapter 8.

**Sampling distribution** The range or array of sample statistics we would obtain if we drew a very large number of samples from a single population. With random sampling, we expect that the sampling
distribution for a particular statistic (mean age, for example) will cluster around the population parameter for mean age. Furthermore, sampling distributions for larger sample sizes will cluster more tightly around the population parameter. See Chapter 8.

**Sampling frame** That list or quasi-list of units composing a population from which a sample is selected. If the sample is to be representative of the population, it is essential that the sampling frame include all (or nearly all) members of the population. See Chapter 8.

**Sampling units** Like sampling elements, these are things that may be selected in the process of sampling; often, sampling units are people. In some types of sampling, however, we often begin by selecting large groupings of the eventual elements we will analyze. Sampling units is a generic term for things that are selected in some stage of sampling but are not necessarily the objects of our ultimate interest. See Chapter 8.

**Scientific realism** An approach to evaluation that studies what’s called “local causality.” Interest focuses more on how interventions and measures of effect are related in a specific situation. This is different from a more traditional social science interest in finding causal relationships that apply generally to a variety of situations. As explained by Ray Pawson and Nick Tilley (1997), scientific realism is especially useful for evaluating justice programs because it centers on analyzing interventions in local contexts. See Chapters 4, 7, and 12.

**Secondary analysis** A form of research in which the data collected and processed by one researcher are reanalyzed—often for a different purpose—by another. This is especially appropriate in the case of survey data. Data archives are repositories or libraries for the storage and distribution of data for secondary analysis. See Chapter 11.

**Self-report survey** Self-report surveys ask people to tell about crimes they have committed. This method is best for measuring drug use and other so-called victimless crimes. Confidentiality is especially important in self-report surveys. See Chapters 6 and 9.

**Simple random sample** A type of probability sample in which the units composing a population are assigned numbers, a set of random numbers is then generated, and the units that have those numbers are included in the sample. Although probability theory and the calculations it provides assume this basic sampling method, it is seldom used for practical reasons. An alternative is the systematic sample (with a random start). See Chapter 8.

**Snowball sampling** A method for drawing a non-probability sample. Snowball samples are often used in field research. Each person interviewed is asked to suggest additional people for interviewing. See Chapters 8 and 10.

**Stakeholders** Individuals with some interest, or stake, in a specific program. Any particular program may have multiple stakeholders with different interests and goals. See Chapter 12.

**Standard deviation** A measure of dispersion about the mean. Conceptually, the standard deviation represents an “average” deviation of all values relative to the mean. See Chapter 13.

**Standard error** A measure of sampling error, the standard error gives us a statistical estimate of how much a member of a sample might differ from the population we are studying, solely by chance. Larger samples usually result in smaller standard errors. See Chapters 8 and 13.

**Statistical conclusion validity** Whether we can find covariation among two variables. This is the first of three requirements for causal inference (see Chapter 4 for the other two). If two variables do not vary together (covariation), there cannot be a causal relationship between them. See Chapters 4 and 7 for more on statistical conclusion validity. Chapter 13 describes the role of sample size in finding statistical significance, which is conceptually related to statistical conclusion validity.

**Statistical significance** A general term for the unlikeliness that relationships observed in a sample could be attributed to sampling error alone. See also test of statistical significance and Chapter 13.

**Statistically discernible difference** Tests of statistical significance should not normally be used unless our data meet certain assumptions about sampling and variation. Citing a finding as a “statistically discernible difference” allows us to use the tools of statistical inference while alerting readers that the use of these tools is not technically appropriate. See Chapter 13.

**Stratification** The grouping of the units composing a population into homogeneous groups (or strata) before sampling. This procedure, which may be used in conjunction with simple random, systematic, or cluster sampling, improves the representativeness of a sample, at least in terms of the stratification variables. See Chapter 8.

**Summary-based measure** Crime measures that report only total crimes for a jurisdiction or other small area are summary-based measures of crime. The FBI Uniform Crime Reports is one well-known

**Surveillance system** Data routinely collected to monitor some phenomenon—accidents, injuries, communicable disease incidents, drug emergencies—are surveillance measures. U.S. agencies use surveillance measures to obtain estimates of drug use and, more recently, injuries produced by violent crime. See Chapter 6.

**Systematic sampling** A type of probability sample in which every \( k \)th unit in a list is selected for inclusion in the sample—for example, every 25th student in the college directory of students. We compute \( k \) (also called the sampling interval) by dividing the size of the population by the desired sample size. Within certain constraints, systematic sampling is a functional equivalent of simple random sampling and is usually easier to do. Typically, the first unit is selected at random. See Chapter 8.

