Series Editor's Note

Missing data are a real bane to researchers across all social science disciplines. For most of our scientific history, we have approached missing data much like a doctor from the ancient world might use bloodletting to cure disease or amputation to stem infection (e.g, removing the infected parts of one's data by using list-wise or pair-wise deletion). My metaphor should make you feel a bit squeamish, just as you should feel if you deal with missing data using the antediluvian and ill-advised approaches of old. Fortunately, Craig Enders is a gifted quantitative specialist who can clearly explain missing data procedures to diverse readers from beginners to seasoned veterans. He brings us into the age of modern missing data treatments by demystifying the arcane discussions of missing data mechanisms and their labels (e.g., MNAR) and the esoteric acronyms of the various techniques used to address them (e.g., FIML, MCMC, and the like).

Enders's approachable treatise provides a comprehensive treatment of the causes of missing data and how best to address them. He clarifies the principles by which various mechanisms of missing data can be recovered, and he provides expert guidance on which method to implement and how to execute it, and what to report about the modern approach you have chosen. Enders's deft balancing of practical guidance with expert insight is refreshing and enlightening. It is rare to find a book on quantitative methods that you can read for its stated purpose (educating the reader about modern missing data procedures) and find that it treats you to a level of insight on a topic that whole books dedicated to the topic cannot match. For example, Enders's discussions of maximum likelihood and Bayesian estimation procedures are the clearest, most understandable, and instructive discussions I have read—your inner geek will be delighted, really.

Enders successfully translates the state-of-the art technical missing data literature into an accessible reference that you can readily rely on and use. Among the treasures of Enders's work are the pointed simulations that he has developed to show you exactly what the technical literature obtusely presents. Because he provides such careful guidance of the foundations and the step-by-step processes involved, you will quickly master the concepts and issues of this critical literature. Another treasure is his use of a common running example that he builds upon as more complex issues are presented. And if these features are not enough, you can also visit the accompanying website (*www.appliedmissingdata.com*), where you will find up-to-date program files for the examples presented, as well as additional examples of the different software programs available for handling missing data.

What you will learn from this book is that missing data imputation is not cheating. In fact, you will learn why the egregious scientific error would be the business-as-usual approaches that still permeate our journals. You will learn that modern missing data procedures are so effective that intentionally missing data designs often can provide more valid and generalizable results than traditional data collection protocols. In addition, you will learn to rethink how you collect data to maximize your ability to recover any missing data mechanisms and that many quandaries of design and analysis become resolvable when recast as a missing data problem. Bottom line—after you read this book you will have learned how to go forth and impute with impunity!

TODD D. LITTLE
University of Kansas
Lawrence, Kansas

1

An Introduction to Missing Data

1.1 INTRODUCTION

Missing data are ubiquitous throughout the social, behavioral, and medical sciences. For decades, researchers have relied on a variety of ad hoc techniques that attempt to “fix” the data by discarding incomplete cases or by filling in the missing values. Unfortunately, most of these techniques require a relatively strict assumption about the cause of missing data and are prone to substantial bias. These methods have increasingly fallen out of favor in the methodological literature (Little & Rubin, 2002; Wilkinson & Task Force on Statistical Inference, 1999), but they continue to enjoy widespread use in published research articles (Bodner, 2006; Peugh & Enders, 2004).

Methodologists have been studying missing data problems for nearly a century, but the major breakthroughs came in the 1970s with the advent of maximum likelihood estimation routines and multiple imputation (Beale & Little, 1975; Dempster, Laird, & Rubin, 1977; Rubin, 1978b; Rubin, 1987). At about the same time, Rubin (1976) outlined a theoretical framework for missing data problems that remains in widespread use today. Maximum likelihood and multiple imputation have received considerable attention in the methodological literature during the past 30 years, and researchers generally regard these approaches as the current “state of the art” (Schafer & Graham, 2002). Relative to traditional approaches, maximum likelihood and multiple imputation are theoretically appealing because they require weaker assumptions about the cause of missing data. From a practical standpoint, this means that these techniques will produce parameter estimates with less bias and greater power.

Researchers have been relatively slow to adopt maximum likelihood and multiple imputation and still rely heavily on traditional missing data handling techniques (Bodner, 2006; Peugh & Enders, 2004). In part, this hesitancy may be due to a lack of software options, as maximum likelihood and multiple imputation did not become widely available in statistical packages until the late 1990s. However, the technical nature of the missing data literature probably represents another significant barrier to the widespread adoption of these techniques. Consequently, the primary goal of this book is to provide an accessible and user-friendly introduction to missing data analyses, with a special emphasis on maximum likelihood and multiple imputation. It is my hope that this book will help address the gap that currently exists between the analytic approaches that methodologists recommend and those that appear in published research articles.

1.2 CHAPTER OVERVIEW

This chapter describes some of the fundamental concepts that appear repeatedly throughout the book. In particular, the first half of the chapter is devoted to missing data theory, as described by Rubin (1976) and colleagues (Little & Rubin, 2002). Rubin is responsible for establishing a nearly universal classification system for missing data problems. These so-called missing data mechanisms describe relationships between measured variables and the probability of missing data and essentially function as assumptions for missing data analyses. Rubin's mechanisms serve as a vital foundation for the remainder of the book because they provide a basis for understanding why different missing data techniques succeed or fail.

The second half of this chapter introduces the idea of planned missing data. Researchers tend to believe that missing data are a nuisance to be avoided whenever possible. It is true that unplanned missing data are potentially damaging to the validity of a statistical analysis. However, Rubin's (1976) theory describes situations where missing data are relatively benign. Researchers have exploited this fact and have developed research designs that produce missing data as an intentional by-product of data collection. The idea of intentional missing data might seem odd at first, but these research designs actually solve a number of practical problems (e.g., reducing respondent burden and reducing the cost of data collection). When used in conjunction with maximum likelihood and multiple imputation, these planned missing data designs provide a powerful tool for streamlining and reducing the cost of data collection.

I use the small data set in Table 1.1 to illustrate ideas throughout this chapter. I designed these data to mimic an employee selection scenario in which prospective employees complete an IQ test and a psychological well-being questionnaire during their interview. The company subsequently hires the applicants who score in the upper half of the IQ distribution, and a supervisor rates their job performance following a 6-month probationary period. Note that the job performance scores are systematically missing as a function of IQ scores (i.e., individuals in the lower half of the IQ distribution were never hired, and thus have no performance rating). In addition, I randomly deleted three of the well-being scores in order to mimic a situation where the applicant's well-being questionnaire is inadvertently lost.

1.3 MISSING DATA PATTERNS

As a starting point, it is useful to distinguish between missing data patterns and missing data mechanisms. These terms actually have very different meanings, but researchers sometimes use them interchangeably. A **missing data pattern** refers to the configuration of observed and missing values in a data set, whereas **missing data mechanisms** describe possible relationships between measured variables and the probability of missing data. Note that a missing

TABLE 1.1. Employee Selection Data Set

IQ	Psychological well-being	Job performance
78	13	—
84	9	—
84	10	—
85	10	—
87	—	—
91	3	—
92	12	—
94	3	—
94	13	—
96	—	—
99	6	7
105	12	10
105	14	11
106	10	15
108	—	10
112	10	10
113	14	12
115	14	14
118	12	16
134	11	12

data pattern simply describes the location of the "holes" in the data and does not explain why the data are missing. Although the missing data mechanisms do not offer a causal explanation for the missing data, they do represent generic mathematical relationships between the data and missingness (e.g., in a survey design, there may be a systematic relationship between education level and the propensity for missing data). Missing data mechanisms play a vital role in Rubin's missing data theory.

Figure 1.1 shows six prototypical missing data patterns that you may encounter in the missing data literature, with the shaded areas representing the location of the missing values in the data set. The **univariate pattern** in panel A has missing values isolated to a single variable. A univariate pattern is relatively rare in some disciplines but can arise in experimental studies. For example, suppose that Y_1 through Y_3 are manipulated variables (e.g., between-subjects factors in an ANOVA design) and Y_4 is the incomplete outcome variable. The univariate pattern is one of the earliest missing data problems to receive attention in the statistics literature, and a number of classic articles are devoted to this topic.

Panel B shows a configuration of missing values known as a **unit nonresponse pattern**. This pattern often occurs in survey research, where Y_1 and Y_2 are characteristics that are available for every member of the sampling frame (e.g., census tract data), and Y_3 and Y_4 are surveys that some respondents refuse to answer. Later in the book I describe a planned missing data design that yields a similar pattern of missing data. In the context of planned missingness, this pattern can arise when a researcher administers two inexpensive measures to the entire sample (e.g., Y_1 and Y_2) and collects two expensive measures (e.g., Y_3 and Y_4) from a subset of cases.

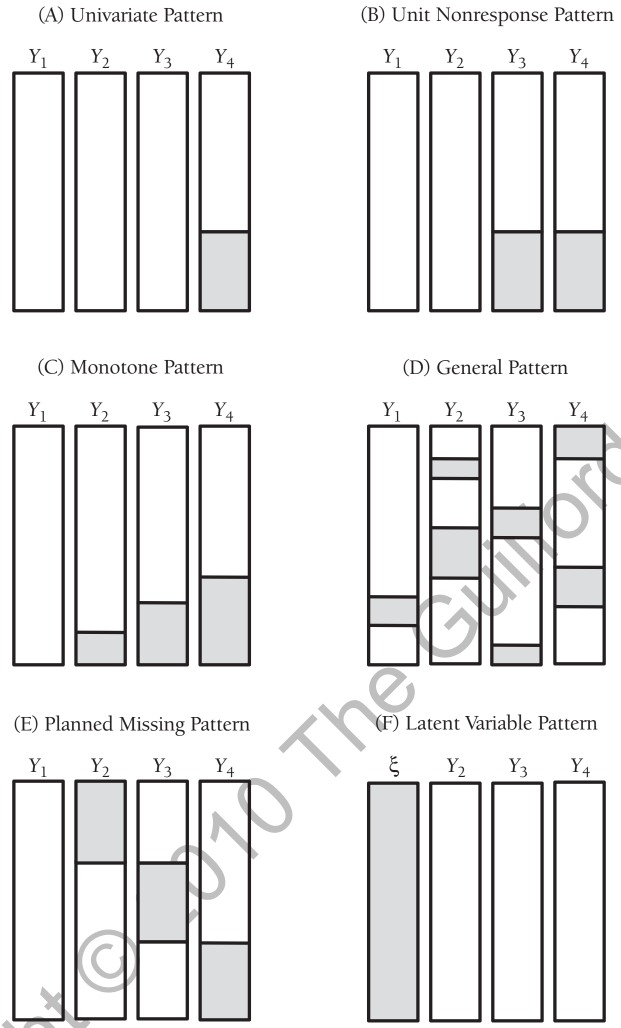

FIGURE 1.1. Six prototypical missing data patterns. The shaded areas represent the location of the missing values in the data set with four variables.

A **monotone missing data pattern** in panel C is typically associated with a longitudinal study where participants drop out and never return (the literature sometimes refers to this as **attrition**). For example, consider a clinical trial for a new medication in which participants quit the study because they are having adverse reactions to the drug. Visually, the monotone pattern resembles a staircase, such that the cases with missing data on a particular assessment are always missing subsequent measurements. Monotone missing data patterns have received attention in the missing data literature because they greatly reduce the mathematical complexity of maximum likelihood and multiple imputation and can eliminate the need for iterative estimation algorithms (Schafer, 1997, pp. 218–238).

