Psychological Methods
2002, Vol. 7, No. 2, 147–177

1082-989X/02/$5.00 DOI: 10.1037//1082-989X.7.2.147

Missing Data: Our View of the State of the Art

Joseph L. Schafer and John W. Graham
Pennsylvania State University

Statistical procedures for missing data have vastly improved, yet misconception and unsound practice still abound. The authors frame the missing-data problem, review methods, offer advice, and raise issues that remain unresolved. They clear up common misunderstandings regarding the *missing at random* (MAR) concept. They summarize the evidence against older procedures and, with few exceptions, discourage their use. They present, in both technical and practical language, 2 general approaches that come highly recommended: maximum likelihood (ML) and Bayesian multiple imputation (MI). Newer developments are discussed, including some for dealing with missing data that are not MAR. Although not yet in the mainstream, these procedures may eventually extend the ML and MI methods that currently represent the state of the art.

Why do missing data create such difficulty in scientific research? Because most data analysis procedures were not designed for them. Missingness is usually a nuisance, not the main focus of inquiry, but handling it in a principled manner raises conceptual difficulties and computational challenges. Lacking resources or even a theoretical framework, researchers, methodologists, and software developers resort to editing the data to lend an appearance of completeness. Unfortunately, ad hoc edits may do more harm than good, producing answers that are biased, inefficient (lacking in power), and unreliable.

Purposes of This Article

This article's intended audience and purposes are varied. For the novice, we review the available methods and describe their strengths and limitations. For those already familiar with missing-data issues, we seek to fill gaps in understanding and highlight recent developments in this rapidly changing field. For methodologists, we hope to stimulate thoughtful discussion and point to important areas for new research. One of our main reasons for writing this article is to familiarize researchers with these newer techniques and encourage them to apply these methods in their own work.

In the remainder of this article, we describe criteria by which missing-data procedures should be evaluated. Fundamental concepts, such as the distribution of missingness and the notion of *missing at random* (MAR), are presented in nontechnical fashion. Older procedures, including case deletion and single imputation, are reviewed and assessed. We then review and compare modern procedures of maximum likelihood (ML) and multiple imputation (MI), describing their strengths and limitations. Finally, we describe new techniques that attempt to relax distributional assumptions and methods that do not assume that missing data are MAR.

In general, we emphasize and recommend two approaches. The first is ML estimation based on all available data; the second is Bayesian MI. Readers who are not yet familiar with these techniques may wish to see step-by-step illustrations of how to apply them to real data. Such examples have already been published, and space limitations do not allow us to repeat them here. Rather, we focus on the underlying motivation and principles and provide references so that interested readers may learn the specifics of applying them later. Many software products (both free and commercial) that implement ML and MI are listed here, but in this rapidly changing field others will

Joseph L. Schafer, Department of Statistics and the Methodology Center, Pennsylvania State University; John W. Graham, Department of Biobehavioral Health, Pennsylvania State University.

This research was supported by National Institute on Drug Abuse Grant 1-P50-DA10075.

Correspondence concerning this article should be addressed to Joseph L. Schafer, The Methodology Center, Pennsylvania State University, Henderson S-159, University Park, Pennsylvania 16802. E-mail: jls@stat.psu.edu

undoubtedly become available soon. A resource Web page maintained by John W. Graham (http://methodology.psu.edu/resources.html) will provide timely updates on missing-data applications and utilities as they evolve, along with step-by-step instructions on MI with NORM (Schafer, 1999b).

Fundamentals

What Is a Missing Value?

Data contain various codes to indicate lack of response: "Don't know," "Refused," "Unintelligible," and so on. Before applying a missing-data procedure, one should consider whether an underlying "true" value exists and, if so, whether that value is unknown. The answers may not be obvious. Consider a questionnaire for adolescents with a section on marijuana use. The first item is "Have you ever tried marijuana?" If the response is "No," the participant is directed to skip items on recent and lifetime use and proceed to the next section. Many researchers would not consider the skipped items to be missing, because never having used marijuana logically implies no recent or lifetime use. In the presence of response error, however, it might be a mistake to presume that all skipped items are zero, because some answers to the initial question may be incorrect.

Interesting issues arise in longitudinal studies in which unfortunate events preclude measurement. If participants die, should we consider their characteristics (e.g., mental functioning) at subsequent occasions to be missing? Some might balk at the idea, but in some contexts it is quite reasonable. If deaths are rare and largely unrelated to the phenomena of interest, then we may want to estimate parameters for an ideal scenario in which no one dies during the study. At other times, we may want to describe the characteristics of live participants only and perhaps perform additional analyses with mortality itself as the outcome. Even then, however, it is sometimes convenient to posit the existence of missing values for the deceased purely as a computational device, to permit the use of missing-data algorithms. These issues are clarified later, after we review the notion of MAR.

Missing values are part of the more general concept of *coarsened data,* which includes numbers that have been grouped, aggregated, rounded, censored, or truncated, resulting in partial loss of information (Heitjan & Rubin, 1991). Latent variables, a concept familiar to psychologists, are also closely related to missing data. Latent variables are unobservable quantities (e.g., intelligence, assertiveness) that are only imperfectly measured by test or questionnaire items. Computational methods for missing data may simplify parameter estimation in latent-variable models; a good example is the expectation-maximization (EM) algorithm for latent class analysis (Clogg & Goodman, 1984).

Psychologists have sometimes made a distinction between missing values on independent variables (predictors) and missing values on dependent variables (outcomes). From our perspective, these two do not fundamentally differ. It is true that, under certain assumptions, missing values on a dependent variable may be efficiently handled by a very simple method such as case deletion, whereas good missing-data procedures for independent variables can be more difficult to implement. We discuss these matters later as we review specific classes of missing-data procedures. However, we caution our readers not to believe general statements such as, "Missing values on a dependent variable can be safely ignored," because such statements are imprecise and generally false.

Historical Development

Until the 1970s, missing values were handled primarily by editing. Rubin (1976) developed a framework of inference from incomplete data that remains in use today. The formulation of the EM algorithm (Dempster, Laird, & Rubin, 1977) made it feasible to compute ML estimates in many missing-data problems. Rather than deleting or filling in incomplete cases, ML treats the missing data as random variables to be removed from (i.e., integrated out of) the likelihood function as if they were never sampled. We elaborate on this point later after introducing the notion of MAR. Many examples of EM were described by Little and Rubin (1987). Their book also documented the shortcomings of case deletion and single imputation, arguing for explicit models over informal procedures. About the same time, Rubin (1987) introduced the idea of MI, in which each missing value is replaced with $m > 1$ simulated values prior to analysis. Creation of MIs was facilitated by computer technology and new methods for Bayesian simulation discovered in the late 1980s (Schafer, 1997). ML and MI are now becoming standard because of implementations in free and commercial software.

The 1990s have seen many new developments. Reweighting, long used by survey methodologists, has been proposed for handling missing values in regression models with missing covariates (Ibrahim, 1990).

New lines of research focus on how to handle missing values while avoiding the specification of a full parametric model for the population (Robins, Rotnitzky, & Zhao, 1994). New methods for nonignorable modeling, in which the probabilities of nonresponse are allowed to depend on the missing values themselves, are proliferating in biostatistics and public health. The primary focus of these nonignorable models is dropout in clinical trials, in which participants may be leaving the study for reasons closely related to the outcomes being measured (Little, 1995). Researchers are now beginning to assess the sensitivity of results to alternative hypotheses about the distribution of missingness (Verbeke & Molenberghs, 2000).

Goals and Criteria

With or without missing data, the goal of a statistical procedure should be to make valid and efficient inferences about a population of interest—not to estimate, predict, or recover missing observations nor to obtain the same results that we would have seen with complete data. Attempts to recover missing values may impair inference. For example, the common practice of *mean substitution*—replacing each missing value for a variable with the average of the observed values—may accurately predict missing data but distort estimated variances and correlations. A missing-value treatment cannot be properly evaluated apart from the modeling, estimation, or testing procedure in which it is embedded.

Basic criteria for evaluating statistical procedures have been established by Neyman and Pearson (1933) and Neyman (1937). Let Q denote a generic population quantity to be estimated, and let $\hat{Q}$ denote an estimate of Q based on a sample of data. If the sample contains missing values, then the method for handling them should be considered part of the overall procedure for calculating $\hat{Q}$. If the procedure works well, then $\hat{Q}$ will be close to Q, both on average over repeated samples and for any particular sample. That is, we want the *bias*—the difference between the average value of $\hat{Q}$ and the true Q—to be small, and we also want the variance or standard deviation of $\hat{Q}$ to be small. Bias and variance are often combined into a single measure called mean square error, which is the average value of the squared distance $(\hat{Q} - Q)^2$ over repeated samples. The mean square error is equal to the squared bias plus the variance.

Bias, variance, and mean square error describe the behavior of an estimate, but we also want honesty in the measures of uncertainty that we report. A reported standard error $SE(\hat{Q})$ should be close to the true standard deviation of $\hat{Q}$. A procedure for confidence intervals—for example, $\hat{Q} \pm 2SE(\hat{Q})$ for a 95% interval—should cover the true Q with probability close to the nominal rate. If the coverage rate is accurate, the probability of Type I error (wrongly rejecting a true null hypothesis) will also be accurate. Subject to correct coverage, we also want the intervals to be narrow, because shorter intervals will reduce the rate of Type II error (failure to accept a true alternative hypothesis) and increase power.

When missing values occur for reasons beyond our control, we must make assumptions about the processes that create them. These assumptions are usually untestable. Good science suggests that assumptions be made explicit and the sensitivity of results to departures be investigated. One hopes that similar conclusions will follow from a variety of realistic alternative assumptions; when that does not happen, the sensitivity should be reported.