**Test of statistical significance** A class of statistical computations that indicate the likelihood that the relationship observed between variables in a sample can be attributed to sampling error only. See also inferential statistics and Chapter 13.

**Theory** A theory is a systematic explanation for the observed facts and laws that relate to a particular aspect of life. For example, routine activities theory (see Cohen and Felson, 1979) explains crime as the result of three key elements coming together: a suitable victim, a motivated offender, and the absence of capable guardians. See Chapter 2.

**Trend study** A type of longitudinal study in which a given characteristic of some population is monitored over time. An example is the series of annual Uniform Crime Report totals for some jurisdiction. See Chapter 4.

**Typology** Classifying observations in terms of their attributes. Sometimes referred to as “taxonomies,” typologies are typically created with nominal variables. For example, a typology of thieves might group them according to the types of cars they steal and the types of locations they search to find targets. See Chapter 5.

**Units of analysis** The what or whom being studied. Units of analysis may be individual people, groupings of people (a juvenile gang), formal organizations (a probation department), or social artifacts (crime reports). See Chapter 4.

**Univariate analysis** The analysis of a single variable for purposes of description. Frequency distributions, averages, and measures of dispersion are examples of univariate analysis, as distinguished from bivariate and multivariate analyses. See Chapter 13.

**Validity** (1) Whether statements about cause and effect are true (valid) or false (invalid). See Chapters 4 and 7; also see validity threats. (2) A descriptive term used for a measure that accurately reflects what it is intended to measure. For example, police records of auto theft are more valid measures than police records of shoplifting. It is important to realize that the ultimate validity of a measure can never be proved. Yet we may agree as to its relative validity on the basis of face validity, criterion-related validity, content validity, and construct validity. This must not be confused with reliability. See Chapter 5.

**Validity threats** Possible sources of invalidity, or making false statements about cause and effect. Four categories of validity threats are linked to fundamental requirements for demonstrating cause: statistical conclusion validity, internal validity, construct validity, and external validity (see separate entries in this Glossary). In general, statistical conclusion validity and internal validity are concerned with bias; construct validity and external validity are concerned with generalization. See Chapters 4 and 7.

**Variable-oriented research** A research strategy whereby a large number of variables are studied for one or a small number of cases or subjects. Time-series designs and case studies are examples. See Chapter 7.

**Variables** Logical groupings of attributes. The variable gender is made up of the attributes male and female. See Chapter 1.

**Victim survey** A sample survey that asks people about their experiences as victims of crime. Victim surveys are one way to measure crime, and they are especially valuable for getting information about crimes not reported to police. The National Crime Victimization Survey is an example. See Chapters 6 and 9.
References


References


Farrington, David P., et al. 1996. “Self-Reported Delinquency and a Combined Delinquency Seriousness Scale Based on Boys, Mothers, and Teachers: Concur-


References


References


Painter, Kate. 1991. *An Evaluation of Public Lighting as a Crime Prevention Strategy with a Special Focus on Women and Elderly People*. Manchester, England: Faculty of Economic and Social Studies, University of Manchester.


References


References

453


References


### Distribution of Chi Square

<table>
<thead>
<tr>
<th>Probability</th>
<th>.99</th>
<th>.98</th>
<th>.95</th>
<th>.90</th>
<th>.80</th>
<th>.70</th>
<th>.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>df</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0157</td>
<td>0.0628</td>
<td>0.00393</td>
<td>0.0158</td>
<td>0.0642</td>
<td>0.148</td>
<td>0.455</td>
</tr>
<tr>
<td>2</td>
<td>0.0201</td>
<td>0.0404</td>
<td>0.103</td>
<td>0.211</td>
<td>0.446</td>
<td>0.713</td>
<td>1.386</td>
</tr>
<tr>
<td>3</td>
<td>0.115</td>
<td>0.185</td>
<td>0.352</td>
<td>0.584</td>
<td>1.005</td>
<td>1.424</td>
<td>2.366</td>
</tr>
<tr>
<td>4</td>
<td>0.297</td>
<td>0.429</td>
<td>0.711</td>
<td>1.064</td>
<td>1.649</td>
<td>2.195</td>
<td>3.357</td>
</tr>
<tr>
<td>5</td>
<td>0.554</td>
<td>0.752</td>
<td>1.145</td>
<td>1.610</td>
<td>2.343</td>
<td>3.000</td>
<td>4.351</td>
</tr>
<tr>
<td>6</td>
<td>0.872</td>
<td>1.134</td>
<td>1.635</td>
<td>2.204</td>
<td>3.070</td>
<td>3.828</td>
<td>5.348</td>
</tr>
<tr>
<td>7</td>
<td>1.239</td>
<td>1.564</td>
<td>2.167</td>
<td>2.833</td>
<td>3.822</td>
<td>4.671</td>
<td>6.346</td>
</tr>
<tr>
<td>8</td>
<td>1.646</td>
<td>2.032</td>
<td>2.733</td>
<td>3.490</td>
<td>4.594</td>
<td>5.528</td>
<td>7.344</td>
</tr>
<tr>
<td>9</td>
<td>2.088</td>
<td>2.532</td>
<td>3.325</td>
<td>4.168</td>
<td>5.380</td>
<td>6.393</td>
<td>8.343</td>
</tr>
</tbody>
</table>