A **general missing data pattern** is perhaps the most common configuration of missing values. As seen in panel D, a general pattern has missing values dispersed throughout the data matrix in a haphazard fashion. The seemingly random pattern is deceptive because the values can still be systematically missing (e.g., there may be a relationship between Y_1 values and the propensity for missing data on Y_2). Again, it is important to remember that the missing data pattern describes the location of the missing values and not the reasons for missingness. The data set in Table 1.1 is another example of a general missing data pattern, and you can further separate this general pattern into four unique missing data patterns: cases with only IQ scores ($n = 2$), cases with IQ and well-being scores ($n = 8$), cases with IQ and job performance scores ($n = 1$), and cases with complete data on all three variables ($n = 9$).

Later in the chapter, I outline a number of designs that produce intentional missing data. The **planned missing data pattern** in panel E corresponds to the three-form questionnaire design outlined by Graham, Hofer, and MacKinnon (1996). The basic idea behind the three-form design is to distribute questionnaires across different forms and administer a subset of the forms to each respondent. For example, the design in panel E distributes the four questionnaires across three forms, such that each form includes Y_1 but is missing Y_2, Y_3, or Y_4. Planned missing data patterns are useful for collecting a large number of questionnaire items while simultaneously reducing respondent burden.

Finally, the **latent variable pattern** in panel F is unique to latent variable analyses such as structural equation models. This pattern is interesting because the values of the latent variables are missing for the entire sample. For example, a confirmatory factor analysis model uses a latent factor to explain the associations among a set of manifest indicator variables (e.g., Y_1 through Y_3), but the factor scores themselves are completely missing. Although it is not necessary to view latent variable models as missing data problems, researchers have adapted missing data algorithms to estimate these models (e.g., multilevel models; Raudenbush & Bryk, 2002, pp. 440–444).

Historically, researchers have developed analytic techniques that address a particular missing data pattern. For example, Little and Rubin (2002) devote an entire chapter to older methods that were developed specifically for experimental studies with a univariate missing data pattern. Similarly, survey researchers have developed so-called hot-deck approaches to deal with unit nonresponse (Scheuren, 2005). From a practical standpoint, distinguishing among missing data patterns is no longer that important because maximum likelihood estimation and multiple imputation are well suited for virtually any missing data pattern. This book focuses primarily on techniques that are applicable to general missing data patterns because these methods also work well with less complicated patterns.

1.4 A CONCEPTUAL OVERVIEW OF MISSING DATA THEORY

Rubin (1976) and colleagues introduced a classification system for missing data problems that is widely used in the literature today. This work has generated three so-called missing data mechanisms that describe how the probability of a missing value relates to the data, if at all. Unfortunately, Rubin's now-standard terminology is somewhat confusing, and researchers often misuse his vernacular. This section gives a conceptual overview of missing data theory that uses hypothetical research examples to illustrate Rubin's missing data mechanisms. In the next section, I delve into more detail and provide a more precise mathematical definition of the missing data mechanisms. Methodologists have proposed additions to

Rubin's classification scheme (e.g., Diggle & Kenward, 1994; Little, 1995), but I focus strictly on the three missing data mechanisms that are common in the literature. As an aside, I try to use a minimal number of acronyms throughout the book, but I nearly always refer to the missing data mechanisms by their abbreviation (MAR, MCAR, MNAR). You will encounter these acronyms repeatedly throughout the book, so it is worth committing them to memory.

Missing at Random Data

Data are **missing at random** (MAR) when the probability of missing data on a variable *Y* is related to some other measured variable (or variables) in the analysis model but not to the values of *Y* itself. Said differently, there is no relationship between the propensity for missing data on *Y* and the values of *Y* after partialling out other variables. The term *missing at random* is somewhat misleading because it implies that the data are missing in a haphazard fashion that resembles a coin toss. However, MAR actually means that a systematic relationship exists between one or more measured variables and the probability of missing data. To illustrate, consider the small data set in Table 1.2. I designed these data to mimic an employee selection scenario in which prospective employees complete an IQ test during their job interview and a supervisor subsequently evaluates their job performance following a 6-month probationary period. Suppose that the company used IQ scores as a selection measure and did not hire applicants that scored in the lower quartile of the IQ distribution. You can see that the job performance ratings in the MAR column of Table 1.2 are missing for the applicants with the lowest IQ scores. Consequently, the probability of a missing job performance rating is solely a function of IQ scores and is unrelated to an individual's job performance.

There are many real-life situations in which a selection measure such as IQ determines whether data are missing, but it is easy to generate additional examples where the propensity for missing data is less deterministic. For example, suppose that an educational researcher is studying reading achievement and finds that Hispanic students have a higher rate of missing data than Caucasian students. As a second example, suppose that a psychologist is studying quality of life in a group of cancer patients and finds that elderly patients and patients with less education have a higher propensity to refuse the quality of life questionnaire. These examples qualify as MAR as long as there is no residual relationship between the propensity for missing data and the incomplete outcome variable (e.g., after partialling out age and education, the probability of missingness is unrelated to quality of life).

The practical problem with the MAR mechanism is that there is no way to confirm that the probability of missing data on *Y* is solely a function of other measured variables. Returning to the education example, suppose that Hispanic children with poor reading skills have higher rates of missingness on the reading achievement test. This situation is inconsistent with an MAR mechanism because there is a relationship between reading achievement and missingness, even after controlling for ethnicity. However, the researcher would have no way of verifying the presence or absence of this relationship without knowing the values of the missing achievement scores. Consequently, there is no way to test the MAR mechanism or to verify that scores are MAR. This represents an important practical problem for missing data analyses because maximum likelihood estimation and multiple imputation (the two techniques that methodologists currently recommend) assume an MAR mechanism.

TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings			
IQ	Complete	MCAR	MAR	MNAR
78	9	—	—	9
84	13	13	—	13
84	10	—	—	10
85	8	8	—	—
87	7	7	—	—
91	7	7	7	—
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	—	7	—
99	7	7	7	—
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	—	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	—	12	12

Missing Completely at Random Data

The **missing completely at random** (MCAR) mechanism is what researchers think of as purely haphazard missingness. The formal definition of MCAR requires that the probability of missing data on a variable *Y* is unrelated to other measured variables and is unrelated to the values of *Y* itself. Put differently, the observed data points are a simple random sample of the scores you would have analyzed had the data been complete. Notice that MCAR is a more restrictive condition than MAR because it assumes that missingness is completely unrelated to the data.

With regard to the job performance data in Table 1.2, I created the MCAR column by deleting scores based on the value of a random number. The random numbers were uncorrelated with IQ and job performance, so missingness is unrelated to the data. You can see that the missing values are not isolated to a particular location in the IQ and job performance distributions; thus the 15 complete cases are relatively representative of the entire applicant pool. It is easy to think of real-world situations where job performance ratings could be missing in a haphazard fashion. For example, an employee might take maternity leave prior to her 6-month evaluation, the supervisor responsible for assigning the rating could be promoted to another division within the company, or an employee might quit because his spouse accepted a job in another state. Returning to the previous education example, note that children could have MCAR achievement scores because of unexpected personal events (e.g., an illness, a funeral, family vacation, relocation to another school district), scheduling difficulties (e.g., the class was away at a field trip when the researchers visited the school), or administrative snafus (e.g., the researchers inadvertently misplaced the tests before the data could be entered). Similar types of issues could produce MCAR data in the quality of life study.

In principle, it is possible to verify that a set of scores are MCAR. I outline two MCAR tests in detail later in the chapter, but the basic logic behind these tests will be introduced here. For example, reconsider the data in Table 1.2. The definition of MCAR requires that the observed data are a simple random sample of the hypothetically complete data set. This implies that the cases with observed job performance ratings should be no different from the cases that are missing their performance evaluations, on average. To test this idea, you can separate the missing and complete cases and examine group mean differences on the IQ variable. If the missing data patterns are randomly equivalent (i.e., the data are MCAR), then the IQ means should be the same, within sampling error. To illustrate, I classified the scores in the MCAR column as observed or missing and compared the IQ means for the two groups. The complete cases have an IQ mean of 99.73, and the missing cases have a mean of 100.80. This rather small mean difference suggests that the two groups are randomly equivalent, and it provides evidence that the job performance scores are MCAR. As a contrast, I used the performance ratings in the MAR column to form missing data groups. The complete cases now have an IQ mean of 105.47, and the missing cases have a mean of 83.60. This large disparity suggests that the two groups are systematically different on the IQ variable, so there is evidence against the MCAR mechanism. Comparing the missing and complete cases is a strategy that is common to the MCAR tests that I describe later in the chapter.

Missing Not at Random Data

Finally, data are **missing not at random** (MNAR) when the probability of missing data on a variable *Y* is related to the values of *Y* itself, even after controlling for other variables. To illustrate, reconsider the job performance data in Table 1.2. Suppose that the company hired all 20 applicants and subsequently terminated a number of individuals for poor performance prior to their 6-month evaluation. You can see that the job performance ratings in the MNAR column are missing for the applicants with the lowest job performance ratings. Consequently, the probability of a missing job performance rating is dependent on one's job performance, even after controlling for IQ.

It is relatively easy to generate additional examples where MNAR data could occur. Returning to the previous education example, suppose that students with poor reading skills have missing test scores because they experienced reading comprehension difficulties during the exam. Similarly, suppose that a number of patients in the cancer trial become so ill (e.g., their quality of life becomes so poor) that they can no longer participate in the study. In both examples, the data are MNAR because the probability of a missing value depends on the variable that is missing. Like the MAR mechanism, there is no way to verify that scores are MNAR without knowing the values of the missing variables.

1.5 A MORE FORMAL DESCRIPTION OF MISSING DATA THEORY

The previous section is conceptual in nature and omits the mathematical details behind Rubin's missing data theory. This section expands the previous ideas and gives a more precise description of the missing data mechanisms. As an aside, the notation and the terminology that I use in this section are somewhat different from Rubin's original work, but they are consistent with the contemporary missing data literature (Little & Rubin, 2002; Schafer, 1997; Schafer & Graham, 2002).

Preliminary Notation

Understanding Rubin's (1976) missing data theory requires some basic notation and terminology. The **complete data** consist of the scores that you would have obtained had there been no missing values. The complete data is partially a hypothetical entity because some of its values are missing. However, in principle, each case has a score on every variable. This idea is intuitive in some situations (e.g., a student's reading comprehension score is missing because she was unexpectedly absent from school) but is somewhat unnatural in others (e.g., a cancer patient's quality of life score is missing because he died). Nevertheless, you have to assume that a complete set of scores does exist, at least hypothetically. I denote the complete data as Y_{com} throughout the rest of this section.

In practice, some portion of the hypothetically complete data set is often missing. Consequently, you can think of the complete data as consisting of two components, the **observed data** and the **missing data** (Y_{obs} and Y_{mis}, respectively). As the names imply, Y_{obs} contains the observed scores, and Y_{mis} contains the hypothetical scores that are missing. To illustrate, reconsider the data set in Table 1.2. Suppose that the company used IQ scores as a selection measure and did not hire applicants that scored in the lower quartile of the IQ distribution. The first two columns of the table contain the hypothetically complete data (i.e., Y_{com}), and the MAR column shows the job performance scores that the human resources office actually collected. For a given individual with incomplete data, Y_{obs} corresponds to the IQ variable and Y_{mis} is the hypothetical job performance rating. As you will see in the next section, partitioning the hypothetically complete data set into its observed and missing components plays an integral role in missing data theory.

The Distribution of Missing Data

The key idea behind Rubin's (1976) theory is that missingness is a variable that has a probability distribution. Specifically, Rubin defines a binary variable R that denotes whether a score on a particular variable is observed or missing (i.e., $r = 1$ if a score is observed, and $r = 0$ if a value is missing). For example, Table 1.3 shows the MAR job performance ratings and the corresponding missing data indicator. A single indicator can summarize the distribution of missing data in this example because the IQ variable is complete. However, multivariate data sets tend to have a number of missing variables, in which case R becomes a matrix of missing data indicators. When every variable has missing values, this $\mathbf{R}$ matrix has the same number of rows and columns as the data matrix.