Finally, one should avoid tricks that apparently solve the missing-data problem but actually redefine the parameters or the population. For example, consider a linear regression of Y on X, where the predictor X is sometimes missing. Suppose we replace the missing values by an arbitrary number (say, zero) and introduce a dummy indicator Z that is one if X is missing and zero if X is observed. This procedure merely redefines the coefficients. In the original model $E(Y) = \beta_0 + \beta_1 X$, where E represents expected value, β_0 and β_1 represent the intercept and slope for the full population; in the expanded model $E(Y) = \beta_0 + \beta_1 X + \beta_2 Z$, β_0 and β_1 represent the intercept and slope for respondents, and $\beta_0 + \beta_2$ represents the mean of Y among nonrespondents. For another example, suppose that missing values occur on a nominal outcome with response categories $1, 2, \ldots, k$. One could treat the missing value as category $k + 1$, but that merely redefines categories $1, \ldots, k$ to apply only to respondents.

Types and Patterns of Nonresponse

Survey methodologists have historically distinguished *unit nonresponse*, which occurs when the entire data collection procedure fails (because the sampled person is not at home, refuses to participate, etc.), from *item nonresponse*, which means that partial data are available (i.e., the person participates but does not respond to certain individual items). Survey statisticians have traditionally handled unit nonresponse by reweighting and item nonresponse by

single imputation. These older methods, which are reviewed in this article, may perform reasonably well in some situations, but more modern procedures (e.g., ML and MI) exhibit good behavior more generally.

In longitudinal studies, participants may be present for some waves of data collection and missing for others. This kind of missingness may be called *wave nonresponse. Attrition,* or *dropout,* which is a special case of wave nonresponse, occurs when one leaves the study and does not return. Overall, dropout or attrition may be the most common type of wave nonresponse. However, it is not uncommon for participants to be absent from one wave and subsequently reappear. Because repeated measurements on an individual tend to be correlated, we recommend procedures that use all the available data for each participant, because missing information can then be partially recovered from earlier or later waves. Longitudinal modeling by ML can be a highly efficient way to use the available data. MI of missing responses is also effective if we impute under a longitudinal model that borrows information across waves.

Many data sets can be arranged in a rectangular or matrix form, where the rows correspond to observational units or participants and the columns correspond to items or variables. With rectangular data, there are several important classes of overall missing-data patterns. Consider Figure 1a, in which missing values occur on an item Y but a set of p other items $X_1, \ldots, X_p$ is completely observed; we call this a *univariate pattern.* The univariate pattern is also meant to include situations in which Y represents a group of items that is either entirely observed or entirely missing for each unit. In Figure 1b, items or item groups $Y_1, \ldots, Y_p$ may be ordered in such a way that if Y_j is missing for a unit, then $Y_{j+1}, \ldots, Y_p$ are missing as well; this is called a *monotone pattern.* Monotone patterns may arise in longitudinal studies with attrition, with Y_j representing variables collected at the jth occasion. Figure 1c shows an *arbitrary pattern* in which any set of variables may be missing for any unit.

The Distribution of Missingness

For any data set, one can define indicator variables R that identify what is known and what is missing. We refer to R as the *missingness.* The form of the missingness depends on the complexity of the pattern. In Figure 1a, R can be a single binary item for each unit indicating whether Y is observed ($R = 1$) or missing ($R = 0$). In Figure 1b, R can be a integer variable ($1, 2, \ldots, p$) indicating the highest j for which Y_j is observed. In Figure 1c, R can be a matrix of binary indicators of the same dimension as the data matrix, with elements of R set to 1 or 0 according to whether the corresponding data values are observed or missing.

In modern missing-data procedures missingness is regarded as a probabilistic phenomenon (Rubin, 1976). We treat R as a set of random variables having a joint probability distribution. We may not have to specify a particular distribution for R, but we must agree that it has a distribution. In statistical literature, the distribution of R is sometimes called the *response mechanism* or *missingness mechanism,* which may be confusing because *mechanism* suggests a real-world process by which some data are recorded and others are missed. To describe accurately all potential causes or reasons for missingness is not realistic. The distribution of R is best regarded as a mathematical device to describe the rates and patterns of missing values and to capture roughly possible relationships between the missingness and the values of the missing items themselves. To avoid suggestions of causality, we

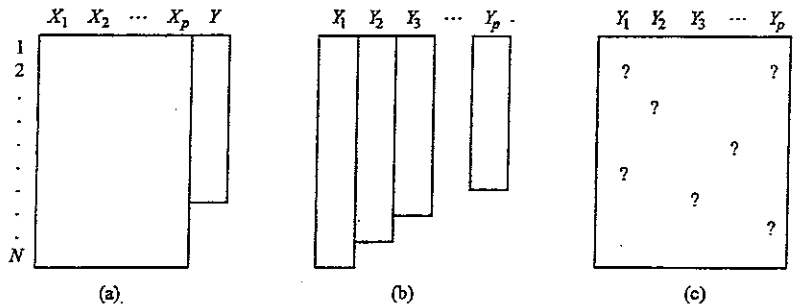

Figure 1. Patterns of nonresponse in rectangular data sets: (a) univariate pattern, (b) monotone pattern, and (c) arbitrary pattern. In each case, rows correspond to observational units and columns correspond to variables.

therefore refer to the probability distribution for R as *the distribution of missingness* or *the probabilities of missingness.*

Missing at Random

Because missingness may be related to the data, we classify distributions for R according to the nature of that relationship. Rubin (1976) developed a typology for these distributions that is widely cited but less widely understood. Adopting a generic notation, let us denote the complete data as Y_{com} and partition it as $Y_{com} = (Y_{obs}, Y_{mis})$, where Y_{obs} and Y_{mis} are the observed and missing parts, respectively. Rubin (1976) defined missing data to be MAR if the distribution of missingness does not depend on Y_{mis},

$$P(R|Y_{com}) = P(R|Y_{obs}). \quad (1)$$

In other words, MAR allows the probabilities of missingness to depend on observed data but not on missing data.[1] An important special case of MAR, called *missing completely at random* (MCAR), occurs when the distribution does not depend on Y_{obs} either,

$$P(R|Y_{com}) = P(R).$$

When Equation 1 is violated and the distribution depends on Y_{mis}, the missing data are said to be *missing not at random* (MNAR). MAR is also called *ignorable nonresponse,* and MNAR is called *nonignorable.*

For intuition, it helps to relate these definitions to the patterns in Figure 1. Consider the univariate pattern of Figure 1a, where variables $X = (X_1, \ldots, X_p)$ are known for all participants but Y is missing for some. If participants are independently sampled from the population, then MCAR, MAR, and MNAR have simple interpretations in terms of X and Y: MCAR means that the probability that Y is missing for a participant does not depend on his or her own values of X or Y (and, by independence, does not depend on the X or Y of other participants either), MAR means that the probability that Y is missing may depend on X but not Y, and MNAR means that the probability of missingness depends on Y. Notice that under MAR, there could be a relationship between missingness and Y induced by their mutual relationships to X, but there must be no residual relationship between them once X is taken into account. Under MNAR, some residual dependence between missingness and Y remains after accounting for X.

Notice that Rubin's (1976) definitions describe statistical relationships between the data and the missingness, not causal relationships. Because we often consider real-world reasons why data become missing, let us imagine that one could code all the myriad reasons for missingness into a set of variables. This set might include variables that explain why some participants were physically unable to show up (age, health status), variables that explain the tendency to say "I don't know" or "I'm not sure" (cognitive functioning), variables that explain outright refusal (concerns about privacy), and so on. These causes of missingness are not likely to be present in the data set, but some of them are possibly related to X and Y and thus by omission may induce relationships between X or Y and R. Other causes may be entirely unrelated to X and Y and may be viewed as external noise. If we let Z denote the component of cause that is unrelated to X and Y, then MCAR, MAR, and MNAR may be represented by the graphical relationships in Figure 2. MCAR requires that the causes of missingness be entirely contained within the unrelated part Z, MAR allows some causes to be related to X, and MNAR requires some causes to be residually related to Y after relationships between X and R are taken into account.

If we move from the univariate pattern of Figure 1a to the monotone pattern of Figure 1b, MCAR means that Y_j is missing with probability unrelated to any variables in the system; MAR means that it may be related only to $Y_1, \ldots, Y_{j-1}$; and MNAR means that it is related to $Y_j, \ldots, Y_p$. If $Y_1, \ldots, Y_p$ are repeated measurements of an outcome variable in a longitudinal study, and missing data arise only from attrition, then MCAR requires dropout to be independent of responses at every occasion, MAR allows dropout to depend on responses at any or all occasions prior to dropout, and MNAR means that it depends on the unseen responses after the participant drops out. In this specialized setting, MAR has been called *noninformative* or *ignorable* dropout, whereas MNAR is called *informative* (Diggle & Kenward, 1994).

[1] In Rubin's (1976) definition, Equation 1 is not required to hold for all possible values of R, but only for the R that actually appeared in the sample. This technical point clarifies certain issues. For example, suppose that an experiment produced no missing values even though it could have. In that case, Equation 1 would hold because Y_{mis} is empty, and Rubin's (1976) results indicate that one should simply analyze the complete data without worrying about the fact that missing values could have arisen in hypothetical repetitions of the experiment.