For larger values of df, the expression \( x^2 = 2x^2 = 2df - 1 \) may be used as a normal deviate with unit variance, remembering that the probability of \( x^2 \) corresponds with that of a single tail of the normal curve.
<table>
<thead>
<tr>
<th>df</th>
<th>.30</th>
<th>.20</th>
<th>.10</th>
<th>.05</th>
<th>.02</th>
<th>.01</th>
<th>.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.074</td>
<td>1.642</td>
<td>2.706</td>
<td>3.841</td>
<td>5.412</td>
<td>6.635</td>
<td>10.827</td>
</tr>
<tr>
<td>13</td>
<td>15.119</td>
<td>16.985</td>
<td>19.812</td>
<td>22.362</td>
<td>25.472</td>
<td>27.688</td>
<td>34.528</td>
</tr>
<tr>
<td>18</td>
<td>20.601</td>
<td>22.760</td>
<td>25.989</td>
<td>28.869</td>
<td>32.346</td>
<td>34.805</td>
<td>42.312</td>
</tr>
<tr>
<td>19</td>
<td>21.689</td>
<td>23.900</td>
<td>27.204</td>
<td>30.144</td>
<td>33.687</td>
<td>36.191</td>
<td>43.820</td>
</tr>
<tr>
<td>20</td>
<td>22.775</td>
<td>25.038</td>
<td>28.412</td>
<td>31.410</td>
<td>35.020</td>
<td>37.566</td>
<td>45.315</td>
</tr>
<tr>
<td>21</td>
<td>23.858</td>
<td>26.171</td>
<td>29.615</td>
<td>32.671</td>
<td>36.343</td>
<td>38.932</td>
<td>46.797</td>
</tr>
<tr>
<td>22</td>
<td>24.939</td>
<td>27.301</td>
<td>30.813</td>
<td>33.924</td>
<td>37.659</td>
<td>40.289</td>
<td>48.268</td>
</tr>
<tr>
<td>23</td>
<td>26.018</td>
<td>28.429</td>
<td>32.007</td>
<td>35.172</td>
<td>38.968</td>
<td>41.638</td>
<td>49.728</td>
</tr>
<tr>
<td>24</td>
<td>27.096</td>
<td>29.553</td>
<td>33.196</td>
<td>36.415</td>
<td>40.270</td>
<td>42.980</td>
<td>51.179</td>
</tr>
<tr>
<td>25</td>
<td>28.172</td>
<td>30.675</td>
<td>34.382</td>
<td>37.652</td>
<td>41.566</td>
<td>43.941</td>
<td>52.620</td>
</tr>
<tr>
<td>26</td>
<td>29.246</td>
<td>31.795</td>
<td>35.563</td>
<td>38.885</td>
<td>42.856</td>
<td>45.642</td>
<td>54.052</td>
</tr>
<tr>
<td>27</td>
<td>30.319</td>
<td>32.912</td>
<td>36.741</td>
<td>40.113</td>
<td>44.140</td>
<td>46.963</td>
<td>55.476</td>
</tr>
<tr>
<td>28</td>
<td>31.391</td>
<td>34.027</td>
<td>37.916</td>
<td>41.337</td>
<td>45.419</td>
<td>48.278</td>
<td>56.893</td>
</tr>
<tr>
<td>29</td>
<td>32.461</td>
<td>35.139</td>
<td>39.087</td>
<td>42.557</td>
<td>46.932</td>
<td>49.588</td>
<td>58.302</td>
</tr>
<tr>
<td>30</td>
<td>35.530</td>
<td>36.250</td>
<td>40.256</td>
<td>43.773</td>
<td>47.962</td>
<td>50.892</td>
<td>59.703</td>
</tr>
</tbody>
</table>

Source: We are grateful to the Literary Executor of the late Sir Ronald A. Fisher, F.R.S., to Dr. Frank Yates, F.R.S., and to Longman Group Ltd., London, for permission to reprint Table IV from their book Statistical Tables for Biological, Agricultural, and Medical Research (6th Edition, 1974).