TABLE 1.3. Missing Data Indicator for MAR Job Performance Ratings

Job performance		
Complete	MAR	Indicator
9	—	0
13	—	0
10	—	0
8	—	0
7	—	0
7	7	1
9	9	1
9	9	1
11	11	1
7	7	1
7	7	1
10	10	1
11	11	1
15	15	1
10	10	1
10	10	1
12	12	1
14	14	1
16	16	1
12	12	1

Rubin's (1976) theory essentially views individuals as having a pair of observations on each variable: a score value that may or may not be observed (i.e., Y_{obs} or Y_{mis}) and a corresponding code on the missing data indicator, *R*. Defining the missing data as a variable implies that there is a probability distribution that governs whether *R* takes on a value of zero or one (i.e., there is a function or equation that describes the probability of missingness). For example, reconsider the cancer study that I described earlier in the chapter. If the quality of life scores are missing as a function of other variables such as age or education, then the coefficients from a logistic regression equation might describe the distribution of *R*. In practice, we rarely know why the data are missing, so it is impossible to describe the distribution of *R* with any certainty. Nevertheless, the important point is that *R* has a probability distribution, and the probability of missing data may or may not be related to other variables in the data set. As you will see, the nature of the relationship between *R* and the data is what differentiates the missing data mechanisms.

A More Precise Definition of the Missing Data Mechanisms

Having established some basic terminology, we can now revisit the missing data mechanisms in more detail. The formal definitions of the missing data mechanisms involve different probability distributions for the missing data indicator, *R*. These distributions essentially describe different relationships between *R* and the data. In practice, there is generally no way to specify the parameters of these distributions with any certainty. However, these details are not important because it is the presence or absence of certain associations that differentiates the missing data mechanisms.

The probability distribution for MNAR data is a useful starting point because it includes all possible associations between the data and missingness. You can write this distribution as

$$p(R|Y_{\text{obs}}, Y_{\text{mis}}, \phi) \qquad (1.1)$$

where p is a generic symbol for a probability distribution, R is the missing data indicator, Y_{obs} and Y_{mis} are the observed and missing parts of the data, respectively, and ϕ is a parameter (or set of parameters) that describes the relationship between R and the data. In words, Equation 1.1 says that the probability that R takes on a value of zero or one can depend on both Y_{obs} and Y_{mis}. Said differently, the probability of missing data on Y can depend on other variables (i.e., Y_{obs}) as well as on the underlying values of Y itself (i.e., Y_{mis}).

To put Equation 1.1 into context, reconsider the data set in Table 1.2. Equation 1.1 implies that the probability of missing data is related to an individual's IQ or job performance score (or both). Panel A of Figure 1.2 is a graphical depiction of these relationships that I adapted from a similar figure in Schafer and Graham (2002). Consistent with Equation1.1, the figure contains all possible associations (i.e., arrows) between R and the data. The box labeled Z represents a collection of unmeasured variables (e.g., motivation, health problems, turnover intentions, and job satisfaction) that may relate to the probability of missing data and to IQ and job performance. Rubin's (1976) missing data mechanisms are only concerned with relationships between R and the data, so there is no need to include Z in Equation 1.1. However, correlations between measured and unmeasured variables can induce spurious associations between R and $Y_,$ which underscores the point that Rubin's mechanisms are not real-world causal descriptions of the missing data.

An MAR mechanism occurs when the probability of missing data on a variable Y is related to another measured variable in the analysis model but not to the values of Y itself. This implies that R is dependent on Y_{obs} but not on Y_{mis}. Consequently, the distribution of missing data simplifies to

$$p(R|Y_{\text{obs}}, \phi) \qquad (1.2)$$

Equation 1.2 says that the probability of missingness depends on the observed portion of data via some parameter ϕ that relates Y_{obs} to R. Returning to the small job performance data set, observe that Equation 1.2 implies that an individual's propensity for missing data depends only on his or her IQ score. Panel B of Figure 1.2 depicts an MAR mechanism. Notice that there is no longer an arrow between R and the job performance scores, but a linkage remains between R and IQ. The arrow between R and IQ could represent a direct relationship between these variables (e.g., the company uses IQ as a selection measure), or it could be a spurious relationship that occurs when R and IQ are mutually correlated with one of the unmeasured variables in Z. Both explanations satisfy Rubin's (1976) definition of MAR, so the underlying causal process is unimportant.

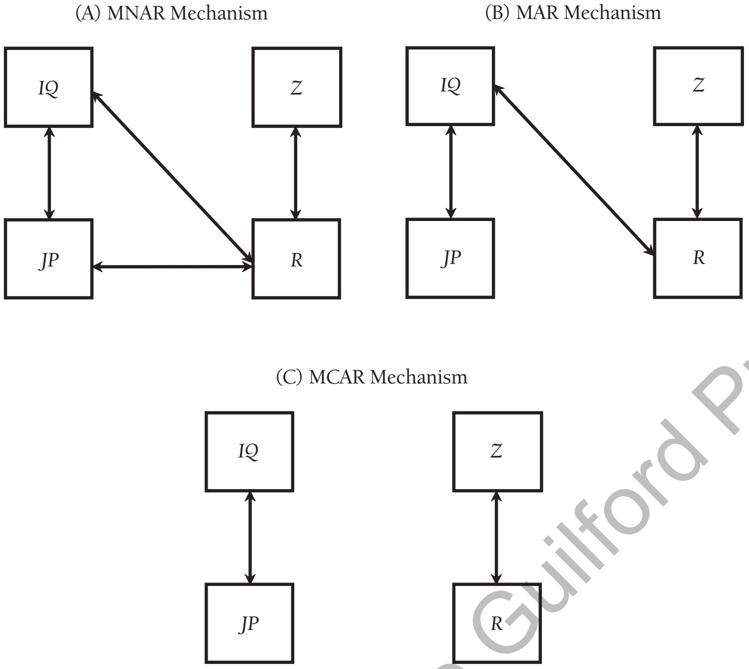

FIGURE 1.2. A graphical representation of Rubin's missing data mechanisms. The figure depicts a bivariate scenario in which IQ scores are completely observed and the job performance scores (*JP*) are missing for some individuals. The double-headed arrows represent generic statistical associations and ϕ is a parameter that governs the probability of scoring a 0 or 1 on the missing data indicator, *R*. The box labeled *Z* represents a collection of unmeasured variables.

Finally, the MCAR mechanism requires that missingness is completely unrelated to the data. Consequently, both Y_{obs} and Y_{mis} are unrelated to *R*, and the distribution of missing data simplifies even further to

$$p(R|\phi) \tag{1.3}$$

Equation 1.3 says that some parameter still governs the probability that *R* takes on a value of zero or one, but missingness is no longer related to the data. Returning to the job performance data set, note that Equation 1.3 implies that the missing data indicator is unrelated to both IQ and job performance. Panel C of Figure 1.2 depicts an MCAR mechanism. In this situation, the ϕ parameter describes possible associations between *R* and unmeasured variables, but there are no linkages between *R* and the data. Although it is not immediately obvious, panel C implies that the unmeasured variables in *Z* are uncorrelated with IQ and job performance because the presence of such a correlation could induce a spurious association between *R* and *Y*.

1.6 WHY IS THE MISSING DATA MECHANISM IMPORTANT?

Rubin's (1976) missing data theory involves two sets of parameters: the parameters that address the substantive research questions (i.e., the parameters that you would have estimated had there been no missing data) and the parameters that describe the probability of missing data (i.e., ϕ). Researchers rarely know why the data are missing, so it is impossible to describe ϕ with any certainty. For example, reconsider the cancer study described in the previous section. Quality of life scores could be missing as an additive function of age and education, as an interactive function of treatment group membership and baseline health status, or as a direct function of quality of life itself. The important point is that there is generally no way to determine or estimate the parameters that describe the propensity for missing data.

The parameters that describe the probability of missing data are a nuisance and have no substantive value (e.g., had the data been complete, there would be reason to worry about ϕ). However, in some situations these parameters may influence the estimation of the substantive parameters. For example, suppose that the goal of the cancer study is to estimate the mean quality of life score. Furthermore, imagine that a number of patients become so ill (i.e., their quality of life becomes so poor) that they can no longer participate in the study. In this scenario, ϕ is a set of parameters (e.g., logistic regression coefficients) that relates the probability of missing data to an individual's quality of life score. At an intuitive level, it would be difficult to obtain an accurate mean estimate because scores are disproportionately missing from the lower tail of the distribution. However, if the researchers happened to know the parameter values in ϕ, it would be possible to correct for the positive bias in the mean. Of course, the problem with this scenario is that there is no way to estimate ϕ.

Rubin's (1976) work is important because he clarified the conditions that need to exist in order to accurately estimate the substantive parameters without also knowing the parameters of the missing data distribution (i.e., ϕ). It ends up that these conditions depend on how you analyze the data. Rubin showed that likelihood-based analyses such as maximum likelihood estimation and multiple imputation do not require information about ϕ if the data are MCAR or MAR. For this reason, the missing data literature often describes the MAR mechanism as **ignorable missingness** because there is no need to estimate the parameters of the missing data distribution when performing analyses. In contrast, Rubin showed that analysis techniques that rely on a sampling distribution are valid only when the data are MCAR. This latter set of procedures includes most of the ad hoc missing data techniques that researchers have been using for decades (e.g., discarding cases with missing data).

From a practical standpoint, Rubin's (1976) missing data mechanisms are essentially assumptions that govern the performance of different analytic techniques. Chapter 2 outlines a number of missing data handling methods that have been mainstays in published research articles for many years. With few exceptions, these techniques assume an MCAR mechanism and will yield biased parameter estimates when the data are MAR or MNAR. Because these traditional methods require a restrictive assumption that is unlikely to hold in practice, they have increasingly fallen out of favor in recent years (Wilkinson & Task Force on Statistical Inference, 1999). In contrast, maximum likelihood estimation and multiple imputation yield unbiased parameter estimates with MCAR or MAR data. In some sense, maximum likelihood and multiple imputation are robust missing data handling procedures because they require less stringent assumptions about the missing data mechanism. However, these methods are not a perfect solution because they too will produce bias with MNAR data. Methodologists have developed analysis methods for MNAR data, but these approaches require strict assumptions that limit their practical utility. Chapter 10 outlines models for MNAR data and shows how to use these models to conduct sensitivity analyses.

1.7 HOW PLAUSIBLE IS THE MISSING AT RANDOM MECHANISM?

The methodological literature recommends maximum likelihood and multiple imputation because these approaches require the less stringent MAR assumption. It is reasonable to question whether this assumption is plausible, given that there is no way to test it. Later in the chapter, I describe a number of planned missing data designs that automatically produce MAR or MCAR data, but these situations are unique because missingness is under the researcher's control. In the vast majority of studies, missing values are an unintentional byproduct of data collection, so the MAR mechanism becomes an unverifiable assumption that influences the accuracy of the maximum likelihood and multiple imputation analyses.

As is true for most statistical assumptions, it seems safe to assume that the MAR assumption will not be completely satisfied. The important question is whether routine violations are actually problematic. The answer to this question is situation-dependent because not all violations of MAR are equally damaging. To illustrate, reconsider the job performance scenario I introduced earlier in the chapter. The definition of MNAR states that a relationship exists between the probability of missing data on *Y* and the values of *Y* itself. This association can occur for two reasons. First, it is possible that the probability of missing data is directly related to the incomplete outcome variable. For example, if the company terminates a number of individuals for poor performance prior to their 6-month evaluation, then there is a direct relationship between job performance and the propensity for missing data. However, an association between job performance and missingness can also occur because these variables are mutually correlated with an unmeasured variable. For example, suppose that individuals with low autonomy (an unmeasured variable) become frustrated and quit prior to their six-month evaluation. If low autonomy is also associated with poor job performance, then this unmeasured variable can induce a correlation between performance and missingness, such that individuals with poor job performance have a higher probability of missing their six-month evaluation.

Figure 1.3 is a graphical depiction of the previous scenarios. Note that I use a straight arrow to specify a causal influence and a double-headed arrow to denote a generic association. Although both diagrams are consistent with Rubin's (1976) definition of MNAR, they are not equally capable of introducing bias. Collins, Schafer, and Kam (2001) showed that a direct relationship between the outcome and missingness (i.e., panel A) can introduce substantial bias, whereas MNAR data that results from an unmeasured variable is problematic only when correlation between the unmeasured variable and the missing outcome is relatively strong (e.g., greater than .40). The situation in panel B seems even less severe when you consider that the IQ variable probably captures some of the variation that autonomy would have explained, had it been a measured variable that was included in the statistical

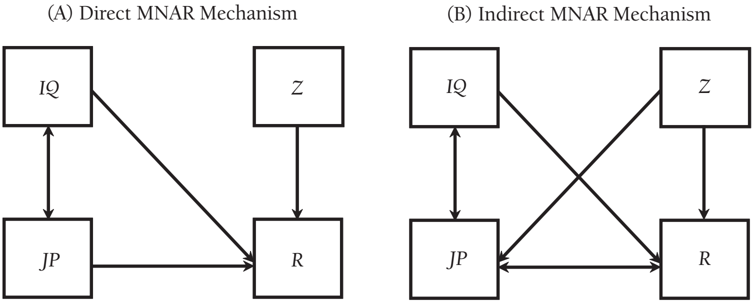

FIGURE 1.3. A graphical representation of two causal processes that produce MNAR data. The figure depicts a bivariate scenario in which IQ scores are completely observed and the job performance scores (*JP*) are missing for some individuals. The double-headed arrows represent generic statistical associations, and the straight arrows specify a causal influences. Panel A corresponds to a situation in which the probability of missing data is directly related to the missing outcome variable (i.e., the straight arrow between *JP* and *R*). Panel B depicts a scenario in which the probability of missing data is indirectly related to the missing outcome variable via the unmeasured cause of missingness in box *Z*.

analysis. This means that an unmeasured cause of missingness is problematic only if it has a strong relationship with the missing outcome after partialling out other measured variables. Schafer and Graham (2002, p. 173) argue that this is unlikely in most situations.

Notice that the MNAR mechanism in Panel B of Figure 1.3 becomes an MAR mechanism if autonomy is a measured variable that is included in the statistical analysis (i.e., the spurious correlation between job performance and *R* disappears once autonomy is partialled out). This suggests that you should be proactive about satisfying the MAR assumption by measuring variables that might explain missingness. For example, Graham, Taylor, Olchowski, and Cumsille (2006) suggest that variables such as reading speed and conscientiousness might explain why some respondents leave questionnaire items blank. In a longitudinal study, Schafer and Graham (2002) recommend using a survey question that asks respondents to report their likelihood of dropping out of the study prior to the next measurement occasion. As noted by Schafer and Graham (2002, p. 173), collecting data on the potential causes of missingness "may effectively convert an MNAR situation to MAR," so you should strongly consider this strategy when designing a study.

Of course, not all MNAR data are a result of unmeasured variables. In truth, the likelihood of the two scenarios in Figure 1.3 probably varies across research contexts. There is often a tendency to assume that data are missing for rather sinister reasons (e.g., a participant in a drug cessation study drops out, presumably because she started using again), and this presumption may be warranted in certain situations. For example, Hedeker and Gibbons (1997) describe data from a psychiatric clinical trial in which dropout was likely a function of response to treatment (e.g., participants in the placebo group were likely to leave the study because their symptoms were not improving, whereas dropouts in a drug condition experienced rapid improvement prior to dropping out). Similarly, Foster and Fang (2004) describe an evaluation of a conduct disorder intervention in which highly aggressive boys were less likely to continue participating in the study. However, you should not discount the possibility that a substantial proportion of the missing observations are MAR or even MCAR. For example, Graham, Hofer, Donaldson, MacKinnon, and Schafer (1997) and Enders, Dietz, Montague, and Dixon (2006) describe longitudinal studies that made systematic attempts to document the reasons for missing data. These studies had a substantial proportion of unplanned missing data, yet intensive follow-up analyses suggested that the missing data were largely benign (e.g., the most common reason for missing data was that students moved out of the school where the study took place).

Some researchers have argued that serious violations of MAR are relatively rare (Graham et al., 1997, p. 354; Schafer & Graham, 2002, p. 152), but the only way to evaluate the MAR assumption is to collect follow-up data from the missing respondents. Of course, this is difficult or impossible in many situations. Sensitivity analyses are also useful for assessing the potential impact of MNAR data. Graham et al. (1997, pp. 354–358) provide a good illustration of a sensitivity analysis; I discuss these procedures in Chapter 10.

1.8 AN INCLUSIVE ANALYSIS STRATEGY

The preceding section is overly simplistic because it suggests that the MAR assumption is automatically satisfied when the "cause" of missingness is a measured variable. In truth, the MAR mechanism is a characteristic of a specific analysis rather than a global characteristic of a data set. That is, some analyses from a given data set may satisfy the MAR assumption, whereas others are consistent with an MCAR or MNAR mechanism. To illustrate the subtleties of the MAR mechanism, consider a study that examines a number of health-related behaviors (e.g., smoking, drinking, and sexual activity) in a teenage population. Because of its sensitive nature, researchers decide to administer the sexual behavior questionnaire to participants who are above the age of 15. At first glance, this study may appear to satisfy the MAR assumption because a measured variable determines whether data are missing. However, this is not necessarily true.

Technically, MAR is satisfied only if the researchers incorporate age into the missing data handling procedure. For example, suppose that the researchers use a simple regression model to examine the influence of self-esteem on risky sexual behavior. Many software packages that implement maximum likelihood missing data handling methods can estimate a regression model with missing data, so this is a relatively straightforward analysis. However, the regression analysis is actually consistent with the MNAR mechanism and may produce biased parameter estimates, particularly if age and sexual activity are correlated. To understand the problem, consider Figure 1.4. This figure depicts an indirect MNAR mechanism that is similar to the one in Panel B of Figure 1.3. Age is not part of the regression model, so it effectively operates an unmeasured variable and induces an association between missingness and the sexual behavior scores; the figure denotes this spurious correlation as a dashed line. The bias that results from omitting age from the regression model may not be problematic and depends on the correlation between age and sexual activity. Nevertheless, the regression analysis violates the MAR assumption.

The challenge of satisfying the MAR assumption has prompted methodologists to recommend a so-called **inclusive analysis strategy** that incorporates a number of **auxiliary variables** into the analysis model or into the imputation process (Collins, Schafer, & Kam,

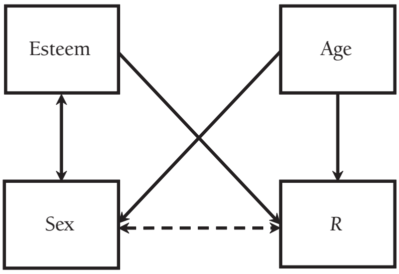

FIGURE 1.4. A graphical representation of an indirect MNAR mechanism. The figure depicts a bivariate scenario in which self-esteem scores are completely observed and sexual behavior questionnaire items are missing for respondents who are less than 15 years of age. If age (the "cause" of missingness) is excluded from the analysis model, it effectively acts as an unmeasured variable and induces an association between the probability of missing data and the unobserved sexual activity scores. The dashed line represents this spurious correlation. Including age in the analysis model (e.g., as an auxiliary variable) converts an MNAR analysis into an MAR analysis.

2001; Rubin, 1996; Schafer, 1997; Schafer & Graham, 2002). Auxiliary variables are variables you include in an analysis because they are either correlates of missingness or correlates of an incomplete variable. Auxiliary variables are not necessarily of substantive interest (i.e., you would not have included these variables in the analysis, had the data been complete), so their primary purpose is to fine-tune the missing data analysis by increasing power or reducing nonresponse bias. In the health study, age is an important auxiliary variable because it is a determinant of missingness, but other auxiliary variables may be correlates of the missing sexual behavior scores. For example, a survey question that asks teenagers to report whether they have a steady boyfriend or girlfriend is a good auxiliary variable because of its correlation with sexual activity. Theory and past research can help identify auxiliary variables, as can the MCAR tests described later in the chapter. Incorporating auxiliary variables into the missing data handling procedure does not guarantee that you will satisfy the MAR assumption, but it certainly improves the chances of doing so. I discuss auxiliary variables in detail in Chapter 5.

1.9 TESTING THE MISSING COMPLETELY AT RANDOM MECHANISM

MCAR is the only missing data mechanism that yields testable propositions. You might question the utility of testing this mechanism given that the majority of this book is devoted to techniques that require the less stringent MAR assumption. In truth, testing whether an entire collection of variables is consistent with MCAR is probably not that useful because some of the variables in a data set are likely to be missing in a systematic fashion. Furthermore, finding evidence for or against MCAR does not change the recommendation to use maximum likelihood or multiple imputation. However, identifying individual variables that are not MCAR is potentially useful because there may be a relationship between these variables and the probability of missingness. As I explained previously, methodologists recommend incorporating correlates of missingness into the missing data handling procedure because doing so can mitigate bias and improve the chances of satisfying the MAR assumption (Collins et al., 2001; Rubin, 1996; Schafer, 1997; Schafer & Graham, 2002).

To illustrate how you might use the information from an MCAR test, suppose that a psychologist is studying quality of life in a group of cancer patients and finds that patients who refused the quality of life questionnaire have a higher average age and a lower average education than the patients who completed the survey. These mean differences provide compelling evidence that the data are not MCAR and suggest a possible relationship between the demographic variables and the probability of missing data. Incorporating the demographic characteristics into the missing data handling procedure (e.g., using the auxiliary variable procedures in Chapter 5) adjusts for age- and education-related bias in the quality of life scores and increases the chances of satisfying the MAR assumption. Consequently, using MCAR tests to identify potential correlates of missingness is often a useful starting point, even if you have no interest in assessing whether an entire set of variables is MCAR.

Rubin's (1976) definition of MCAR requires that the observed data are a simple random sample of the hypothetically complete data set. This implies that the cases with missing data belong to the same population (and thus share the same mean vector and covariance matrix) as the cases with complete data. Kim and Bentler (2002) refer to this condition as **homogeneity of means and covariances**. One way to test for homogeneity of means is to separate the missing and the complete cases on a particular variable and examine group mean differences on other variables in the data set. Testing for homogeneity of covariances follows a similar logic and examines whether the missing data subgroups have different variances and covariances. Finding that the missing data patterns share a common mean vector and a common covariance matrix provides evidence that the data are MCAR, whereas group differences in the means or the covariances provide evidence that the data are not MCAR.

Methodologists have proposed a number of methods for testing the MCAR mechanism (Chen & Little, 1999; Diggle, 1989; Dixon, 1988; Kim & Bentler, 2002; Little, 1988; Muthén, Kaplan, & Hollis, 1987; Park & Lee, 1997; Thoemmes & Enders, 2007). This section describes two procedures that evaluate mean differences across missing data patterns. I omit procedures that assess homogeneity of covariances because it seems unlikely that covariance differences would exist in the absence of mean differences. In addition, simulation studies offer conflicting evidence about the performance of covariance-based tests (Kim & Bentler, 2002; Thoemmes & Enders, 2007). It therefore seems safe to view these procedures with caution until further research accumulates. Interested readers can consult Kim and Bentler (2002) for an overview of covariance-based tests.

Univariate *t*-Test Comparisons

The simplest method for assessing MCAR is to use a series of independent *t* tests to compare missing data subgroups (Dixon, 1988). This approach separates the missing and the complete cases on a particular variable and uses a *t* test to examine group mean differences on other variables in the data set. The MCAR mechanism implies that the cases with observed data should be the same as the cases with missing values, on average. Consequently, a nonsignificant *t* test provides evidence that the data are MCAR, whereas a significant *t* statistic (or alternatively, a large mean difference) suggests that the data are MAR or MNAR.

To illustrate the *t*-test approach, reconsider the data in Table 1.1. To begin, I used the job performance scores to create a binary missing data indicator and subsequently used independent *t* tests to assess group mean differences on IQ and psychological well-being. The missing and complete cases have IQ means of 88.50 and 111.50, respectively, and Welch's *t* test indicated that this mean difference is statistically significant, $t(14.68) = 6.43$, $p < .001$. Considering psychological well-being, the means for the missing and complete cases are 9.13 and 11.44, respectively, and the *t* test was not significant, $t(11.70) = 1.39$, $p = .19$. Collectively, these tests suggest that the job performance ratings are not MCAR because the missing and observed cases systematically differ with respect to IQ. This conclusion is correct because I deleted job performance scores for the cases in the lower half of the IQ distribution. Next, I repeated this procedure by forming a missing data indicator from the psychological well-being scores and by testing whether the resulting groups had different IQ means (it was impossible to compare the job performance means because only one case from the missing data group had a job performance score). The *t* test indicated that the group means are equivalent, $t(3.60) = .50$, $p = .65$, which correctly provides support for the MCAR mechanism.

The *t*-test approach has a number of potential problems to consider. First, generating the test statistics can be very cumbersome unless you have a software package that automates the process (e.g., the SPSS Missing Values Analysis module). Second, the tests do not take the correlations among the variables into account, so it is possible for a missing data indicator to produce mean differences on a number of variables, even if there is only a single cause of missingness in the data. Related to the previous points, the potential for a large number of statistical tests and the possibility of spurious associations seem to warrant some form of type I error control. The main reason for implementing the *t*-test approach is to identify auxiliary variables that you can later adjust for in the missing data handling procedure. I would argue against any type of error control procedure because there is ultimately no harm in using auxiliary variables that are unrelated to missingness (Collins et al., 2001). Another problem with the *t*-test approach is the possibility of very small group sizes (e.g., there are only three cases in Table 1.1 with missing well-being scores). This can decrease power and make it impossible to perform certain comparisons. To offset a potential loss of power, it might be useful to augment the *t* tests with a measure of effect size such as Cohen's (1988) standardized mean difference. Finally, it is important to note that mean comparisons do not provide a conclusive test of MCAR because MAR and MNAR mechanisms can produce missing data subgroups with equal means.

Little's MCAR Test

Little (1988) proposed a multivariate extension of the *t*-test approach that simultaneously evaluates mean differences on every variable in the data set. Unlike univariate *t* tests, Little's procedure is a global test of MCAR that applies to the entire data set. Omnibus tests of the MCAR mechanism are probably not that useful because they provide no way to identify potential correlates of missingness (i.e., auxiliary variables). Nevertheless, Little's test is available in some statistical software packages (e.g., the SPSS Missing Values Analysis module), so the procedure warrants a description.

Like the *t*-test approach, Little's test evaluates mean differences across subgroups of cases that share the same missing data pattern. The test statistic is a weighted sum of the standardized differences between the subgroup means and the grand means, as follows:

$$d^2 = \sum_{j=1}^{J} n_j \left(\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_j^{(\mathrm{ML})} \right)^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_j^{-1} \left(\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_j^{(\mathrm{ML})} \right) \tag{1.4}$$

where n_j is the number of cases in missing data pattern j, $\hat{\boldsymbol{\mu}}_j$ contains the variable means for the cases in missing data pattern j, $\hat{\boldsymbol{\mu}}_j^{(\mathrm{ML})}$ contains maximum likelihood estimates of the grand means, and $\hat{\boldsymbol{\Sigma}}_j$ is the maximum likelihood estimate of the covariance matrix. The j subscript indicates that the number of elements in the parameter matrices vary across missing data patterns. The d^2 statistic is essentially a weighted sum of J squared z scores. Specifically, the parentheses contain deviation scores that capture differences between pattern j's means and the corresponding grand means. With MCAR data, the subgroup means should be within sampling error of the grand means, so small deviations are consistent with an MCAR mechanism. In matrix algebra, multiplying by the matrix inverse is analogous to division, so the $\hat{\boldsymbol{\Sigma}}_j^{-1}$ term functions like the denominator of the z score formula by converting the raw deviation values to a standardized metric. Finally, multiplying the squared z values by n_j weights each pattern's contribution to the test statistic. When the null hypothesis is true (i.e., the data are MCAR), d^2 is approximately distributed as a chi-square statistic with $\Sigma k_j - k$ degrees of freedom, where k_j is the number of complete variables for pattern j, and k is the total number of variables. Consistent with the univariate t-test approach, a significant d^2 statistic provides evidence against MCAR.

To illustrate Little's MCAR test, reconsider the small data set in Table 1.1. The data contain four missing data patterns: cases with only IQ scores ($n_j = 2$), cases with IQ and well-being scores ($n_j = 8$), cases with IQ and job performance scores ($n_j = 1$), and cases with complete data on all three variables ($n_j = 9$). The test statistic in Equation 1.4 compares the subgroup means to the maximum likelihood estimates of the grand means. I outline maximum likelihood missing data handling in Chapter 4, but for now, the necessary parameter estimates are as follows:

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \hat{\ }_{IQ} \\ \hat{\ }_{JP} \\ \hat{\ }_{WB} \end{bmatrix} = \begin{bmatrix} 100.00 \\ 10.23 \\ 10.27 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\sigma}^2_{IQ} & \hat{\sigma}_{IQ,JP} & \hat{\sigma}_{IQ,WB} \\ \hat{\sigma}_{JP,IQ} & \hat{\sigma}^2_{JP} & \hat{\sigma}_{JP,WB} \\ \hat{\sigma}_{WB,IQ} & \hat{\sigma}_{WB,JP} & \hat{\sigma}^2_{WB} \end{bmatrix} = \begin{bmatrix} 189.60 & 22.31 & 12.21 \\ 22.31 & 8.68 & 5.61 \\ 12.21 & 6.50 & 11.04 \end{bmatrix}$$

To begin, consider the group of cases with data on only IQ ($n_j = 2$). This pattern has an IQ mean of 91.50, so its contribution to the d^2 statistic is as follows:

$$d_j^2 = 2(91.50 - 100.00)(189.60^{-1})(91.50 - 100.00) = 0.762$$

Next, consider the subgroup of cases with complete data on IQ and well-being ($n_j = 8$). The IQ and well-being means for this pattern are 87.75 and 9.13, respectively, and the contribution to the d^2 statistic is

$$d_j^2 = 8\left(\begin{bmatrix} 87.75 \\ 9.13 \end{bmatrix} - \begin{bmatrix} 100.00 \\ 10.27 \end{bmatrix} \right)^T \begin{bmatrix} 189.60 & 12.21 \\ 12.21 & 11.04 \end{bmatrix}^{-1} \left(\begin{bmatrix} 87.75 \\ 9.13 \end{bmatrix} - \begin{bmatrix} 100.00 \\ 10.27 \end{bmatrix} \right) = 6.432$$

In both of the previous examples, notice that $\hat{\boldsymbol{\mu}}_j$ and $\hat{\boldsymbol{\Sigma}}_j$ contain the maximum likelihood estimates that correspond to the observed variables for a particular pattern (i.e., the estimates that correspond to the missing variables do not appear in the matrices). Repeating the computations for the remaining missing data patterns and summing the resulting d_j^2 values yields $d^2 = 14.63$, and referencing the test statistic to a chi-square distribution with 5 degrees of freedom returns a probability value of $p = .01$. The null hypothesis for Little's test states that the data are MCAR, so a statistically significant test statistic provides evidence against the MCAR mechanism.

Like the *t*-test approach, Little's test has a number of problems to consider. First, the test does not identify the specific variables that violate MCAR, so it is only useful for testing an omnibus hypothesis that is unlikely to hold in the first place. Second, the version of the test outlined above assumes that the missing data patterns share a common covariance matrix. MAR and MNAR mechanisms can produce missing data patterns with different variances and covariances, and the test statistic in Equation 1.4 would not necessarily detect covariance-based deviations from MCAR. Third, simulation studies suggest that Little's test suffers from low power, particularly when the number of variables that violate MCAR is small, the relationship between the data and missingness is weak, or the data are MNAR (Thoemmes & Enders, 2007). Consequently, the test has a propensity to produce Type II errors and can lead to a false sense of security about the missing data mechanism. Finally, mean comparisons do not provide a conclusive test of MCAR because MAR and MNAR mechanisms can produce missing data subgroups with equal means.

1.10 PLANNED MISSING DATA DESIGNS

The next few sections outline research designs that produce MCAR or MAR data as an intentional by-product of data collection. The idea of intentional missing data might seem odd at first, but you may already be familiar with a number of these designs. For example, in a randomized study with two treatment conditions, each individual has a hypothetical score from both conditions, but participants only provide a response to their assigned treatment condition. The unobserved response to the other condition (i.e., the potential outcome or counterfactual) is MCAR. Viewing randomized experiments as a missing data problem is popular in the statistics literature and is a key component of Rubin's Causal Model (Rubin, 1974, 1978a; West & Thoemmes, in press). A fractional factorial design (Montgomery, 1997) is another research design that yields MCAR missing data. In a fractional factorial, you purposefully select a subset of experimental conditions from a full factorial design and randomly assign participants to these conditions. A classic example of intentional MAR data occurs in selection designs where scores on one variable determine whether respondents provide data on a second variable. For example, universities frequently use the Graduate Record Exam (GRE) as a selection tool for graduate school admissions, so first-year grade point averages are subsequently missing for students who score below some GRE threshold. A related issue arises in survey designs where the answer to a screener question dictates a particular skip pattern. Selection problems such as this have received considerable attention in the methodological literature (Sackett & Yang, 2000) and date back to Pearson's (1903) work on range restriction.

The previous designs are classic examples of intentional missing data that do not necessarily require missing data techniques. The advent of maximum likelihood estimation and multiple imputation has prompted methodologists to develop specialized **planned missing data designs** that address a number of practical problems (Graham et al., 2006). For example, researchers often face constraints on the number of questionnaire items that they can reasonably expect respondents to answer, and this problem becomes more acute in longitudinal studies where respondents fill out questionnaire batteries on multiple occasions. Limiting the number of variables is one obvious solution to this problem, but introducing planned missing data is another possibility. In a planned missing data design, you distribute the questionnaire items across different forms and administer a subset of the forms to each respondent. This strategy allows you to collect data on the full set of questionnaire items while simultaneously reducing respondent burden.

Planned missingness is not limited to questionnaire data and has a number of other interesting applications. For example, suppose that a researcher wants to use two data collection methods, one of which is very expensive. To illustrate, imagine a study in which a researcher is collecting brain image data. Ideally, she would like to collect her data using magnetic resonance imaging (MRI), but the MRI is very expensive and she has difficulty accessing it for extended periods. However, she can readily collect data using the less expensive computed tomography (CT) scan. Planned missingness is ideally suited for this situation because the researcher can collect CT data from every participant and restrict the MRI data to a subset of her sample. A similar example occurs with body fat measurements from an exercise physiology study. A researcher can readily use a set of calipers to take skinfold measurements from all of his subjects but might use a more expensive technique (e.g., air displacement in a BOD POD) on a subset of the participants. Importantly, maximum likelihood and multiple imputation allow researchers to analyze data from planned missingness designs without having to discard the incomplete cases. For example, the exercise physiologist can use the entire sample to estimate the associations between the expensive measure and other study variables, even though a subset of the cases has missing data on the expensive measure.

Planned missing data strategies have been available for many years and have a number of interesting applications (Johnson, 1992; Lord, 1962; Raghunathan & Grizzle, 1995; Shoemaker, 1973). I focus primarily on the planned missing data designs outlined by John Graham and his colleagues (Graham et al., 1996; Graham, Taylor, & Cumsille, 2001; Graham et al., 2006). In particular, the subsequent sections describe a three-form design that is widely applicable to questionnaire data collection and planned missingness designs for longitudinal studies. Readers interested in additional details on planned missingness designs can consult Graham et al. (2006).

As an aside, my experience suggests that researchers tend to view the idea of planned missing data with some skepticism and are often reluctant to implement this strategy. This skepticism probably stems from a presumption that missing data can bias the analysis results. However, the planned data designs in this section produce MCAR data, so the only potential downside is a loss of statistical power. Planned missingness designs are very flexible and allow you to address power concerns by restricting the missing data to certain variables. Every research study involves compromises, so you have to decide whether collecting additional variables offsets the resulting loss of power. Of course, increasing the sample size will always improve power, but this may not be feasible. In any case, planned missing data designs are highly useful and underutilized tools that will undoubtedly increase in popularity as researchers become familiar with their benefits.

1.11 THE THREE-FORM DESIGN

Researchers in many disciplines use multiple-item questionnaires to measure complex constructs. For example, psychologists routinely use several questionnaire items to measure depression, each of which taps into a different depressive symptom (e.g., sadness, lack of energy, sleep difficulties, feelings of hopelessness). Using multiple-item questionnaires to measure even a relatively small number of variables can introduce a substantial respondent burden. Graham et al. (1996) addressed this problem with a **three-form design** that distributes a subset of questionnaire items to each respondent. The design divides the item pool into four sets (X, A, B, and C) and allocates these sets across three questionnaire forms, such that each form includes X and is missing A, B, or C. Table 1.4 shows the distribution of the item sets across the three questionnaire forms. Note that each item set can include multiple questionnaires or combinations of items from multiple questionnaires (e.g., item set X can include a depression questionnaire and a self-esteem questionnaire).

To illustrate the three-form design, suppose that a researcher plans to use eight questionnaires, each of which has 10 items. Concerned that her study participants will not have time to complete all 80 questions, she uses a three-form design to reduce respondent burden. Table 1.5 shows what the three-form design would look like if the researcher equally distributed her questionnaires across the four item sets (i.e., she assigns two questionnaires to set X, A, B, and C). Notice that the 3-form design allows the researcher to collect data on 80 questionnaire items, even though any given respondent only answers 60 items. Importantly, maximum likelihood estimation and multiple imputation allow the researcher to analyze the data without discarding incomplete cases.

The three-form design is flexible and does not require an equal number of questionnaire items in each item set. For example, the researcher could use the three-form design in Table 1.6 if she wanted to increase the number of variables in the X set, although this would require each participant to answer 70 items. In addition, there is no need to group questionnaire items together in the same set (e.g., assign all Q_1 items to set X), and it is possible to distribute questionnaire items across more than one item set (e.g., assign five of the Q_1 items to set

TABLE 1.4. Missing Data Pattern for a Three-Form Design

	Item sets			
Form	X	A	B	C
1	✓	—	✓	✓
2	✓	✓	—	✓
3	✓	✓	✓	—

Note. A check mark denotes complete data.

TABLE 1.5. Missing Data Pattern for a Three-Form Design with Eight Questionnaires

	Item Sets							
	X		A		B		C	
Form	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8
1	✓	✓	✓	✓	✓	✓	—	—
2	✓	✓	✓	✓	—	—	✓	✓
3	✓	✓	—	—	✓	✓	✓	✓
Items	10	10	10	10	10	10	10	10

Note. A check mark denotes complete data.

X and the remaining five items to set A). Graham et al. (1996) used a computer simulation study to investigate this issue and found that splitting the questionnaire across multiple item sets reduced the standard errors from a regression analysis. Despite this slight power advantage, Graham et al. (2006) recommend grouping the questionnaire items together in the same item set because this strategy facilitates the statistical analyses, particularly with a large number of variables.

How Does the Three-Form Design Impact Power?

The main downside of planned data designs is a potential loss of statistical power. Fortunately, you can mitigate this power loss by carefully aligning the questionnaire forms to your substantive goals. However, doing so requires an understanding of some of the subtleties of the three-form design and its influence on statistical power. This section describes a number of these subtleties and illustrates the influence of planned missing data on statistical power. For simplicity, I restrict the subsequent discussion to correlations, but the basic ideas generalize to other analyses. Interested readers can find a more thorough discussion of power in Graham et al. (2006).

There are essentially three tiers of power in the three-form design, and the power of a given statistical test depends on the particular combination of item sets that are involved. To illustrate, reconsider the three-form design in Table 1.5. Table 1.7 shows a **covariance coverage matrix** that gives the percentage of respondents with complete data on a given question-

TABLE 1.6. Missing Data Pattern for a Three-Form Design with Unequal Item Sets

	Item sets							
	X					A	B	C
Form	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8
1	✓	✓	✓	✓	✓	✓	✓	—
2	✓	✓	✓	✓	✓	✓	—	✓
3	✓	✓	✓	✓	✓	—	✓	✓
Items	10	10	10	10	10	10	10	10

Note. A check mark denotes complete data.

TABLE 1.7. Covariance Coverage Matrix for a Three-Form Design

		Item set							
		X		A		B		C	
Set	Scale	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8
X	Q_1	100%							
	Q_2	100%	100%						
A	Q_3	66%	66%	66%					
	Q_4	66%	66%	66%	66%				
B	Q_5	66%	66%	33%	33%	66%			
	Q_6	66%	66%	33%	33%	66%	66%		
C	Q_7	66%	66%	33%	33%	33%	33%	66%	
	Q_8	66%	66%	33%	33%	33%	33%	66%	66%

Note. The percentages represent the amount of complete data for a variable or variable pair.

naire (the diagonal elements) or pair of questionnaires (the off-diagonal elements). The entire sample has complete data on a single pair of questionnaires (i.e., Q_1 and Q_2), 15 of the questionnaire pairs have a 33% missing data rate (e.g., Q_1 and Q_3), and 12 pairs have 66% missing data (e.g., Q_3 and Q_5). Not surprisingly, the percentages in Table 1.7 have an impact on statistical power. Analyses that involve two variables from the X set (e.g., the correlation between Q_1 and Q_2) have the highest power because these variables have no missing data. A second tier of associations has somewhat less power and includes correlations between an X variable and a variable from item set A, B, or C (e.g., the correlation between Q_1 and Q_3) and relationships between variables within set A, B, or C (e.g., the correlation between Q_3 and Q_4). Finally, any correlations between A, B, or C variables (e.g., the correlation between Q_3 and Q_5) will have the lowest power.

With such a large proportion of missing data, you might expect certain associations to produce abysmal power. However, this is not necessarily true. To illustrate, I performed two computer simulation studies that examined the influence of the three-form design on power. To mimic the previous research scenario, I generated 5,000 samples of $N = 300$, each with eight normally distributed variables. The first simulation generated variables with a population correlation of $\rho = .30$, and the second simulation generated data from a population with $\rho = .10$. These population correlations correspond to Cohen's (1988) benchmarks for a medium and a small effect size, respectively. I subsequently deleted data according to the three-form design in Table 1.5 and then used maximum likelihood missing data handling to estimate the sample correlation matrix for each of the 5,000 replicates. Because I generated the data from a population with a nonzero correlation, the proportion of the 5,000 replications that produced a statistically significant correlation is an estimate of power.

Table 1.8 gives the power estimates from the simulation studies. To begin, consider the power values from the $\rho = .30$ simulation. Notice that the correlation between Q_1 and Q_3 (the two X set variables) had a power of 1.00. These variables had complete data, so this power estimate serves as a useful benchmark for assessing the impact of planned missingness. It may be somewhat surprising and counterintuitive to find that the decrease in power was not

TABLE 1.8. Correlation Power Estimates from a Three-Form Design

			Item set							
			X		A		B		C	
ρ	Set	Scale	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8
0.3	X	Q_1	—							
		Q_2	1.00	—						
	A	Q_3	.99	.99	—					
		Q_4	.99	.99	.99	—				
	B	Q_5	.99	.99	.90	.90	—			
		Q_6	.99	.99	.90	.90	.99	—		
	C	Q_7	.99	.99	.91	.91	.90	.91	—	
		Q_8	.99	.99	.90	.91	.91	.90	.99	—
0.1	X	Q_1	—							
		Q_2	.41	—						
	A	Q_3	.29	.30	—					
		Q_4	.30	.30	.28	—				
	B	Q_5	.30	.30	.18	.18	—			
		Q_6	.30	.30	.19	.18	.29	—		
	C	Q_7	.30	.31	.20	.19	.19	.18	—	
		Q_8	.29	.29	.18	.18	.18	.19	.29	—

commensurate with overall reduction in sample size. For example, consider the correlation between Q_1 and Q_3. A 33% missing data rate on Q_3 produced a 1% drop in power. The correlation between Q_3 and Q_5 is even more remarkable because one-third of the sample had complete data on this variable pair, yet power decreased by only 10%. The fact that power did not decrease dramatically is largely a by-product of maximum likelihood estimation. As you will see in Chapter 4, maximum likelihood uses the entire sample to estimate the parameters, so estimation effectively borrows information from the observed data to estimate the parameters of the incomplete variables (e.g., cases with missing Q_3 scores have Q_1 data that can help estimate the correlation between Q_1 and Q_3). Consequently, the loss of power from a planned missing data design is not necessarily as extreme as you might expect.

Next, consider the power estimates from the ρ = .10 simulation. In this situation, the correlation between the two complete variables (i.e., Q_1 and Q_2) had a power value of .41. Again, this power estimate serves as a useful benchmark for assessing the impact of planned missingness. Consistent with the previous simulation results, the decrease in power was not commensurate with overall reduction in sample size, although it was more nearly so. For example, the variable pairs with 33% missing data had an average power decrease of approximately 28%, while power dropped by roughly 55% for the variable pairs with a 66% missing data rate. In this simulation, relatively weak correlations limited the amount of information that maximum likelihood could borrow from the observed data, so the drop in power more closely approximates the missing data rate. As a rule, the impact of missing data on power will diminish as the correlations among the variables increase in magnitude.

Increasing the number of variables in the X set is one way to improve the power of a planned missingness design because it will increase the number of hypotheses that you can test with the full sample. Fortunately, the three-form design is flexible and does not require an equal distribution of questionnaire items across the four item sets. For example, the three-form design in Table 1.6 assigns five questionnaires to the X set and one questionnaire to each of the remaining sets. This design dramatically increases the number of variable pairs with complete data and decreases the number of tests with low power. Effect size is another factor that you can use to manipulate the power of a planned missing data design. For example, variables that you expect to produce a large effect size are good candidates for the A, B, or C set because they have lower sample size requirements. Conversely, you should consider placing a variable in the X set if you expect it to produce a small effect size because doing so will maximize power. Implementing a planned missingness design clearly requires some careful preparation, but these designs are very flexible and allow you to balance substantive and power concerns. Graham et al. (2006) provide additional details on the power of a three-form design.

Estimating Interaction Effects from a Three-Form Design

There are a number of nuances to consider when deciding how to distribute questionnaires across the four item sets. The previous section clearly suggests that the placement of a questionnaire influences statistical power. Questionnaire placement becomes even more critical when the goal is to estimate interaction effects. Unlike some planned missing data designs, the three-form design allows you to estimate every zero-order association in the data. However, the design does have limitations for testing higher-order effects.

Returning to the three-form design in Table 1.5, suppose that the researcher wants to examine whether Q_5 moderates the relationship between Q_3 and Q_7 (i.e., a B variable moderates the association between an A variable and a C variable). One way to address this question is to estimate a regression model with Q_3, Q_5, and the Q_3Q_5 product term as predictors of Q_7 (Aiken & West, 1991). However, it is impossible to estimate this regression model from the three-form design in Table 1.5. To illustrate the problem, Table 1.9 shows the missing data patterns that result when you form a product term between an A variable and a B variable (e.g., the Q_3Q_5 product term). Notice that one-third of the sample has complete data on both A and B (and thus the AB product term), but this subset of cases does not have data on the criterion variable from the C set. Consequently, there is no way to estimate the association between the outcome variable and the product term.

The three-form design does allow for two-way interactions, but one or more of the analysis variables must be from the X set (it does not matter whether this variable is a predictor or the criterion). To illustrate, suppose that an X variable moderates the association between a B variable and a C variable (e.g., a regression model with X, B, and the XB product term as predictors of C). Table 1.9 shows the missing data patterns for this new configuration of variables. Notice that every bivariate relationship among the regression model variables appears

TABLE 1.9. Missing Data Pattern for a Three-Form Design with Interaction Terms

	Item sets				Interaction terms	
Form	X	A	B	C	AB	XB
1	✓	—	✓	✓	—	✓
2	✓	✓	—	✓	—	—
3	✓	✓	✓	—	✓	✓

Note. A check mark denotes complete data.

in at least one questionnaire form, so it is now possible to estimate the model. Not surprisingly, questionnaire placement becomes more complex with three-way interactions. The three-form design does allow you to estimate certain three-way interactions, but the X set must include the criterion variable and at least one of the predictor variables.

1.12 PLANNED MISSING DATA FOR LONGITUDINAL DESIGNS

The problem of respondent burden can be particularly acute in longitudinal studies where participants fill out questionnaire batteries on multiple occasions. Graham et al. (2001) applied the logic of the three-form design to longitudinal data and investigated the power of several planned missingness designs. The basic idea behind these designs is to split the sample into a number of random subgroups and impose planned missing data patterns on each subgroup. Table 1.10 is an example of one such design where the random subgroups have missing data at a single wave.

Graham et al. (2001) outlined a number of planned missing data designs and examined each design's power to detect an intervention effect in a longitudinal analysis. The design in Table 1.10 was 94% as powerful as a complete-data analysis, but there were other designs that produced comparable power with fewer data points. For example, Table 1.11 shows a design that was 91% as powerful as a complete-data analysis but eliminated 44% of the total data points. (By data points, I mean the total number of observations in the data matrix.) The interesting thing about these results is that the planned missing data designs were actually

TABLE 1.10. Planned Missing Data Pattern 1 for a Longitudinal Design

	Data collection wave					
Group	1	2	3	4	5	% of N
1	✓	✓	✓	✓	✓	16.7
2	✓	✓	✓	✓	—	16.7
3	✓	✓	✓	—	✓	16.7
4	✓	✓	—	✓	✓	16.7
5	✓	—	✓	✓	✓	16.7
6	—	✓	✓	✓	✓	16.7

Note. A check mark denotes complete data.

TABLE 1.11. Planned Missing Data Pattern 2 for a Longitudinal Design

	Data collection wave					
Group	1	2	3	4	5	% of *N*
1	✓	✓	✓	✓	✓	9.1
2	✓	✓	✓	—	—	10.1
3	✓	✓	—	✓	—	10.1
4	✓	—	✓	✓	—	10.1
5	✓	✓	—	—	✓	20.2
6	✓	—	✓	—	✓	20.2
7	✓	—	—	✓	✓	20.2

Note. A check mark denotes complete data.

more powerful than a complete-data design that used the same number of data points. This has important implications for designing a longitudinal study. For example, suppose that each assessment (i.e., data point) costs $50 to administer and your grant budget allows you to collect 1,000 assessments. Graham et al.'s simulation results suggest that collecting complete data from *N* participants will actually yield less power than collecting incomplete data from a larger number of respondents.

The Graham et al. (2001) designs are particularly useful for studies that examine change following an intervention or a treatment. However, many researchers are interested in developmental processes that involve age-related change (e.g., the development of reading skills in early elementary school, the development of religiousness throughout the life span, the development of behavioral problems during the teenage years). The so-called **cohort-sequential design** (Duncan, Duncan, & Hops, 1996; Nesselroade & Baltes, 1979) is a common planned missing data design that is ideally suited for this type of research question.

The basic idea behind the cohort-sequential design is to combine a number of short-term longitudinal studies into a single longitudinal data analysis. You do this by sampling different age cohorts at the initial data collection wave and following each cohort over the same period. Table 1.12 shows the cohort-sequential design from a 3-year study of teenage alcohol use (Duncan et al., 1996). Notice that each age cohort has three waves of intentional missing data (e.g., the 12-year-olds have missing data at ages 15, 16, and 17, the 13-year-olds have missing data at ages 12, 16, and 17, and so on). Maximum likelihood missing data handling allows you to combine data from multiple cohorts into a single data analysis, so you can examine change over a developmental span that exceeds the data collection period. For example, Duncan et al. (1996) used the design in Table 1.12 to examine the change in alcohol use over the 5-year period between ages 12 and 17. Like other planned missingness designs, the cohort-sequential design yields MCAR data.

The cohort-sequential design is extremely useful for developmental research but has an important limitation. Unlike the other designs in this section, the cohort-sequential design includes variable pairs that are completely missing. For example, the design in Table 1.12 yields missing data for six variable pairs: ages 12 and 15, 12 and 16, 12 and 17, 13 and 16, 13 and 17, and 14 and 17. These missing data patterns pose no problem for a longitudinal growth curve analysis, but they limit your ability to estimate zero-order correlations. The only way to eliminate this problem is to collect data across the entire developmental span, but this

TABLE 1.12. Missing Data Pattern for a Cohort-Sequential Design

	Yearly data collection points					
Cohort	12	13	14	15	16	17
12	✓	✓	✓	—	—	—
13	—	✓	✓	✓	—	—
14	—	—	✓	✓	✓	—
15	—	—	—	✓	✓	✓

Note. A check mark denotes complete data.

defeats the purpose of the design. Despite this important limitation, the cohort-sequential design is a useful tool for examining age-related change that is quite common, particularly in psychological research. As an aside, the fact that certain correlations are inestimable rules out multiple imputation as a missing data handling technique for this design (the sample covariance matrix plays an integral role in the imputation process). This problem is not a concern when using maximum likelihood to estimate a growth curve model.

1.13 CONDUCTING POWER ANALYSES FOR PLANNED MISSING DATA DESIGNS

Estimating power is one of the difficulties associated with implementing a planned missing data design. The power loss in these designs is generally not proportional to the decrease in the sample size and depends on the magnitude of the correlations among the methods. This makes it very difficult to get accurate power estimates from standard analysis techniques. Researchers have outlined power analysis techniques that account for missing data, but these approaches are limited in scope (Hedeker, Gibbons, & Waternaux, 1999; Tu et al., 2007). Monte Carlo computer simulations are a useful alternative that you can use to estimate power for virtually any analysis. This section describes how to use computer simulations to estimate power for the three-form design, but the basic approach generalizes to any number of power analyses, with or without missing data. Paxton, Curran, Bollen, Kirby, and Chen (2001) give a more detailed overview of Monte Carlo methodology, and Muthén and Muthén (2002) illustrate Monte Carlo power simulations.

A Monte Carlo simulation generates a large number of samples from a population with a hypothesized set of parameter values. Estimating a statistical model on each artificial sample and saving the resulting parameter estimates yield an empirical sampling distribution for each model parameter. The ultimate goal of a power simulation is to determine the proportion of statistically significant parameter estimates in this distribution. Many statistical software packages have built-in data generation routines that do not require much programming, so it is relatively straightforward to perform power simulations. Structural equation modeling packages are particularly useful because they offer a variety of different data generation and analysis options. Some of these packages also have a number of built-in routines for simulating missing data.*

*Analysis syntax is available on the companion website, *www.appliedmissingdata.com*.

The first step of a computer simulation is to specify the population parameters. In my previous power simulations, I specified eight standardized variables from a normally distributed population with correlations of $\rho = .10$ and .30. This is a very straightforward data generation model, but specifying the population parameters is typically the most difficult aspect of a computer simulation. For example, a Monte Carlo power analysis for a regression model requires population values for all model parameters (i.e., the regression coefficients, correlations among predictors, and residual variance). This is not unique to Monte Carlo power simulations, and standard power analyses effectively require the same information expressed in the form of an effect size. For example, Cohen's (1988) approach converts the regression model parameters into an f^2 effect size metric. The population correlations that I used are convenient because they align with Cohen's small and medium effect size benchmarks, but deriving parameter values from published research studies or meta-analyses is a much better approach.

The next step of the simulation process is to generate a large number of samples from the specified population model. For example, my previous simulations generated 5,000 samples of $N = 300$ cases each. Software packages with built-in Monte Carlo routines typically require only a couple of key words or commands to specify the number of samples and the size of each sample. Simulating missing values can be a difficult aspect of a power simulation. Some software packages have built-in routines for generating missing data, whereas others do not. Again, structural equation modeling packages are particularly useful because some programs offer a number of options for simulating missing data. The availability of such a routine may be a factor to consider when choosing a software package.

The next step of the simulation is to estimate a statistical model on each artificial data set. In my previous power simulations, I used maximum likelihood missing data handling to estimate the correlation matrix for each of the 5,000 samples. As you will see in Chapter 4, maximum likelihood estimation is very easy to implement and typically requires only a single additional key word or line of code. Maximum likelihood missing data handling is implemented in virtually every structural equation modeling program, and I rely heavily on these packages throughout the book.

Describing the empirical sampling distribution of the parameter estimates is the final step of a computer simulation. For the purpose of a power analysis, you would always generate the data from a population where the null hypothesis is false (e.g., the population correlation is nonzero). Consequently, power is the proportion of samples that produce a statistically significant parameter estimate. Programs that have built-in Monte Carlo facilities often report the proportion of significant replications as part of their standard output, so obtaining the power estimates often requires no additional programming.

Using Monte Carlo simulations to estimate power sounds tedious, but software packages tend to automate the process. Generating the power estimates in Table 1.8 was actually quite easy and took just a few lines of code. Specifying reasonable values for the population parameters is by far the most time-consuming part of the process. Once you write the program, the software package automatically generates the data, estimates the model, and summarizes the simulation results. For many common statistical models, this entire process takes just a few minutes to complete.

As an aside, you can also use standard analysis techniques to estimate the power for planned missingness designs (Graham et al., 2006, p. 340), but this is a less accurate approach. As an illustration, reconsider the three-form design in Table 1.5. Suppose that you were considering a total sample size of $N = 300$ and wanted to estimate power for the correlation between Q_3 and Q_5 (an A variable and a B variable). This portion of the design has 66% missing data, so you could simply use $N = 100$ to estimate power. The power of a two-tailed significance test with $\alpha = .05$ and $\rho = .30$ is approximately .86 (Cohen, 1988, p. 93). Standard power analyses do not account for the fact that maximum likelihood estimation borrows strength from other analysis variables, so they underestimate the true power (e.g., the Monte Carlo power estimate in Table 1.8 is slightly higher at .90). Nevertheless, standard power analysis methods are a viable option for generating conservative power estimates.

1.14 DATA ANALYSIS EXAMPLE

This section presents a data analysis example that illustrates how to use MCAR tests to identify potential correlates of missingness.* The analyses use artificial data from a questionnaire on eating disorder risk. Briefly, the data contain the responses from 400 college-aged women on 10 questions from the Eating Attitudes Test (EAT; Garner, Olmsted, Bohr, & Garfinkel, 1982), a widely used measure of eating disorder risk. The 10 questions measure two constructs, Drive for Thinness (e.g., "I avoid eating when I'm hungry") and Food Preoccupation (e.g., "I find myself preoccupied with food"), and mimic the two-factor structure proposed by Doninger, Enders, and Burnett (2005). Figure 4.3 shows a graphic of the EAT factor structure and abbreviated descriptions of the item stems. The data set also contains an anxiety scale score, a variable that measures beliefs about Western standards of beauty (e.g., high scores indicate that respondents internalize a thin ideal of beauty), and body mass index (BMI) values.

Variables in the EAT data set are missing for a variety of reasons. I simulated MCAR data by randomly deleting scores from the anxiety variable, the Western standards of beauty scale, and two of the EAT questions (EAT_2 and EAT_{21}). It seems reasonable to expect a relationship between body weight and missingness, so I created MAR data on five variables (EAT_1, EAT_{10}, EAT_{12}, EAT_{18}, and EAT_{24}) by deleting the EAT scores for a subset of cases in both tails of the BMI distribution. These same EAT questions were also missing for individuals with elevated anxiety scores. Finally, I introduced a small amount of MNAR data by deleting a number of the high body mass index scores (e.g., to mimic a situation where females with high BMI values refuse to be weighed). The deletion process typically produced a missing data rate of 5 to 10% on each variable.

I began the analysis by computing Little's (1988) MCAR test. The test was statistically significant, $\chi^2(489) = 643.32$, $p < .001$, which indicates that the missing data patterns produced mean differences that are inconsistent with the MCAR mechanism. This is an appropriate conclusion given that a number of variables in the data set are either MAR or MNAR. Little's procedure is essentially an omnibus test that evaluates whether all of the missing data patterns in a data set are mutually consistent with the MCAR mechanism. Consequently, the test is not particularly useful for identifying individual variables that are potential correlates of missingness.

*Analysis syntax and data are available on the companion website, *www.appliedmissingdata.com*.

A more focused approach for testing MCAR is to classify individuals as observed or missing on a particular variable and then test for group mean differences on other measured variables (Dixon, 1988). To illustrate, I created a missing data indicator for each of the seven incomplete EAT questionnaire items, such that $r = 1$ if an individual's score was observed and $r = 0$ if the value was missing. I then used each indicator as the grouping variable in a series of independent *t* tests that compared the means of the remaining variables. Table 1.13 shows the *t* statistics and the standardized mean difference values for these comparisons. The table lists the grouping variables (i.e., the missing data indicators) in the rows and uses bold typeface to denote the *t* statistics that exceed an approximate critical value of plus or minus two. I computed the standardized mean difference values by dividing the raw mean difference by the maximum likelihood estimate of the standard deviation. Cohen (1988) suggested values of .20, .50, and .80 as thresholds for a small, medium, and large standardized mean difference, respectively.

Table 1.13 illustrates several important points. To begin, 20 of the 91 *t* statistics are statistically significant, and several others are very nearly so. You would expect a collection of tests this large to produce about five type I errors, so the sheer number of significant comparisons provides compelling evidence that the EAT variables are not MCAR. Again, this is an appropriate conclusion given that five of the questionnaire items are MAR. Although the *t* tests correctly rule out the MCAR mechanism, they do a poor job of identifying the cause of missing data. For example, notice that several pairs of EAT variables produced significant *t* tests. In reality, the probability of missing data is solely a function of body mass index and anxiety, so these results are a spurious by-product of the mutual associations among the variables. Finally, notice that the *t* tests fail to identify body mass index as a cause of missingness on the five EAT variables with MAR data. Deleting the EAT scores for cases in both tails of body mass index distribution produced missing data groups with roughly equal BMI means. It is therefore not surprising that the *t* tests fail to identify the relationship between body mass index and missingness. Any test that evaluates homogeneity of means would fail to detect BMI as a correlate of missingness, so this underscores the fact that these procedures are not definitive tests of MCAR.

The primary benefit of performing MCAR tests is to identify potential correlates of missingness (i.e., auxiliary variables) that you can subsequently incorporate into the missing data handling procedure. The *t* tests are useful in this regard because they identify specific variables that are not MCAR. To illustrate, suppose that the ultimate analysis goal is to fit a confirmatory factor analysis model to the EAT questionnaire data. The MAR assumption is automatically satisfied if missingness on an EAT variable is related to another questionnaire item in the factor model. Consequently, you can ignore any *t* test that has an EAT question as the outcome because these correlates of missingness are already in the analysis. The bigger concern is whether probability of missing data relates to variables outside of the analysis model because excluding these correlates of missingness violates the MAR assumption and can produce biased parameter estimates. For example, the *t* test results in the three rightmost columns of Table 1.13 suggest that body mass index, anxiety, and beliefs about Western standards of beauty are potential correlates of missingness because each variable is significantly related to at least one of the EAT indicators. In truth, the Western standards of beauty variable is unrelated to missingness, but Collins et al. (2001) showed that mistakenly using an auxiliary variable that is unrelated to missingness has no negative impact on the subsequent

TABLE 1.13. Comparison of Missing and Complete Cases from the Data Analysis Example

		Comparison variable												
Indicator	Statistic	EAT_1	EAT_2	EAT_{10}	EAT_{11}	EAT_{12}	EAT_{14}	EAT_{24}	EAT_3	EAT_{18}	EAT_{21}	*BMI*	*WSB*	*ANX*
EAT_1	*t*	—	–1.17	–0.28	–0.35	–1.64	–1.93	–1.31	**–2.13**	**–2.50**	–1.56	–0.09	–0.49	**–2.38**
	d	—	–0.29	–0.07	–0.07	–0.43	–0.39	–0.4	**–0.47**	**–0.58**	–0.34	–0.02	–0.11	**–0.53**
EAT_2	*t*	1.28	—	0.22	0.51	0.02	1.14	0.08	1.49	0.74	1.15	–0.66	0.11	0.24
	d	0.31	—	0.04	0.09	0.00	0.16	0.01	0.27	0.14	0.20	–0.15	0.03	0.05
EAT_{10}	*t*	**–2.32**	–1.27	—	–1.19	0.98	–1.40	–0.63	–1.37	–1.79	–1.87	–1.06	–0.92	**–3.66**
	d	**–0.57**	–0.21	—	–0.22	0.20	–0.25	–0.14	–0.25	–0.39	–0.36	–0.22	–0.20	**–0.47**
EAT_{12}	*t*	**–2.38**	–1.62	–0.88	–1.97	—	–1.84	–1.55	–1.98	–1.32	–1.57	0.80	0.06	**–5.39**
	d	**–0.47**	–0.28	–0.18	–0.42	—	–0.33	–0.25	–0.43	–0.28	–0.32	0.16	0.01	**–0.85**
EAT_{24}	*t*	**–2.41**	**–3.19**	–1.96	**–3.03**	–1.95	**–2.28**	—	–1.37	–1.20	–0.60	–0.89	**–2.03**	**–4.16**
	d	**–0.42**	**–0.71**	–0.38	**–0.60**	–0.39	**–0.46**	—	–0.30	–0.31	–0.12	–0.19	**–0.33**	**–0.70**
EAT_{18}	*t*	–1.45	**–2.89**	**–2.02**	**–2.48**	–0.71	–1.62	**–2.73**	–1.98	—	–1.76	–1.69	**–3.03**	**–4.65**
	d	–0.28	**–0.60**	**–0.39**	**–0.47**	–0.10	–0.31	**–0.54**	–0.37	—	–0.35	–0.34	**–0.51**	**–0.79**
EAT_{21}	*t*	1.77	0.38	0.20	–0.72	0.72	1.31	–1.09	0.30	0.04	—	**–2.10**	–0.33	–0.01
	d	0.31	0.10	0.06	–0.19	0.18	0.29	–0.30	0.08	0.01	—	**–0.53**	–0.09	0.00

Note. *d* = standardized mean difference; *BMI* = body mass index; *WSB* = Western standards of beauty; *ANX* = anxiety. Positive values of *d* indicate that the observed cases had a higher mean, and negative *d* values indicate that the missing cases had a higher mean. **Bold** typeface denotes statistically significant comparisons.

analysis results. This suggests that you can be liberal when using the *t* tests to identify potential correlates of missingness because there is ultimately no harm in committing a type I error. However, my experience suggests that there is little benefit to using a large number of auxiliary variables. Consequently, you may want to identify a small set of variables that produce the largest standardized mean difference values.

1.15 SUMMARY

This chapter described some of the fundamental concepts that you will encounter repeatedly throughout the book. In particular, the first half of the chapter outlined missing data theory. Rubin (1976) and colleagues (Little & Rubin, 2002) introduced a classification system for missing data problems that is widely used in the literature today. This work has generated three so-called missing data mechanisms that describe how the probability of a missing value relates to the data, if at all. First, data are MAR when the probability of missing data on a variable *Y* is related to some other measured variable (or variables) but not to the values of *Y* itself. Second, the MCAR mechanism is stricter because it requires that the probability of missing data on a variable *Y* is unrelated to other measured variables and to the values of *Y* itself (i.e., the observed scores are a random sample of the hypothetically complete data set). Finally, the data are MNAR when the probability of missing data on a variable *Y* is related to the values of *Y* itself, even after controlling for other variables.

Rubin's missing data mechanisms are important because they essentially operate as assumptions that govern the performance of different missing data handling methods. For example, most of the ad hoc missing data techniques that researchers have been using for decades (e.g., discarding cases with incomplete data) require MCAR data. In contrast, the two state-of-the-art techniques—maximum likelihood estimation and multiple imputation—require the less stringent MAR assumption. Rubin's mechanisms are of great practical importance because all missing data techniques produce biased parameter estimates when their requisite assumptions do not hold.

The second half of the chapter introduced the idea of planned missing data. Researchers have proposed a number of designs that produce MCAR or MAR data as an intentional by-product of data collection. These so-called planned missingness designs use benign missing data to solve a number of practical problems. Among other things, planned missing data can reduce respondent burden in questionnaire designs, lower the cost associated with data collection, and diminish the data collection burden in longitudinal designs. Maximum likelihood and multiple imputation allow researchers to analyze data from planned missingness designs without having to discard the incomplete cases, and the power loss from the missing data is generally not proportional to the missing data rate. Planned missing data designs are highly useful and underutilized tools that will undoubtedly increase in popularity in the future.

Having established the basic theory behind missing data analyses, in the next chapter I describe a number of traditional missing data techniques that are still common in published research articles. These approaches typically assume an MCAR mechanism and yield biased parameter estimates with MAR and MNAR data. The methods in Chapter 2 have increasingly fallen out of favor in recent years, but the widespread availability and use of these techniques make it important to understand when and why they fail.

1.16 RECOMMENDED READINGS

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods, 11,* 323–343.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

Guilford Publications, 72 Spring Street, New York, NY 10012, 212-431-9800. www.guilford.com/p/enders