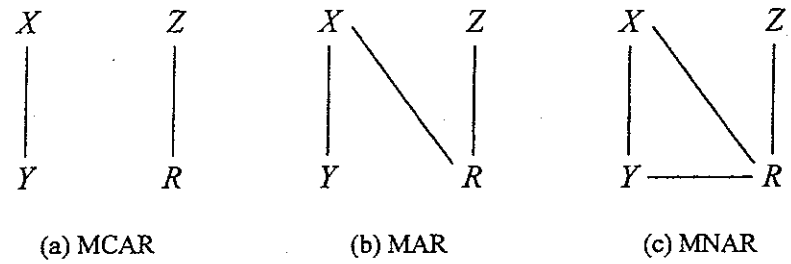

Figure 2. Graphical representations of (a) missing completely at random (MCAR), (b) missing at random (MAR), and (c) missing not at random (MNAR) in a univariate missing-data pattern. X represents variables that are completely observed, Y represents a variable that is partly missing, Z represents the component of the causes of missingness unrelated to X and Y, and R represents the missingness.

With the arbitrary pattern of Figure 1c, MCAR still requires independence between missingness and $Y_1, \ldots, Y_p$. However, MAR is now more difficult to grasp. MAR means that a participant's probabilities of response may be related only to his or her own set of observed items, a set that may change from one participant to another. One could argue that this assumption is odd or unnatural, and in many cases we are inclined to agree. However, the apparent awkwardness of MAR does not imply that it is far from true. Indeed, in many situations, we believe that MAR is quite plausible, and the analytic simplifications that result from making this assumption are highly beneficial.

In discussions with researchers, we have found that most misunderstandings of MCAR, MAR, and MNAR arise from common notions about the meaning of *random.* To a statistician, random suggests a process that is probabilistic rather than deterministic. In that sense, MCAR, MAR, and MNAR are all random, because they all posit probability distributions for R. To a psychologist, random may suggest a process that is unpredictable and extraneous to variables in the current study (e.g., tossing a coin or rolling a die), a notion that agrees more closely with MCAR than with MAR. In retrospect, Rubin's (1976) choice of terminology seems a bit unfortunate, but these terms are now firmly established in the statistical literature and are unlikely to change.

The Plausibility of MAR

In certain settings, MAR is known to hold. These include *planned missingness* in which the missing data were never intended to be collected in the first place: cohort-sequential designs for longitudinal studies (McArdle & Hamagami, 1991; Nesselroade & Baltes, 1979) and the use of multiple questionnaire forms containing different subsets of items (Graham, Hofer, & Piccinin, 1994; Graham, Hofer, & MacKinnon, 1996). Planned missingness in a study may have important advantages in terms of efficiency and cost (Graham, Taylor, & Cumsille, 2001). Planned missing values are usually MCAR, but MAR situations sometimes arise—for example, if participants are included in a follow-up measure only if their pretest scores exceed a cutoff value. Latent variables are missing with probability one and are therefore also known to be MAR.

When missingness is beyond the researcher's control, its distribution is unknown and MAR is only an assumption. In general, there is no way to test whether MAR holds in a data set, except by obtaining follow-up data from nonrespondents (Glynn, Laird, & Rubin, 1993; Graham & Donaldson, 1993) or by imposing an unverifiable model (Little & Rubin, 1987, chapter 11). In most cases we should expect departures from MAR, but whether these departures are serious enough to cause the performance of MAR-based methods to be seriously degraded is another issue entirely (Graham, Hofer, Donaldson, MacKinnon, & Schafer, 1997). Recently, Collins, Schafer, and Kam (2001) demonstrated that in many realistic cases, an erroneous assumption of MAR (e.g., failing to take into account a cause or correlate of missingness) may often have only a minor impact on estimates and standard errors.

Example

Suppose that systolic blood pressures of N participants are recorded in January (X). Some of them have a second reading in February (Y), but others do not. Table 1 shows simulated data for $N = 30$ participants drawn from a bivariate normal population with means $\mu_x = \mu_y = 125$, standard deviations $\sigma_x = \sigma_y = 25$, and correlation $\rho = .60$. The first two columns of the

Table 1
Simulated Blood Pressure Measurements (N = 30 Participants) in January (X) and February (Y) With Missing Values Imposed by Three Different Methods

	Y			
X	Complete	MCAR	MAR	MNAR
	Data for individual participants			
169	148	148	148	148
126	123	—	—	—
132	149	—	—	149
160	169	—	169	169
105	138	—	—	—
116	102	—	—	—
125	88	—	—	—
112	100	—	—	—
133	150	—	—	150
94	113	—	—	—
109	96	—	—	—
109	78	—	—	—
106	148	—	—	148
176	137	—	137	—
128	155	—	—	155
131	131	—	—	—
130	101	101	—	—
145	155	—	155	155
136	140	—	—	—
146	134	—	134	—
111	129	—	—	—
97	85	85	—	—
134	124	124	—	—
153	112	—	112	—
118	118	—	—	—
137	122	122	—	—
101	119	—	—	—
103	106	106	—	—
78	74	74	—	—
151	113	—	113	—
	Summary data: Mean (with standard deviation in parentheses)			
125.7 (23.0)	121.9 (24.7)	108.6 (25.1)	138.3 (21.1)	153.4 (7.5)

Note. Dashes indicate missing values. MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random.

table show the complete data for X and Y. The other columns show the values of Y that remain after imposing missingness by three methods. In the first method, the 7 measured in February were randomly selected from those measured in January; this mechanism is MCAR. In the second method, those who returned in February did so because their January measurements exceeded 140 ($X > 140$), a level used for diagnosing hypertension; this is MAR but not MCAR. In the third method, those recorded in February were those whose February measurements exceeded 140 ($Y > 140$). This could happen, for example, if all individuals returned in February, but the staff person in charge decided to record the February value only if it was in the hypertensive range. This third mechanism is an example of MNAR. (Other MNAR mechanisms are possible; e.g., the February measurement may be recorded only if it is substantially different from the January reading.) Notice that as we move from MCAR to MAR to MNAR, the observed Y values become an increasingly select and unusual group relative to the population; the sample mean increases, and the standard deviation decreases. This phenomenon is not a universal feature of MCAR, MAR, and MNAR, but it does happen in many realistic examples.

We frequently return to this example throughout this article to illustrate the performance of various methods. Because the rate of missing values is high, the chosen method will exert a high degree of influence over the results, and differences among competing methods will be magnified. Effects will also be large because of the unusually strong nature of the MAR and MNAR processes. In psychological studies, one would rarely expect a datum to be missing if and only if its value exceeds a sharp cutoff. More commonly, one might expect a gradual increasing or decreasing or perhaps curvilinear relationship between X or Y and the probability of missingness. Also, in most cases, the reasons or causes of missingness are not X and Y themselves but external factors that are merely related to X and Y, perhaps not strongly. Differences in results due to varying missing-data treatments in this example should therefore be regarded as more extreme than what we usually see in practice.

We would ideally like to have a single procedure for estimating all the parameters (μ_X, μ_Y, σ_X, σ_Y, ρ) from the observed data that performs well over repeated samples regardless of how missingness is distributed. Technically speaking, this is not possible. We see below that some data editing methods perform well for some parameters under MCAR and occasionally MAR. Using likelihood or Bayesian procedures, one may achieve good performance for all parameters without knowing the distribution of missingness, provided that it is MAR. Under MNAR data the task becomes more difficult; one must specify a model for the missingness that is at least approximately correct,

and even then the performance may be poor unless the sample is very large. However, even though there is no single procedure that is technically correct in every situation, all is not lost. We do believe that good performance is often achievable through likelihood or Bayesian methods without specifically modeling the probabilities of missingness, because in many psychological research settings the departures from MAR are probably not serious.

Implications of MAR, MCAR, and MNAR

With a complete-data Y_{com}, statistical methods are usually motivated by an assumption that the data are randomly sampled from a population distribution $P(Y_{com}; \theta)$, where θ represents unknown parameters. To a statistician, $P(Y_{com}; \theta)$ has two very different interpretations. First, it is regarded as the repeated-sampling distribution for Y_{com}; it describes the probability of obtaining any specific data set among all the possible data sets that could arise over hypothetical repetitions of the sampling procedure and data collection. Second, it is regarded as a likelihood function for θ. In the likelihood interpretation, we substitute the realized value of Y_{com} into $P(Y_{com}; \theta)$, and the resulting function for θ summarizes the data's evidence about parameters. Certain classes of modern statistical procedures, including all ML methods, are motivated by viewing $P(Y_{com}; \theta)$ as a likelihood function. Other procedures, including nonparametric and semiparametric methods, are justified only from the repeated-sampling standpoint.

When portions of Y_{com} are missing, it is tempting to base all statistical procedures on $P(Y_{obs}; \theta)$, the distribution of the observed part only. This distribution is obtained by calculus as the definite integral of $P(Y_{com}; \theta)$ with respect to Y_{mis},

$$P(Y_{obs}; \theta) = \int P(Y_{com}; \theta)\, dY_{mis}. \qquad (2)$$

Details and examples of integrating probability distributions are given in texts on probability theory (e.g., Hogg & Tanis, 1997). In missing-data problems, however, it is not automatically true that Equation 2 is the correct sampling distribution for Y_{obs} and the correct likelihood for θ based on Y_{obs}. Rubin (1976) first identified the conditions under which it is a proper sampling distribution and a proper likelihood; interestingly, the conditions are not identical. For Equation 2 to be a correct sampling distribution, the missing data should be MCAR. For Equation 2 to be a correct likelihood, we need only MAR. The weaker conditions for the latter suggest that missing-data procedures based on likelihood principles are generally more useful than those derived from repeated-sampling arguments only. We believe that to be true. Many of the older data-editing procedures bear no relationship to likelihood and may be valid only under MCAR. Even when MCAR does hold, these methods may be inefficient. Methods motivated by treating Equation 2 as a likelihood tend to be more powerful and better suited to real-world applications in which MCAR is often violated.

Finally, we note that the attractive properties of likelihood carry over to the Bayesian method of MI, because in the Bayesian paradigm we combine a likelihood function with a prior distribution for the parameters. As the sample size grows, the likelihood dominates the prior, and Bayesian and likelihood answers become similar (Gelman, Rubin, Carlin, & Stern, 1995).

Missing Values That Are Not MAR

What happens when the missing data are not MAR? It is then not appropriate to use Equation 2 either as a sampling distribution or as a likelihood. From a likelihood standpoint, the correct way to proceed is to choose an explicit model for the missingness, $P(R|Y_{com}; \xi)$, where ξ denotes unknown parameters of the missingness distribution. For example, if missingness is confined to a single variable, then we may suppose that the binary indicators in R are described by a logistic regression on the variable in question, and ξ would consist of an intercept and slope. The joint model for the data and missingness becomes the product of $P(Y_{com}; \theta)$ and $P(R|Y_{com}; \xi)$. The correct likelihood function is then given by the integral of this product over the unseen missing values,

$$P(Y_{obs}, R; \theta, \xi) = \int P(Y_{com}; \theta)\, P(R|Y_{com}; \xi)\, dY_{mis}, \qquad (3)$$

where d is the calculus differential. The practical implication of MNAR is that the likelihood for θ now depends on an explicit model for R. In most cases, this missingness model is a nuisance; questions of substantive interest usually pertain to the distribution of Y_{com}, not the distribution of R. Nevertheless, under MNAR the model for R contributes information about θ, and the evidence about θ from Equation 3 may present a very different picture from that given by Equation 2. Likelihoods for MNAR models are often difficult to handle from a computational standpoint, but some interesting work in this area has been pub-

lished recently. Methods for MNAR data are reviewed at the end of this article.

Missing Values That Are Out of Scope

In addition to MAR, there is another situation in which Equation 2 is an appropriate likelihood: when the fact that an observation is missing causes it to leave the universe of interest. Consider a questionnaire with an item, "How well do you get along with your siblings?" Responses for some participants are missing because they have no siblings. Literally speaking, there are no missing data in this problem at all. However, if the intended analysis could be carried out more simply if the data were balanced (i.e., if responses to this item were available for each participant), then for computational reasons it may be worthwhile to write the likelihood $P(Y_{obs}; \theta)$ in the form of Equation 2, where Y_{obs} denotes responses for those who have siblings and Y_{mis} represents hypothetical responses for those who do not. In this case, the hypothetical missing data could be regarded as MAR and θ would be the parameters of the distribution of responses for the population of those with siblings. We need not worry about whether missingness depends on the characteristics of nonexistent siblings; these missing values are introduced merely as a mathematical device to simplify the computations.

To a certain extent, this discussion also applies to longitudinal studies in which some participants die. Outcome measures of physical or mental status (e.g., cognitive functioning) have meaning for live persons only. One who dies automatically leaves the universe of interest, so values that are "missing" because of death may often be regarded as MAR. However, if the measurements of these outcomes are spaced far apart in time—for example, if they are taken annually—then the death of a participant may provide indirect evidence of an unmeasured steep decline in the outcome prior to death, and response trajectories estimated under an MAR assumption may be somewhat optimistic. In those cases, joint modeling of the outcome and death events may be warranted (Hogan & Laird, 1997).

Older Methods

Case Deletion

Among older methods for missing data, the most popular is to discard units whose information is incomplete. *Case deletion,* also known commonly as *listwise deletion* (LD) and *complete-case analysis,* is used by default in many statistical programs, but details of its implementation vary. LD confines attention to units that have observed values for all variables under consideration. For example, suppose we are computing a sample covariance matrix for items $X_1, \ldots, X_p$. LD omits from consideration any case that has a missing value on any of the variables $X_1, \ldots, X_p$.

Available-case (AC) *analysis,* in contrast to LD, uses different sets of sample units for different parameters. For estimating covariances, this is sometimes called *pairwise deletion* or *pairwise inclusion.* For example, we may use every observed value of X_j to estimate the standard deviation of X_j, and every observed pair of values (X_j, X_k) to estimate the covariance of X_j and X_k. For the correlation between X_j and X_k, we might compute the sample correlation coefficient using the same set of units that we used to estimate the covariance. On the other hand, we could also divide our estimated covariance by the estimated standard deviations. The latter seems more efficient, but it could conceivably yield a correlation outside of the interval $[-1, 1]$, causing one or more eigenvalues to be negative. We believe that the underlying principle of AC analysis—to make use of all the available data—is eminently sensible, but deleting cases is a poor way to operationalize it. Another limitation of AC analysis is that, because parameters are estimated from different sets of units, it is difficult to compute standard errors or other measures of uncertainty; analytic methods are troublesome, and other procedures (e.g., bootstrapping) are at least as tedious as they would be for other estimates with better properties.

Properties of case deletion. Case deletion can be motivated by viewing Equation 2 as a sampling distribution for Y_{obs} and is generally valid only under MCAR. In a few circumstances, it produces inferences that are optimal under MAR. For example, under the univariate missingness pattern of Figure 1a, the parameters of the regression of Y on any subset of $X_1, \ldots, X_p$ can be estimated from the complete cases and the estimates are both valid and efficient under MAR (e.g., see Graham & Donaldson, 1993). However, this result does not extend to other measures of association between Y and X such as correlation coefficients, nor does it extend to parameters of the marginal distribution of Y. When the missing data are not MCAR, results from case deletion may be biased, because the complete cases can be unrepresentative of the full population. If the departures from MCAR are not serious, then the impact of this bias might be

unimportant, but in practice it can be difficult to judge how large the biases might be.

When MCAR holds, case deletion can still be inefficient. Consider again the univariate pattern in Figure 1a, and suppose that we want to estimate aspects of the marginal distribution of Y (e.g., the population mean). Furthermore, suppose that Y and the Xs are highly related, so that the missing values of Y can be predicted from X with near certainty. Case deletion bases estimates on the reduced sample of Y values, ignoring the strong predictive information contained in the Xs. Researchers become acutely aware of the inefficiency of case deletion in multivariate analyses involving many items, in which mild rates of missing values on each item may cause large portions of the sample to be discarded.

The main virtue of case deletion is simplicity. If a missing-data problem can be resolved by discarding only a small part of the sample, then the method can be quite effective. However, even in that situation, one should explore the data to make sure that the discarded cases are not unduly influential. In a study of a rare disease or condition, for example, one should verify that the small group being discarded does not contain a large proportion of the participants possessing that condition. For more discussion on the properties of case deletion and further references, see chapter 3 of Little and Rubin (1987).

Example. For the systolic blood pressure data shown in Table 1, LD removes all participants whose Y values are missing. To demonstrate the properties of LD over repeated samples, we performed a simulation experiment. One thousand samples were drawn from the bivariate normal population, and missing values were imposed on each sample by each of the three mechanisms. For MAR and MNAR, Y was made missing if $X \leq 140$ and $Y \leq 140$, respectively, producing an average rate of 73% missing observations for Y. For MCAR, each Y value was made missing with probability .73. The total number of participants in each sample was increased to 50 to ensure that, after case deletion, a sufficient number remained to support the estimation of correlation and regression coefficients. After deletion, standard techniques were used to calculate estimates and 95% intervals for five population parameters: the mean of Y ($\mu_Y = 125$), the standard deviation of Y ($\sigma_Y = 25$), the correlation coefficient ($\rho_{XY} = .60$), the slope for the regression of Y on X ($\beta_{Y|X} = .60$), and the slope for the regression of X on Y ($\beta_{X|Y} = .60$). For σ_Y, the confidence interval consisted of the square roots of the endpoints for the classical interval for the variance of a normal population. For ρ, the interval was calculated by applying Fisher's transformation $z = \tanh^{-1}(r)$ to the sample correlation r, adding and subtracting $1.96(N-3)^{-1/2}$, and applying the inverse transformation $r = \tanh(z)$ to the endpoints.

Results of the simulation are summarized in Table 2. The top panel of the table reports the average values of the parameter estimates under each mechanism. A discrepancy between this average and the true pa-

Table 2
Performance of Listwise Deletion for Parameter Estimates and Confidence Intervals Over 1,000 Samples (N = 50 Participants)

Parameter	MCAR	MAR	MNAR
Average parameter estimate (with RMSE in parentheses)			
$\mu_Y = 125.0$	125.0 (6.95)	**143.3** (19.3)	**155.5** (30.7)
$\sigma_Y = 25.0$	24.6 (5.26)	**20.9** (5.84)	**12.2** (13.2)
$\rho = .60$	.59 (.19)	**.33** (.37)	**.34** (.36)
$\beta_{Y\|X} = .60$	.61 (.27)	.60 (.51)	**.21** (.43)
$\beta_{X\|Y} = .60$	.60 (.25)	**.20** (.44)	.60 (.52)
Coverage (with average interval width in parentheses)			
μ_Y	94.3 (30.0)	**18.8** (25.0)	**0.0** (14.7)
σ_Y	94.3 (23.3)	90.7 (19.4)	**17.4** (11.4)
ρ	95.4 (0.76)	**82.5** (0.93)	**82.7** (0.94)
$\beta_{Y\|X}$	94.6 (1.10)	95.9 (2.20)	**40.0** (0.73)
$\beta_{X\|Y}$	95.3 (1.08)	**37.7** (0.71)	96.6 (2.23)

Note. Parameters: μ is the population mean; σ is the population standard deviation; ρ is the population correlation; β is the population regression slope. Coverage represents the percentage of confidence intervals that include the parameter value; values near 95 represent adequate coverage. Use of boldface type in the top panel indicates problematic levels of bias (i.e., bias whose absolute size is greater than about one half of the estimate's standard error); use of boldface in the bottom panel indicates seriously low levels of coverage (i.e., coverage that falls below 90%, which corresponds to a doubling of the nominal rate of error). MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root-mean-square error.

rameter indicates bias. A rule of thumb that we have found useful is that bias becomes problematic if its absolute size is greater than about one half of the estimate's standard error; above this threshold, bias begins to noticeably degrade the coverage of 95% confidence intervals. Biases that are practically important according to this rule are indicated by boldface type. Some may disagree with this rule, because the standard error of the estimate depends on the sample size ($N = 50$), which was chosen arbitrarily; those readers may compare the average of the parameter estimate to the true value of the parameter and judge for themselves whether the bias seems practically important. The top panel of Table 2 also displays the root-mean-square error (RMSE), which measures the typical distance between the estimate and the true value. Small values of RMSE are desirable. Examining these results, we find that LD is unbiased under MCAR. Under MAR and MNAR, the complete cases are unrepresentative of the population, and biases are substantial, with two exceptions: The estimate of $\beta_{Y|X}$ is unbiased under MAR, and the estimate of $\beta_{X|Y}$ is unbiased under MNAR.

The bottom panel of Table 2 summarizes the performance of confidence intervals. For each parameter and each mechanism, we report the coverage—the percentage of intervals that covered the true parameter—and the average interval width. Narrow intervals are desirable provided that their coverage is near 95%. As a rough rule of thumb, we consider the coverage to be seriously low if it falls below 90%, which corresponds to a doubling of the nominal rate of error; these values are indicated by boldface type. Examining the bottom panel of Table 2, we see that coverage is acceptable for all parameters under MCAR but low for most parameters under MAR and MNAR. Undercoverage has two possible sources: bias, which causes the interval on average to be centered to the left or to the right of the target, and underestimation of the estimate's true variability, which causes the interval to be narrower than it should be. Under MAR and MNAR, both phenomena are occurring; the observed Y values tend to be higher and less variable than those of the full population, which biases both the parameter estimates and their standard errors.

This simulation clearly illustrates that case detection may produce bias under non-MCAR conditions. Some might argue that these results are unrealistic because a missingness rate of 73% for Y is too high. However, it is not difficult to find published analyses in which 73% or more of the sample cases were omitted because of incomplete information. On the basis of a survey of articles in major political science journals, King, Honaker, Joseph, and Scheve (2001) reported alarmingly high rates of case deletion with serious implications for parameter bias and inefficiency. Other simulations revealing the shortcomings of case deletion have been reported by Brown (1994), Graham et al. (1996), and Wothke (2000).

Reweighting

In some non-MCAR situations, it is possible to reduce biases from case deletion by the judicious application of weights. After incomplete cases are removed, the remaining complete cases are weighted so that their distribution more closely resembles that of the full sample or population with respect to auxiliary variables. Weights are derived from the probabilities of response, which must be estimated from the data (e.g., by a logistic or probit regression). Weighting can eliminate bias due to differential response related to the variables used to model the response probabilities, but it cannot correct for biases related to variables that are unused or unmeasured. For a review of weighting in the context of sample surveys, see Little and Rubin (1987, section 4.4).

Weighting is nonparametric, requiring no model for the distribution of the data values in the population. It does, however, require some model for the probabilities of response. Weights are easy to apply for univariate and monotone missing-data patterns. For the arbitrary pattern of missing values shown in Figure 1c, weighting becomes unattractive because one must potentially compute a different set of weights for each variable. Recent years have seen a resurgence of interest in weighting, with new methods for parametric and semiparametric regression appearing in biostatistics; some of these newer methods are reviewed near the end of this article.

Averaging the Available Items

Many characteristics of interest to psychologists—for example, self-esteem, depression, anxiety, quality of life—cannot be reliably measured by a single item, so researchers may create a scale by averaging the responses to multiple items. An average can be motivated by the idea that the items are exchangeable, equally reliable measures of a unidimensional trait. The items are typically standardized to have a mean of zero and a standard deviation of one before averaging. If a participant has missing values for one or more items, it seems more reasonable to average the items

that remain rather than report a missing value for the entire scale. This practice is widespread, but its properties remain largely unstudied; it does not even have a well-recognized name. Undoubtedly, some researchers may have used this method without realizing that it is a missing-data technique, choosing instead to regard it as part of the scale definition. Some might call it *case-by-case item deletion.* A colleague has suggested the term *ipsative mean imputation,* because it is equivalent to substituting the mean of a participant's own observed items for each of his or her missing items (H. B. Bosworth, personal communication, February 2001).

Averaging the available items is difficult to justify theoretically either from a sampling or likelihood perspective. Unlike case deletion, it may introduce bias under MCAR. For example, suppose that a scale is defined as the average of six items, and we compute an average for each participant who responds to at least three. With missing data the variance of the scale tends to increase, because it becomes a mixture of the averages of three, four, five, or six items rather than the average of all six. The scale also becomes less reliable, because reliability decreases as the number of items drops. The method also raises fundamental conceptual difficulties. The scale has been redefined from the average of a given set of items to the average of the available items, a definition that now depends on the particular rates and patterns of nonresponse in the current sample and that also varies from one participant to another. This violates the principle that an estimand be a well-defined aspect of a population, not an artifact of a specific data set.

Despite these theoretical problems, preliminary investigations suggest that the method can be reasonably well behaved. As an illustration, suppose that a scale is defined as the average of standardized items $Y_1, \ldots, Y_4$, but these items are not equally correlated with each other or with other items. In particular, suppose that the covariance matrix for $Y_1, \ldots, Y_4$ is

$$\Sigma = \begin{bmatrix} 1.0 & 0.5 & 0.2 & 0.2 \\ 0.5 & 1.0 & 0.2 & 0.2 \\ 0.2 & 0.2 & 1.0 & 0.5 \\ 0.2 & 0.2 & 0.5 & 1.0 \end{bmatrix},$$

and suppose that the correlation between these and another standardized item X is .50 for Y_1 and Y_2 and .10 for Y_3 and Y_4. Under these circumstances, the true slope for the regression of $\bar{Y} = (Y_1 + Y_2 + Y_3 + Y_4)/4$ on X is $\beta = .30$. Now suppose that Y_1 and Y_2 are missing fairly often, whereas Y_3 and Y_4 are nearly always observed. We imposed missing values in an MCAR fashion at a rate of 30% for Y_1 and Y_2 and 5% for Y_3 and Y_4. We then averaged the items provided that at least two of the four were available (participants with fewer than two were deleted) and regressed the scale on X. Over 1,000 samples of $N = 100$ cases each, the average value of the estimated slope was .26, and 903 of the nominal 95% confidence intervals covered the population value $\beta = .30$. The bias is noticeable but not dramatic. Modifying this example, we found that bias tends to decrease as the Y_js become more equally correlated with each other and with X and as the intercorrelations among the Y_js increase even if they are unequal.

Newer methods of MI provide a more principled solution to this problem. With MI, one imputes missing items prior to forming scales. Using modern software, it is now routinely possible to impute 100 or so items simultaneously and preserve the intercorrelations between the items, provided that enough sample cases are available to estimate the joint covariance structure. If MI is not feasible, then averaging the available items may be a reasonable choice, especially if the reliability is high (say, $\alpha > .70$) and each group of items to be averaged seems to form a single, well-defined domain.

Single Imputation

When a unit provides partial information, it is tempting to replace the missing items with plausible values and proceed with the desired analysis rather than discard the unit entirely. Imputation, the practice of filling in missing items, has several desirable features. It is potentially more efficient than case deletion, because no units are sacrificed; retaining the full sample helps to prevent loss of power resulting from a diminshed sample size. Moreover, if the observed data contain useful information for predicting the missing values, an imputation procedure can make use of this information and maintain high precision. Imputation also produces an apparently complete data set that may be analyzed by standard methods and software. To a data user, the practical value of being able to apply a favorite technique or software product can be immense. Finally, when data are to be analyzed by multiple persons or entities, imputing once, prior to all analyses, helps to ensure that the same set of units is being considered by each entity, facilitating the comparison of results. On the negative side, imputa-

tion can be difficult to implement well, particularly in multivariate settings. Some ad hoc imputation methods can distort data distributions and relationships. The shortcomings of single imputation have been documented by Little and Rubin (1987) and others. Here we briefly classify and review some popular single-imputation methods.

Imputing unconditional means. Consider the popular practice of mean substitution, in which missing values are replaced by the average of the observed values for that item. The average of the variable is preserved, but other aspects of its distribution—variance, quantiles, and so forth—are altered with potentially serious ramifications. Consider a large-sample 95% confidence interval for the population mean,

$$\bar{y} \pm 1.96 \sqrt{\frac{S^2}{N}},$$

where $\bar{y}$ and S^2 are the sample mean and variance and N is the sample size. Mean substitution narrows this interval in two ways: by introducing a downward bias into S^2 and by overstating N. Under MCAR, the coverage probability after mean substitution is approximately $2\Phi(1.96r) - 1$, where Φ is the standard normal cumulative distribution function and r is the response rate. With 25% missing values ($r = .75$) the coverage drops to 86%, and the error rate is nearly three times as high as it should be. In addition to reducing variances, the method also distorts covariances and intercorrelations between variables.

Imputing from unconditional distributions. The idea underlying mean substitution—to predict the missing data values—is somewhat misguided; it is generally more desirable to preserve a variable's distribution. Survey methodologists, who have long been aware of this, have developed a wide array of single-imputation methods that more effectively preserve distributional shape (Madow, Nisselson, & Olkin, 1983). One popular class of procedures known as *hot deck imputation* fills in nonrespondents' data with values from actual respondents. In a simple univariate hot deck, we replace each missing value by a random draw from the observed values. Hot-deck imputation has no parametric model. It partially solves the problem of understating uncertainty, because the variability of the item is not distorted. However, without further refinements, the method still distorts correlations and other measures of association.

Imputing conditional means. In the univariate situation of Figure 1a, a regression model for predicting Y from $X = (X_1, \ldots, X_p)$ may provide a basis for imputation. The model is first fit to the cases for which Y is known. Then, plugging values of X for the nonrespondents into the regression equation, we obtain predictions $\hat{Y}$ for the missing values of Y. Replacing Y with $\hat{Y}$ is called *conditional mean imputation,* because $\hat{Y}$ estimates the conditional mean of Y given X. Conditional mean imputation is nearly optimal for a limited class of estimation problems if special corrections are made to standard errors (Schafer & Schenker, 2000). The method is not recommended for analyses of covariances or correlations, because it overstates the strength of the relationship between Y and the X variables; the multiple regression R^2 among the imputed values is 1.00. If there is no association between Y and the covariates X, then the method reduces to ordinary mean substitution.

Imputing from a conditional distribution. Distortion of covariances can be eliminated if each missing value of Y is replaced not by a regression prediction but by a random draw from the conditional or predictive distribution of Y given X. With a standard linear model, we may add to $\hat{Y}$ a residual error drawn from a normal distribution with mean zero and variance estimated by the residual mean square. In a logistic regression for a dichotomous Y, one may calculate the fitted probability $\hat{p}$ for each case, draw a random uniform variate u, and set $Y = 1$ if $u \leq \hat{p}$ and $Y = 0$ if $u > \hat{p}$.

More generally, suppose that we have data $Y_{com} = (Y_{obs}, Y_{mis})$ from distribution $P(Y_{obs}, Y_{mis}; \theta)$. Imputing from the conditional distribution means simulating a draw from

$$P(Y_{mis}|Y_{obs}; \theta) = \frac{P(Y_{obs}, Y_{mis}; \theta)}{P(Y_{obs}; \theta)}, \qquad (4)$$

where the denominator is given by Equation 2. In practice, the parameters are unknown and must be estimated, in which case we would draw from $P(Y_{mis}|Y_{obs}; \hat{\theta})$, where $\hat{\theta}$ is an estimate of θ obtained from Y_{obs}. Imputing from Equation 4 assumes MAR. The method produces nearly unbiased estimates for many population quantities under MAR if the model (Equation 4) is correctly specified. Formulating the conditional distribution and drawing from it tends to be easiest for univariate missing-data patterns. With a monotone pattern, the conditional distribution can be expressed as a sequence of regressions for Y_j given $Y_1, \ldots, Y_{j-1}$ for $j = 1, \ldots, p$, which is not too difficult. With the arbitrary patterns, the conditional dis-

tribution can be quite complicated, and drawing from it may require nearly as much effort as full MI, which has superior properties.

Example. Consider the problem of imputing the missing blood pressure readings in Table 1. For the MAR condition, we imputed the missing values of Y by four methods: (a) performing mean substitution, (b) using a simple hot deck, (c) performing conditional mean imputation based on the linear regression of Y on X, and (d) drawing from the estimated predictive distribution of Y given X based on the same regression. Bivariate scatter plots of Y versus X for the four imputation methods are displayed in Figure 3. These plots clearly reveal the shortcomings of the first three methods. Because the variables are jointly normal with correlation .60, a plot of complete data should resemble an elliptical cloud with a moderate positive slope. Mean substitution (see Figure 3a) causes all the imputed values of Y to fall on a horizontal line, whereas conditional mean imputation (see Figure 3c) causes them to fall on a regression line. The hot deck (see Figure 3b) produces an elliptical cloud with too little correlation. The only method that produces a reasonable point cloud is (see Figure 3d) imputation from the conditional distribution of Y and X.

To illustrate the operating characteristics of these methods, we performed a simulation similar to the one we did for case deletion. One thousand samples of $N = 50$ were drawn from the bivariate normal population, and missing values were imposed on each sample by the MCAR, MAR, and MNAR methods. Each incomplete data set was then imputed by the four methods shown in Figure 3. The results are summarized in Table 3. As before, the top panel of the table reports the average of the parameter estimates and the RMSE for each mechanism–imputation method combination; estimates with substantial bias are displayed in boldface. Coverage and average width of the nominal 95% confidence intervals are shown in the bottom panel, with seriously low coverages in boldface. On the basis of these results in the top panel, we make the following general observations. Mean substitution and the hot deck produce biased estimates for many parameters under any type of missingness. Conditional mean imputation performs slightly better but still may introduce bias. Im-

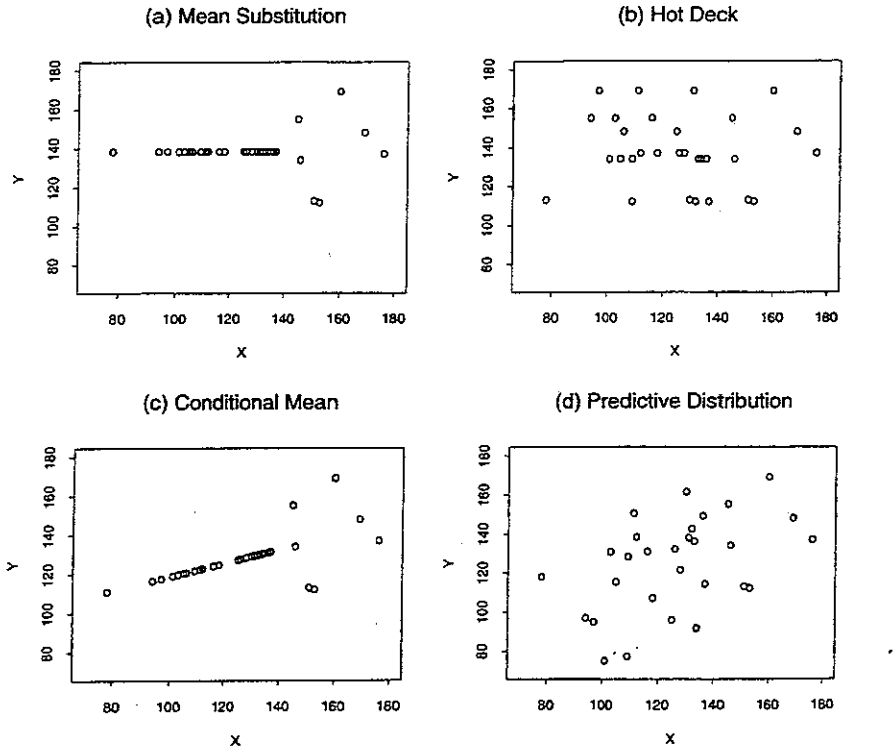

Figure 3. Example scatter plots of blood pressure measured at two occasions, with missing at random missing values imputed by four methods.

Table 3
Performance of Single-Imputation Methods for Parameter Estimates and Confidence Intervals Over 1,000 Samples (N = 50 Participants)

Parameter	MCAR				MAR				MNAR			
	MS	HD	CM	PD	MS	HD	CM	PD	MS	HD	CM	PD
	Average parameter estimate (with RMSE in parentheses)											
μ_Y = 125.0	125.1 (7.18)	125.2 (7.89)	125.2 (6.26)	125.1 (6.57)	**143.5** (19.4)	**143.5** (19.5)	124.9 (18.1)	124.8 (18.3)	**155.5** (30.7)	**155.5** (30.73)	**151.6** (26.9)	**151.6** (26.9)
σ_Y = 25.0	**12.3** (13.0)	23.4 (5.40)	**18.2** (8.57)	24.7 (5.37)	**10.6** (14.6)	**20.0** (6.68)	20.4 (10.7)	27.0 (8.77)	**6.20** (18.9)	**11.7** (13.7)	**8.42** (16.9)	**12.9** (12.7)
ρ = .60	**.30** (.32)	**.16** (.46)	**.79** (.27)	.59 (.20)	**.08** (.52)	**.04** (.57)	.64 (.48)	.50 (.40)	**.15** (.47)	**.08** (.53)	.55 (.40)	**.38** (.37)
$\beta_{Y|X}$ = .60	**.16** (.45)	**.16** (.47)	.61 (.25)	.60 (.27)	**.04** (.56)	**.04** (.57)	.61 (.57)	.62 (.57)	**.04** (.56)	**.04** (.56)	**.21** (.43)	**.21** (.43)
$\beta_{X|Y}$ = .60	.61 (.26)	**.17** (.46)	**1.12** (.64)	.60 (.24)	**.20** (.44)	**.06** (.56)	.78 (.75)	.45 (.40)	.61 (.55)	**.19** (.53)	**1.63** (1.72)	.76 (.68)
	Coverage (with average interval width in parentheses)											
μ_Y	**39.2** (7.0)	**60.0** (13.3)	**58.5** (10.4)	**71.0** (14.1)	**0.2** (6.0)	**2.4** (11.4)	**25.7** (11.6)	**32.3** (15.3)	**0.0** (3.5)	**0.0** (6.7)	**0.0** (4.8)	**0.0** (7.3)
σ_Y	**0.7** (5.1)	**63.7** (9.6)	**31.3** (7.5)	**65.4** (10.2)	**0.1** (4.4)	**45.3** (8.2)	**30.0** (8.4)	**49.4** (11.1)	**0.0** (2.5)	**1.7** (4.8)	**0.7** (3.5)	**4.4** (5.3)
ρ	**25.5** (0.50)	**5.5** (0.53)	**21.7** (0.19)	**65.0** (0.35)	**0.0** (0.55)	**0.0** (0.55)	**19.6** (0.21)	**40.7** (0.34)	**2.2** (0.54)	**0.5** (0.54)	**37.6** (0.31)	**50.0** (0.43)
$\beta_{Y|X}$	**1.2** (0.27)	**16.5** (0.54)	**38.6** (0.22)	**63.5** (0.44)	**0.0** (0.25)	**0.8** (0.47)	**17.2** (0.22)	**33.5** (0.45)	**0.0** (0.14)	**0.1** (0.27)	**3.1** (0.13)	**7.4** (0.26)
$\beta_{X|Y}$	98.1 (1.18)	**23.9** (0.63)	**8.9** (0.50)	**71.1** (0.47)	91.2 (1.43)	**14.5** (0.75)	**18.6** (0.56)	**60.0** (0.46)	97.4 (2.50)	**71.3** (1.30)	**19.1** (1.46)	**56.2** (1.05)

Note. Parameters: μ is the population mean; σ is the population standard deviation; ρ is the population correlation; β is the population regression slope. Coverage represents the percentage of confidence intervals that include the parameter value; values near 95 represent adequate coverage. Use of boldface type in the top panel indicates problematic levels of bias (i.e., bias whose absolute size is greater than about one half of the estimate's standard error); use of boldface in the bottom panel indicates seriously low levels of coverage (i.e., coverage that falls below 90%, which corresponds to a doubling of the nominal rate of error). MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; MS = mean substitution; HD = hot deck; CM = conditional mean imputation; PD = predictive distribution imputation; RMSE = root-mean-square error.

puting from a conditional distribution is essentially unbiased under MCAR or MAR but potentially biased under MNAR.

The problem of undercoverage. Examining the bottom panel of Table 3, we find that the performance of estimated intervals is a disaster. In nearly all cases, the actual coverage is much lower than 95%. (The only exceptions to this rule occur for $\beta_{X|Y}$ under mean substitution, which causes the intervals to be exceptionally wide). This shortcoming of single imputation is well documented (Rubin, 1987). Even if it preserves marginal and joint distributions, it still tends to understate levels of uncertainty, because conventional uncertainty measures ignore the fact the imputed values are only guesses. With single imputation, there is no simple way to reflect missing-data uncertainty. In this example, the implications are very serious because the amount of missing information is unusually high, but even in less extreme situations undercoverage is still an issue. The problem of undercoverage is solved by MI, to be described later.

When single imputation is reasonable. Despite the discouraging news in the bottom panel of Table 3, there are situations in which single imputation is reasonable and better than case deletion. Imagine a data set with 25 variables in which 3% of all the data values are missing. If missing values are spread uniformly across the data matrix, then LD discards more than half of the participants ($1 - .97^{25} = .53$). On the other hand, imputing once from a conditional distri-

bution permits the use of all participants with only a minor negative impact on estimates and uncertainty measures.

ML Estimation

The principle of drawing inferences from a likelihood function is widely accepted. Under MAR, the marginal distribution of the observed data (Equation 2) provides the correct likelihood for the unknown parameters θ, provided that the model for the complete data is realistic. Little and Rubin (1987, p. 89) referred to this function as "the likelihood ignoring the missing-data mechanism," but for brevity we simply call it the *observed-data likelihood.* The logarithm of this function,

$$l(\theta; Y_{obs}) = \log L(\theta; Y_{obs}), \quad (5)$$

plays a crucial role in estimation. The ML estimate $\hat{\theta}$, the value of θ for which Equation 5 is highest, has attractive theoretical properties just as it does in complete-data problems. Under rather general regularity conditions, it tends to be approximately unbiased in large samples. It is also highly efficient; as the sample size grows, its variance approaches the theoretical lower bound of what is achievable by any unbiased estimator (e.g., Cox & Hinkley, 1974).

Confidence intervals and regions are often computed by appealing to the fact that, in regular problems with large samples, $\hat{\theta}$ is approximately normally distributed about the true parameter θ with approximate covariance matrix

$$V(\hat{\theta}) \approx [-l''(\hat{\theta})]^{-1}, \quad (6)$$

where $l''(\hat{\theta})$ is the matrix of second partial derivatives of Equation 5 with respect to the elements of θ. The matrix $-l''(\hat{\theta})$, which is often called *observed information,* describes how quickly the log-likelihood function drops as we move away from the ML estimate; a steep decline indicates that the ML estimate is apparently precise, whereas a gradual decline implies there is considerable uncertainty about where the true parameter lies. This matrix is sometimes replaced by its expected value, which is called *expected information* or *Fisher information,* because the expected value is sometimes easier to compute. In complete-data problems, the approximation (Equation 6) is still valid when the observed information is replaced by the expected information. However, as recently pointed out by Kenward and Molenberghs (1998), this is not necessarily true with missing data. Expected information implicitly uses Equation 5 as a sampling distribution for Y_{obs}, which is valid only if the missing data are MCAR. In missing-data problems, therefore, if we want to obtain standard errors and confidence intervals that are valid under the general MAR condition, we should base them on an observed rather than an expected information matrix. If a software package that performs ML estimation with missing data offers the choice of computing standard errors from the observed or the expected information, the user should opt for the former.

Log likelihood also provides a method for testing hypotheses about elements or functions of θ. Suppose that we wish to test the null hypothesis that θ lies in a certain area or region of the parameter space versus the alternative that it does not. Under suitable regularity conditions, this test may be performed by comparing a difference in log likelihoods to a chi-square distribution. More specifically, let $\hat{\theta}$ denote the maximizer of the log likelihood over the full parameter space, and let $\tilde{\theta}$ be the maximizer over the region defined by the null hypothesis. We would reject the null at the designated alpha level if $2\,[l(\hat{\theta}; Y_{obs}) - l(\tilde{\theta}; Y_{obs})]$ exceeds the 100 $(1 - \alpha)$ percentile of the chi-square distribution. The degrees of freedom are given by the difference in the number of free parameters under the null and alternative hypotheses—that is, the number of restrictions that must be placed on the elements of θ to ensure that it lies in the null region. These likelihood-ratio tests are quite attractive for missing-data problems, because they only require that we be able to compute the maximizers $\hat{\theta}$ and $\tilde{\theta}$; no second derivatives are needed.

Computing ML Estimates

In a few problems, the maximizer of the log likelihood (Equation 5) can be computed directly. One famous example is the bivariate normal monotone problem presented by Anderson (1957). Suppose that a portion of sample units (Part A) has values recorded for variables X and Y, and the remainder of sample units (Part B) has values only for X. The basic idea is as follows: Use the full sample (A + B) to estimate the mean and variance of X, use the reduced sample (A only) to estimate the parameters of the linear regression of Y and X, and then combine the two sets of estimates to obtain the parameters for the full joint distribution. For details on this procedure, see Little and Rubin (1987, section 6.2).

Except for these special cases, expressions for ML estimates cannot in general be written down in closed

form, and computing them requires iteration. A general method for ML in missing-data problems was described by Dempster et al. (1977) in their influential article on the EM algorithm. The key idea of EM is to solve a difficult incomplete-data estimation problem by iteratively solving an easier complete-data problem. Intuitively, we "fill in the missing data" with a best guess at what it might be under the current estimate of the unknown parameters, then reestimate the parameters from the observed and filled-in data. To obtain the correct answer, we must clarify what it means to "fill in the missing data." Dempster et al. showed that, rather than filling in the missing data values per se, we must fill in the complete-data sufficient statistics. The form of these statistics depends on the model under consideration. Overviews of EM have been given by Little and Rubin (1987), Schafer (1997), and McLachlan and Krishnan (1996).

Little and Rubin (1987) catalogued EM algorithms for any missing-data problems. EM has also been applied to many situations that are not necessarily thought of as missing-data problems but can be formulated as such: multilevel linear models for unbalanced repeated measures data, where not all participants are measured at all time points (Jennrich & Schluchter, 1986; Laird & Ware, 1982); latent class analysis (Clogg & Goodman, 1984) and other finite-mixture models (Titterington, Smith, & Makov, 1985); and factor analysis (Rubin & Thayer, 1983). For some of these problems, non-EM methods are also available. Newton–Raphson and Fisher scoring are now considered by many to be the preferred method for fitting multilevel linear models (Lindstrom & Bates, 1988). However, in certain classes of models—finite mixtures, for example—EM is still the method of choice (McLachlan & Peel, 2000).

Software for ML Estimation in Missing-Data Problems

An EM algorithm for ML estimation of an unstructured covariance matrix is available in several programs. The first commercial implementation was released by BMDP (BMDP Statistical Software, 1992), now incorporated into the missing-data module of SPSS (Version 10.0). The procedure is also found in EMCOV (Graham & Hofer, 1991), NORM, SAS (Y. C. Yuan, 2000), Amelia (King et al., 2001), S-PLUS (Schimert, Schafer, Hesterberg, Fraley, & Clarkson, 2001), LISREL (Jöreskog & Sörbom, 2001), and Mplus (L. K. Muthén & Muthén, 1998).

ML is also available for normal models with structured covariance matrices. Multilevel linear models can be fit with HLM (Bryk, Raudenbush, & Congdon, 1996), MLWin (Multilevel Models Project, 1996), the SAS procedure PROC MIXED (Littell, Milliken, Stroup, & Wolfinger, 1996), Stata (Stata, 2001), and the lme function in S-PLUS (Insightful, 2001). Any of these may be used for repeated measures data. In some cases, the documentation and accompanying literature do not mention missing values specifically but describe "unbalanced" data sets, in which participants are not measured at a common set of time points. We must emphasize that if the imbalance occurs not by design but as a result of uncontrolled nonresponse (e.g., attrition), all of these programs will assume MAR. ML estimates for structural equation models with incomplete data are available in Mx (Neale, Boker, Xie, & Maes, 1999), AMOS (Arbuckle & Wothke, 1999), LISREL, and Mplus, which also assume MAR. These programs provide standard errors based on expected or observed information. If offered a choice, the user should opt for observed rather than expected, because the latter is appropriate only under MCAR. The producers of EQS (Bentler, in press) have also announced plans for a new version with missing-data capabilities, but as of this writing it has not yet been released.

Latent class analysis (LCA) is a missing-data problem in the sense that the latent classification is missing for all participants. A variety of software packages for LCA are available; one of the most popular is LEM (Vermunt, 1997). Latent transition analysis (LTA), an extension of LCA to longitudinal studies with a modest number of time points, is available in WinLTA (Collins, Flaherty, Hyatt, & Schafer, 1999). The EM algorithm in the most recent version of LTA allows missing values to occur on the manifest variables in an arbitrary pattern. ML estimates for a wider class of latent-variable models with incomplete data, including finite mixtures and models with categorical responses, are available in Mplus.

Example

To illustrate the properties of likelihood methods, we conducted another simulation using the blood pressure example. One thousand samples of size $N = 50$ were generated, and missing values were imposed on each sample by the MCAR, MAR, and MNAR methods. From each incomplete data set, ML estimates were computed by the technique of Anderson (1957). Standard errors were obtained by inverting the observed information matrix, and approximate 95%

confidence intervals for each parameter were computed by the normal approximation (estimate ±1.96 *SE*s).

Results from this simulation are summarized in Table 4. The top panel of Table 4 reports the average and RMSE of the estimates, and the bottom panel reports the coverage and average width of the intervals. Once again, estimates whose bias exceeds one half of a standard error and coverage values below 90% are displayed in boldface. Examining the top panel, we find that the ML estimates are not substantially biased under MCAR or MAR but are quite biased under MNAR. This agrees with the theoretical result that likelihood inferences are appropriate under any MAR situation. In the bottom panel, however, notice that the coverage for some intervals under MAR is quite poor. The reason for this is that a sample of $N = 50$ with a high rate of missing values is not large enough for the normal approximation to work well. Another simulation was performed with the sample size increased to 250. Results from that simulation, which we omit, show that the intervals for all parameters achieve coverage near 95% under MCAR and MAR but have seriously low coverage under MNAR.

Table 4
Performance of Maximum Likelihood for Parameter Estimates and Confidence Intervals Over 1,000 Samples (N = 50 Participants)

Parameter	MCAR	MAR	MNAR
Average parameter estimate (with RMSE in parentheses)			
$\mu_Y = 125.0$	124.8 (6.52)	125.2 (16.9)	**151.6** (26.9)
$\sigma_Y = 25.0$	24.2 (5.73)	25.5 (7.45)	**12.3** (13.2)
$\rho = .60$	.61 (.19)	.52 (.38)	**.39** (.36)
$\beta_{Y\|X} = .60$	.61 (.27)	.60 (.51)	**.21** (.43)
$\beta_{X\|Y} = .60$	.63 (.23)	.49 (.38)	.79 (.68)
Coverage (with average interval width in parentheses)			
μ_Y	91.2 (22.2)	91.6 (58.2)	**0.9** (16.4)
σ_Y	**86.1** (17.8)	90.2 (28.6)	**7.4** (9.94)
ρ	**84.2** (0.65)	**76.7** (1.20)	**89.2** (0.99)
$\beta_{Y\|X}$	90.3 (0.88)	90.7 (1.78)	**28.2** (0.59)
$\beta_{X\|Y}$	91.6 (0.80)	93.0 (1.26)	**80.5** (2.18)

Note. Parameters: μ is the population mean; σ is the population standard deviation; ρ is the population correlation; β is the population regression slope. Coverage represents the percentage of confidence intervals that include the parameter value; values near 95 represent adequate coverage. Use of boldface type in the top panel indicates problematic levels of bias (i.e., bias whose absolute size is greater than about one half of the estimate's standard error); use of boldface in the bottom panel indicates seriously low levels of coverage (i.e., coverage that falls below 90%, which corresponds to a doubling of the nominal rate of error). MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root-mean-square error.

General Comments on ML for Missing-Data Problems

In theory, likelihood methods are more attractive than ad hoc techniques of case deletion and single imputation. However, they still rest on a few crucial assumptions. First, they assume that the sample is large enough for the ML estimates to be approximately unbiased and normally distributed. In missing-data problems the sample may have to be larger than usual, because missing values effectively reduce the sample size. Second, the likelihood function comes from an assumed parametric model for the complete data $P(Y_{obs}, Y_{mis}; \theta)$. Depending on the particular application, likelihood methods may or may not be robust to departures from model assumptions. Sometimes (e.g., in structural equation models) departures might not have a serious effect on estimates but could cause standard errors and test statistics to be very misleading (Satorra & Bentler, 1994). If one dispenses with the full parametric model, estimation procedures with incomplete data are still possible, but they typically require the missing values to be MCAR rather than MAR (K. H. Yuan & Bentler, 2000; Zeger, Liang, & Albert, 1988). For an evaluation of these new procedures for structural equation models, see the recent article by Enders (2001).

Finally, the likelihood methods described in this section assume MAR. When missingness is not controlled by the researcher, it is unlikely that MAR is precisely satisfied. In many realistic applications, however, we believe that departures from MAR are not large enough to effectively invalidate the results of an MAR-based analysis (Collins et al., 2001). When the reasons for missingness seem strongly related to the data, one can formulate a likelihood or

Bayesian solution to take this into account. However, these methods—which are reviewed in the last section of this article—are not a panacea, because they still rest on unverifiable assumptions and may be sensitive to departures from the assumed model.

Multiple Imputation

MI, proposed by Rubin (1987), has emerged as a flexible alternative to likelihood methods for a wide variety of missing-data problems. MI retains much of the attractiveness of single imputation from a conditional distribution but solves the problem of understating uncertainty. In MI, each missing value is replaced by a list of $m > 1$ simulated values as shown in Figure 4. Substituting the jth element of each list for the corresponding missing value, $j = 1, \ldots, m$, produces m plausible alternative versions of the complete data. Each of the m data sets is analyzed in the same fashion by a complete-data method. The results, which may vary, are then combined by simple arithmetic to obtain overall estimates and standard errors that reflect missing-data uncertainty as well as finite-sample variation. Reviews of MI have been published by Rubin (1996) and Schafer (1997, 1999a). Graham, Cumsille, and Elek-Fisk (in press) and Sinharay, Stern, and Russell (2001) have provided less technical presentations for researchers in psychology.

MI has many attractive features. Like single imputation, it allows the analyst to proceed with familiar complete-data techniques and software. One good set of m imputations may effectively solve the missing-data problems in many analyses; one does not necessarily need to re-impute for every new analysis. Unlike other Monte Carle methods, with MI we do not need a large number of repetitions for precise estimates. Rubin (1987) showed that the efficiency of an estimate based on m imputations, relative to one based on an infinite number, is $(1 + \lambda/m)^{-1}$, where λ is the rate of missing information.[2] For example, with 50% missing information, $m = 10$ imputations is $100/(1 + .05) = 95\%$ efficient; additional imputations do little to remove noise from the estimate itself. In some cases, researchers also like to remove noise from other statistical summaries (e.g., significance levels or probability values); in many practical applications, we have found that $m = 20$ imputations can effectively do this. Once a procedure for managing multiple versions of the data has been established, the additional time and effort required to handle $m = 20$ versions rather than $m = 10$ is often of little consequence.

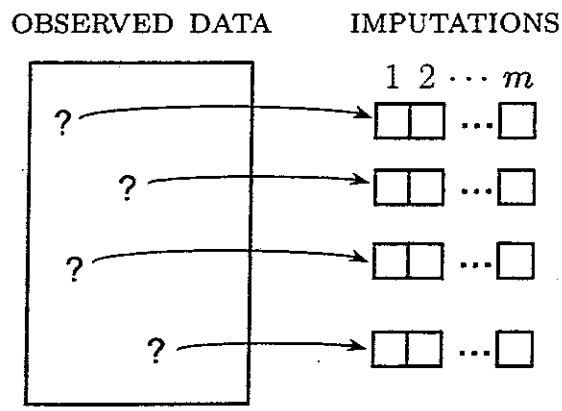

Figure 4. Schematic representation of multiple imputation, where m is the number of imputations.

Rubin's Rules for Combining Estimates and Standard Errors

The simplest method for combining the results of m analyses is Rubin's (1987) method for a scalar (one-dimensional) parameter. Suppose that Q represents a population quantity (e.g., a regression coefficient) to be estimated. Let $\hat{Q}$ and $\sqrt{U}$ denote the estimate of Q and the standard error that one would use if no data were missing. The method assumes that the sample is large enough so that $\sqrt{U}(\hat{Q} - Q)$ has approximately a standard normal distribution, so that $\hat{Q} \pm 1.96\sqrt{U}$ has about 95% coverage. Of course, we cannot compute $\hat{Q}$ and U; rather, we have m different versions of them, $[\hat{Q}^{(j)}, U^{(j)}]$, $j = 1, \ldots, m$. Rubin's (1987) overall estimate is simply the average of the m estimates,

$$\overline{Q} = m^{-1}\sum_{j=1}^{m}\hat{Q}^{(j)}.$$

The uncertainy in $\overline{Q}$ has two parts: the average within-imputation variance,

$$\overline{U} = m^{-1}\sum_{j=1}^{m}U^{(j)},$$

and the between-imputations variance,

[2] The rate of missing information, as distinct from the rate of missing observations, measures the increase in the large-sample variance of a parameter estimate (Equation 6) due to missing values. It may be greater or smaller than the rate of missing values in any given problem.

$$B = (m-1)^{-1}\sum_{j=1}^{m}[\hat{Q}^{(j)} - \overline{Q}]^2.$$

The total variance is a modified sum of the two components,

$$T = \overline{U} + (1 + m^{-1})B,$$

and the square root of T is the overall standard error. For confidence intervals and tests, