1

A Gentle Introduction to Missing Data

As the old saying goes, the only certainties are death and taxes. We would like to add one more to that list: missing data. As any social scientist can attest, missing data are virtually guaranteed in research studies. However, why data are missing, how missing data may affect research outcomes, and what to do about missing data are far less certain. Although the problem of missing data has been addressed in the statistical literature for decades, it still remains a significant conundrum for social scientists who are not methodologists or statistical experts. This state of affairs is at least partially due to their lack of familiarity with the statistical literature on missing data. This lack of familiarity is likely due to a variety of reasons, including a lack of understanding about the importance of missing data and an inability to interpret what is generally a complex and technical literature. The purpose of this book is to bridge the gap between the technical missing data literature and social scientists. The purpose of this introductory chapter is to both familiarize the reader with the concept of missing data and stress the importance of missing data when interpreting research results. In addition, we provide the reader an overview of the remainder of the book.

THE CONCEPT OF MISSING DATA

Broadly, the term *missing data* means that we are missing some type of information about the phenomena in which we are interested. In general, missing data hinder our ability to explain and understand the phenomena that we study. We seek to explain and understand these phenomena by collecting observations. Research results hinge largely on the analyses of these observations or data. This empirical or data-driven approach finds its roots in Sir Francis Bacon's *Novum Organum* Book II (1620), in which he details the inductive method of scientific inquiry. Bacon argued that three Tables of Comparative Instances involving the presence, absence, and degrees of a phenomenon are the hallmarks for discovering the truth about nature. Observations of conditions where the phenomenon of interest is present, where it is absent, and where it is present in varying degrees define and remain hallmarks of the scientific method. When observations are missing for any reason, our ability to understand the nature of the phenomena is reduced, the extent to which is not often known. All three of Bacon's "Tables" may be affected by these missed observations, therefore affecting our ability to either infer or deduce the nature of the phenomenon of interest. We believe, therefore, that missing data in general pose a threat to the validity of scientific inquiry.

THE PREVALENCE OF MISSING DATA

If knowledge about the nature of particular phenomena is adversely affected by any missing data, then one would expect that researchers would attend to the problem of missing data. In other words, both prevention and remediation of missing data should be normative behavior for social scientists. Evidence, empirical and otherwise, suggests that this might not be the case. For example, the American Psychological Association (APA) Task Force on Statistical Inference (Wilkinson and APA Task Force on Statistical Inference, 1999) published recommendations for presenting empirical results in published manuscripts. Recommendations for analyzing and reporting missing data were included, which were:

> Before presenting results, report complications [including] missing data, attrition, and nonresponse. Discuss analytic techniques devised to ameliorate these problems. Describe nonrepresentativeness statistically by reporting pat-

> terns and distributions of missing data and contaminations. Document how the actual analysis differs from the analysis planned before complications arose. The use of techniques to ensure that the reported results are not produced by anomalies in the data (e.g., outliers, points of high influence, nonrandom missing data, selection bias, attrition problems) should be a standard component of all analyses. (p. 597)

Inspired by this report, three of the present authors gathered data on the prevalence, type, and treatment of missing data across 3 years of publications within a prominent psychological journal to assess both prevalence and treatment of missing data in the psychological literature. Estimates of amounts and types of missing data were recorded from over 300 articles across the 3-year period. The results were dramatic. Not only were missing data prevalent across studies (approximately 90% of the articles had missing data), but the average amount of missing data for studies in our sample well exceeded 30%. Additionally, few of the articles included explicit mention of missing data, and even fewer indicated that the authors attended to the missing data, either by performing statistical procedures or by making disclaimers regarding the studies in the results and conclusions. We suspected that there would be a fairly high rate of missing data, but we did not anticipate such lack of attention to missing data. The prevalence of missing data and the general lack of attention to it suggest a probable impact on the validity and interpretability of research results.

Our review of journal articles suggests that missing data are common and often not given adequate attention by social scientists; the problem is either ignored or finessed. That is, researchers are aware of the missing data and attend to it by rationalizing why it is irrelevant to the particular study. Sometimes this rationale involves a statistical sleight of hand in which statistics are used to convince us that missing data have no significant impact on study results.[1] Yet on closer inspection, the statistics are often used inappropriately and are therefore prone to misinterpretation.

Such a situation is not limited to social science or to science in general. Missing data are ubiquitous and often ignored or finessed in many disciplines, both within and outside science. For example, in softball, there exists a "mercy" rule to eliminate demoralization of the losing team. Teams that are behind the winning team by a specific number of runs at the conclusion of any inning are declared the loser. This rule has subtle consequences involving missing data. The losing team never gets the opportunity to attempt a comeback, which, although improbable, is possi-

ble. Thus, in softball, values for the missing data are inferred and few argue about those values or the consequences resulting from the conclusions drawn from incomplete data.

Other examples come from historical biographies and archaeology. In historical biographies, detailed information about a remarkable person's life quite often is undocumented and therefore must be extrapolated or imputed based on the norms of the time. For example, Thomas Jefferson's food preferences were not likely to have been recorded in any great detail, yet if that feature of his life were to become relevant in a biography or history, the author would likely infer that his preferences were similar to the norm of his day. Unless his habits were unusual and/or noteworthy, the contemporary normative behavior is likely to be imputed to replace the missing observations. Imputation of missing data can easily be justified. Richard Dawkins (1998) remarked that humans are designed to perceive stimuli that are novel in the environment. If Jefferson's dietary practices were notable in his day, his contemporaries would likely have noted these oddities. Otherwise, there is no compelling reason to think that the missing information would have been anything out of the ordinary.

Archaeologists and paleontologists almost always have more missing than complete data. They rely on logic, theory, and conjecture to piece together the bits of available data and produce a plausible and worthwhile story. Paleontologists piece together the skeletal remains of species long extinct. Rarely are there enough remains for the researcher to resurrect the entire bone structure, posture, or external features of the creature. However, museum curators impute that missing information and infer the physical structure when they build models of these creatures for public display. Sometimes these imputations are incorrect. A recent report noted that a major change in the inferred facial structure of a particular dinosaur was necessary to explain both new fossil discoveries as well as likely foraging habits of the animal. The change, as it turned out, was logical, but also revolutionary in the study of that type of dinosaur and of other dinosaurs from the same period.

These examples show the pervasiveness and ordinariness of missing data. There is no reason to consider missing data as a unique feature of social science. What makes missing data noteworthy is the influence, whether known or unknown, that it has on our conclusions and ultimately on our knowledge. Therefore, the fact that it is not unusual for social scientists to make little or no mention of the potential impact of missing data on research conclusions is worrisome. Scientists in fields where missing data are both common and obvious (e.g., paleontology) expect that future

research will uncover the errors made because of incorrect imputations or extrapolations. Social science researchers, however, do not always have the luxury of these future corrections.

WHY DATA MIGHT BE MISSING

There are a variety of reasons data would be missing. We classify those into three broad categories related to (1) the study participants, (2) the study design, and (3) the interaction of the participants and the study design. For example, data might be missing because some participants were offended by certain questions on a survey (participant characteristics), or because a study required too much of the participants' time (design characteristics), or because those who were the sickest were unable to complete the more burdensome aspects of the study (participant and design characteristics). As we discuss in subsequent chapters, the reasons why data are missing can have important consequences on the amount and pattern of missing data, the selection of appropriate missing data handling techniques, and the interpretation of research results.

The stage of the study in which missing data occurs is also informative. Data can be lost at the study recruitment stage, the implementation stage, or the follow-up stage. Data missing from the recruitment stage could be due to exclusionary criteria for the study, dropout prior to assignment to experimental conditions (e.g., treatment groups), or participants losing interest in the study prior to signing a consent form. Data missing during the implementation stage might be due to skipped items on questionnaires, to absence during a data collection period, or to refusal to participate after being recruited. Data missing at follow-up is a familiar situation for longitudinal researchers: data could be missing due to participants dropping out of the study or to the inability to contact participants for follow-up data.

Another important aspect of missing data is the different units of analysis and different levels of measurement within the study. For example, a distinction is made in the missing data literature between "unit missing data," which refers to data for an entire unit of analysis (e.g., study participant) that is missing, and "missing values," which refers to scores on a particular variable (e.g., questionnaire item) that are missing. Moreover, in longitudinal studies, there can be "missing wave" data, that is, data that are missing at a particular occasion of measurement. However, even this seemingly fine-grained distinction does not convey a sufficient level of specific-

ity. In multilevel studies (i.e., those in which participants are grouped into larger units), "unit missing data" can occur at the individual or participant level, at the group level (e.g., males or females), and/or at the organization or community level (e.g., clinics, hospitals, or schools). Similarly, "missing values" can occur for single items within a measure, for subscales, for entire test scores, or for multivariate latent variable scores.

Moreover, data can be missing cross-sectionally (for persons or variables observed at a single occasion) or across time in longitudinal studies (for persons, variables, and occasions of measurement). Noting all of these sources of missing data can assist researchers in determining the reasons for missing data and the amount and pattern of missing data. In turn, this information can help researchers as they develop methods of handling missing data and appropriately interpreting study results.

It is important to note that these different types of missing data exist for researchers collecting data directly from participants and for researchers collecting data from existing records such as police records or medical files. As with data collected directly from study participants, record data can be missing data for entire "cases" (e.g., individuals), for single items, for variables, for an occasion of measurement, and so on.

THE IMPACT OF MISSING DATA

The most pressing concern regarding missing data is the extent to which the missing information influences study results. For example, if the majority of study participants who fared poorly in an experimental intervention dropped out, the results would be based largely on the participants who responded positively. The missing information about the poorer outcomes would then lead to an overestimation of the benefits of the treatment. Yet because the data are missing, it is difficult to determine the impact of data that might have been present in the study. There are two aspects of missing data that can provide us with clues regarding the extent of the influence of the missing information on study results. First, the *amount* of missing data (see Chapter 5 for details) is related to its impact on research conclusions. In general, greater amounts of missing data are expected to have a large impact on study generalizability and statistical inference; however, as we will discuss later, these expectations are not always warranted. Under most conditions, data sets in which large amounts of data are missing result in smaller sample sizes and potentially

unrepresentative samples of the population to which we wish to generalize. Further, the available data for the remaining sample might reflect a bias, thus resulting in biased parameter estimates and misleading statistical conclusions. Second, the actual process that causes missing data can affect the validity of the inferences made from the analyses. Depending on the causal origin, missing data can have dramatic influences on the validity of study findings. For example, in a study assessing the effects of two mathematics curricula on test performance, the students' current math abilities could be related to missing data; that is, if those likely to fail the test chose not to take it, then inferences based on study results about the effectiveness of the curricula would likely be misleading. Moreover, generalizability would be compromised because results would not include the poorer math students.

WHAT IS MISSING IN THE LITERATURE ON MISSING DATA?

The subject of missing data has been widely addressed in the statistical literature as well as in other relevant bodies of literature. Statisticians have attempted to assess and reconcile the problems associated with missing data theoretically and with practical solutions. In general, the statistical literature reflects an appreciation for the type and magnitude of problems associated with missing data, particularly with respect to how missing data affect statistical results. Much of the literature focuses on how to identify missing data and correct potential biases attributable to missing data. The collective effort has produced numerous working solutions to many missing data problems.

Unfortunately, the statistical literature appears to have had negligible impact on the research practices of social scientists when it comes to handling missing data. We offer several possible reasons for this lack of impact. One quite plausible reason is the fact that many social scientists lack the level of training and expertise in statistics or mathematics required to understand this highly technical literature, which includes proofs, theorems, and definitions expressed in mathematical notation. Another reason is the paucity of user-friendly tools available for handling missing data. Although most statistical programs now offer missing data handling procedures, these procedures are often difficult to use for a novice data analyst. Missing data "add-on" programs for statistical software

packages often suffer from poor documentation or limited functionality. Stand-alone programs for handling missing data also exist (e.g., NORM; Schafer, 1997), but again, these programs are often complex and difficult to use. The tools are thus generally not easily adopted by most researchers.

A third reason that the statistical literature on missing data has had little impact on the practices of social scientists is that social scientists have not had much of a mandate to attend to the literature. Results from our previously mentioned analysis of missing data in journal articles suggest that such articles can be and are being published without their authors paying attention to the problem of missing data. This situation suggests that either reviewers and/or editors are not requiring investigators to address the issue of missing data. Our own direct experience with journal reviewers has led us to conclude that many reviewers who do comment on missing data tend to be misinformed about missing data issues, particularly with regard to statistical techniques for handling missing data. Our observations are consistent with those of other social scientists who are well informed about both the statistical literature and social science research in general. David Krantz, a renowned statistician and social scientist, remarked a while ago that most social scientists who consider themselves statistically sophisticated often are not (Krantz, 1999).

Prevention and Remediation

The contemporary and classical view of missing data is largely ensconced in statistics and primarily focused on prescribing remedies. While we subscribe to much of this work, we feel that the matter of missing data is far more complex than a remedial statistical conundrum. Missing data become a statistical consideration only after data have been collected and are ready to be analyzed. Prior to data collection, there are many available strategies that decrease the likelihood of missing data. For example, decreasing respondent burden or increasing benefits to participants tends to decrease the incidence of missing data. We believe that the problem of missing data is a larger issue that involves not only statistics but also logic and research methodology. Therefore, in this book we not only discuss statistical solutions for handling missing data, but we also address research design and measurement strategies for preventing the occurrence of missing data. Our goals are to combine the major findings of a technically daunting statistical literature with those of the research methodology literature, and to redirect the focus from solely remedial solutions for handling

missing data to preventive ones as well. We thus provide a comprehensive view of solutions for handling missing data that cover the planning, implementation, and analysis phases of research.

Of course, we recognize that prevention of missing data is not always possible—for example, in the context of secondary data analysis, where the data have been collected by someone else and the goal is to analyze those data. Figueredo, Sales, Russell, Becker, and Kaplan (2000) conducted a study using data from male adolescent sex offenders (ages 13 to 18) who had been referred to an outpatient evaluation and treatment clinic for sexual offenders in New York City, for assessment and treatment, from 1985 to 1990. As a result of inconsistent data collection over the years by the clinicians, a large number of offenders had missing data. In such cases, only remediation of missing data remains an option, and it is necessary to carefully consider one's options for doing so.

A Structural versus Functional Approach

Much of the existing literature on missing data reflects an approach that is structural rather than functional. In other words, the traditional classification schemes for missing data emphasize what missing data are rather than what missing data do and, in turn, what one might do about them. The traditional terminology hails from the statistical literature and focuses on abstract mechanisms of missing data, for example, missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). As we discuss in detail in Chapter 3, these terms are defined by theoretical relationships between missing values and observed or unobserved variables, reflecting a focus on the structure of the missing data.

A functional approach would instead focus on how missing data functions with respect to research results and conclusions. For example, what are the implications for possible sampling bias of data being MCAR? Is the sample of nonmissing data that is available for analysis still representative of the original population? What are the implications of the data being MAR for the generalizability of our results? These are the types of questions that would be addressed by a functional approach as opposed to just describing the mathematical properties of the missing data. Our discussion will address both structural and functional approaches, with an emphasis on the functional approach (see Table 3.2 in Chapter 3 for an illustration of this integration).

A COST–BENEFIT APPROACH TO MISSING DATA

In emphasizing a functional approach, however, we do not provide a "cookbook" for handling the problem of missing data. Instead, throughout this book we advocate sensible and thoughtful attention to missing data. For two reasons we do not promote a "one-size-fits-all" method to solve the problem of missing data. First, there is no single method for handling missing data, and second, every researcher has his or her own priorities. As in any research study, there is no perfect design or method for handling missing data. The research goals and priorities ought to drive the selection of methods, not the reverse. In short, we discourage the tendency to allow tools or methods to drive decisions and to allow convention to dictate how we handle the problem of missing data. Although reliance on convention might be the most expedient means by which to handle missing data, it is often not the optimal choice.

What constitutes a thoughtful approach regarding the handling of missing data falls within the familiar framework of a cost–benefit analysis. Every decision has associated costs and benefits. We use this guiding principle to convey why no single solution is appropriate for handling missing data. As an illustration, consider a researcher who decides to gather a convenience sample by collecting data from participants who are easily recruited but not necessarily best suited for the research study. The researcher benefits by saving time and energy as well as possibly increasing the sample size. The costs of this sampling method may include one of many threats to validity, such as selection bias, maturation, and so on, which in turn call into question the study results. Thus, each researcher should consider the costs and benefits of study-related decisions in order to make more informed and potentially better decisions than would be made by conducting research in a standard manner with the same design methods and analysis. With respect to missing data, costs may come in the form of increased demands on resources such as time and money, threats to other aspects of the study such as statistical power, or opportunity costs (i.e., requiring resources that may be used more productively elsewhere). These costs may be associated with actions taken to prevent, treat, or diagnose missing data. Likewise, benefits may be expressed in the same forms and associations.

How Costs and Benefits Are Valued

Unfortunately, costs and benefits are often unknown to researchers. Common practice and research history or lore dictate decisions more often than

researchers realize. Research fields often operate with an implicit valuing of inferential errors. For example, most formally trained biostatisticians are loath to make Type I errors, and thus guard against being too liberal in statistical testing by invoking the most stringent test. Conversely, clinical researchers are trained to guard against Type II errors by avoiding overly stringent statistical tests when evaluating treatments. Thus, the biostatistician would opt for more conservative methods while the clinical researcher opts for more liberal methods. These perspectives often lead to substantially different decisions and practices, though both groups are trained to recognize and deal with reliability and validity issues. We as researchers are often unaware of these implicit values, because we tend to follow convention in our field, doing what we were trained to do and being surrounded by people who do the same. Thus, biostatisticians are surrounded by others who maintain the same inferential risk values, as are clinical researchers. Neither perspective is necessarily wrong, but they do produce different decisions. One primary goal of this book is to make researchers aware of the implicit values behind the selection of various missing data strategies in order to facilitate informed decision making with respect to missing data. Treating missing data can be characterized by a series of decisions that are closely tied to both implicit and explicit values.

With missing data, a researcher ought to change the level of inferential risk based on these values. For example, if the main priority of a study is generalizability, it ought to have a different approach to missing data than a study whose main priority is high internal validity. In the first study, missing data that limit sample variability are detrimental, whereas in the second study, limiting sample variability would be optimal (not to suggest you should limit variability by tolerating missing data!). If researchers adhere to the process of explicitly stating the purpose or goals of a study, no single procedure will be a panacea for missing data. Rule-bound or practice-bound researchers will find little comfort in the pages ahead, since we stress that missing data situations are often unique. It behooves the researcher to carefully consider the costs and benefits of the missing data decisions.

The Facets of Costs and Benefits

The two facets of research that concern scientists are reliability and validity.[2] Reliability addresses the replicability or consistency of the observed findings and conclusions based on the data. If study findings fail to replicate, then the phenomena of interest are not well addressed by the meth-

ods, theory, and/or data used in the study. Similarly, if the findings are not consistent across all related outcomes, then the results may lack a suitable level of reliability. Validity—internal, external, and construct—concerns the issue of sound causal inference. Internal validity relates to the extent to which we can infer that the observed outcome is related to the independent or manipulated variable of interest. Internal validity is decreased when plausible rival hypotheses exist that the researcher has either failed to control for or failed to anticipate. External validity relates to the applicability of the findings to other observations or samples, settings, and constructs. The term *causal generalization* (e.g., Cook, 2004) refers to the broad notion of causal stability and external validity. Construct validity relates to the appropriateness of the measurements to yield accurate and worthy indicators of the constructs of interest. For example, whether an IQ test is a valid measure of intelligence is a question of construct validity. If measures are poor indicators of the theoretical constructs of interest, then information provided by statistical procedures may provide a poor test of the theory.

Regardless of the research content, researchers strive to maximize all three of these facets. Therefore, our discussion of costs and benefits with respect to addressing missing data will be constrained to these facets. There are other more worldly costs to these decisions, such as economic costs. If researchers considered the cost per unit analyzed, most investigators might be more motivated to carefully scrutinize lost data. Costs, however, we leave to economists and to the reader to address independently, since they are likely to be unique as well as quite variable between studies.

Missing data constitute threats to different forms of reliability, validity, and generalizability of study results. As detailed elsewhere in this volume, the application or nonapplication of different solutions to those problems can impact these threats directly. Therefore, handling of missing data is directly relevant to competing reliability, validity, and generalizability concerns. Because these concerns need to be considered relative to each other's competing demands, the judicious selection of treatment for missing data becomes likewise relative to this optimal tradeoff.

The Relationship between Costs and Benefits

Decisions that enhance one facet (e.g., external validity) may do so at the expense of another facet (e.g., internal validity). Simply put, costs and

benefits may not be completely independent and may, in fact, be negatively correlated. Social science researchers have argued among themselves about the tradeoffs between experimental control and ecological validity. The argument provides insight into the mutual exclusivity of internal validity and generalizability. Researchers who value internal validity may operate at the expense of generalizability or, more importantly, at the expense of mundane realism.[3] Therefore, the valuing of each facet forms the basis for decision making when preventing, diagnosing, and treating missing data. Without knowledge of study priorities regarding these facets, the researcher is left to accept prescriptions from others who may not share the same priorities. For example, some critics appear to consider it akin to "cheating" to estimate missing data. We acknowledge that there is indeed an inferential risk in any method of estimating missing data. Nevertheless, this risk has to be considered in relation to the inferential risk of *not* estimating missing data and leaving unaddressed the possible errors in hypothesis testing and the invalid generalizations that may ensue from not imputing missing data. In these cases, it might emphatically *not* be the case that a bird in the hand (observed data) is worth two in the bush (imputed data). The available data, although seemingly more solid, might be less representative, generalizable, and valid than a reconstructed sample including some merely "estimated" data.

MISSING DATA—NOT JUST FOR STATISTICIANS ANYMORE

In addition to taking a cost–benefit approach, we view missing data from a comprehensive perspective. The majority of the extant missing data literature focuses almost exclusively on treatment. As a result, almost all discussions of missing data begin and end with statistical procedures. The following chapters detail our more comprehensive perspective, beginning with the formulation of the study and ending with the final step of disseminating the results. Each step along the way involves decisions that impact missing data. Carefully considering the potential influences of missing data in each step provides researchers with more information on which to base decisions. In addition, understanding the values placed on each facet (i.e., reliability, validity, and generalizability) will make the decision process easier to communicate to others. We hope that after reading this book readers will come see the presence of missing data as something that can

be described, prevented, treated, and discussed without relying solely on statistical methods.

A Functional Emphasis

The functional approach taken in this book reflects a fourfold approach to missing data:

1. Explore the possibilities of where missing data originate so that their presence might be minimized by thoughtful prevention strategies;
2. Derive the implications missing data might have for the outcome of any analysis in terms of the reliability, validity, and generalizability of study conclusions;
3. Consider the various statistical options for dealing with missing data, including methods of data deletion, imputation, and model estimation;
4. Investigate the probable consequences of the options considered in the previous step on the study's results and conclusions.

We believe that this functional approach to handling missing data has the additional advantage of further enhancing the comprehensibility of scientific contributions. It is likely that many social scientists are not only confused by the technical vocabulary surrounding statistical descriptions of missing data but also befuddled by the lack of a pragmatic frame of reference for this technical information. By framing the problem in a functional rather than purely structural context, we hope to make the "bottom-line" relevance of missing data issues clearer to social science researchers. Our intention is to render the information understandable and thus highly usable.

PURPOSE OF THIS BOOK

The primary goal of this book is to eliminate at least some of the plausible reasons why social scientists fail to attend to missing data. We present the technical information on missing data accurately but with minimal reliance on statistical or mathematical notation. When information is presented in formalized mathematical ways, we provide explanatory information so that the reader can interpret the notation without need-

ing to know advanced mathematics. We also provide practical examples that convey many of these abstract descriptions more concretely. In addition to making the information more available to a non-technically inclined audience, we present practical solutions for handling missing data. We include all information that is relevant to implementing these solutions, regardless of the research design or statistical package used. Our goal is to raise researchers' awareness of the importance of missing data by eliminating at least some of the obstacles to their understanding of this typically technical information and to their using all the available tools for handling missing data to ameliorate the problems posed by it.

Another purpose of this book is to dispel the implicit myth that missing data is just a statistical matter. We will discuss missing data as a threat to the internal validity of causal inference and to the external validity or generalizability of research results (see Chapter 2). We hope that this book will serve as an introduction to the conceptual and methodological issues involved with handling missing data as well as a reference for students, researchers, and reviewers.

Chapter 2 addresses the consequences of missing data. We describe the many possible consequences of missing data to emphasize that missing data is an important issue in the social sciences. In Chapter 3, we turn our attention to the classification of missing data. Since statisticians and other formally trained or knowledgeable data analysts discuss missing data using specific terminology, it is important that the reader become familiar with that terminology. The information is presented in a manner that requires little if any knowledge of statistics or statistical formulas. At the conclusion of our discussion of these topics, we direct the reader to our approach to missing data. In Chapter 4, we discuss how to prevent missing data by designing studies to decrease the likelihood of missing data. We also discuss the elimination of errors in data handling that often result in missing data. Chapter 5 presents diagnostic procedures that will enable the data analyst and researcher to better appreciate the extent, pattern, and nature of the missing data. Since a great amount of thought must go into choosing the methods for treating missing data, we cover the decision making process in great detail in the following chapter (Chapter 6). The next group of chapters (Chapters 7–10) discuss statistical methods for handling missing data. Broadly, these chapters address deletion procedures, augmentation methods, single imputation, and multiple imputation procedures. In each chapter, the methods are described and examples are given. Finally, we conclude the book in Chapter 11 with recommendations for how to report missing data.

NOTES

1. We discuss these statistical "sleights of hand" in subsequent chapters that address diagnosing (Chapter 5) and handling missing data (Chapter 6).

2. Information about these concepts is covered in greater detail by Cook and Campbell (1979) and Shadish, Cook, and Campbell (2002). The information provided in this text is cursory and ought to serve only as a review.

3. "Mundane realism" is a term used by Aronson and Carlsmith (1968) in a chapter on experimental research methods in social psychology. The term refers to the correspondence of the research situation to the situation most likely to be observed outside the research setting.

Guilford Publications
72 Spring Street
New York, NY 10012
212-431-9800
800-365-7006
www.guilford.com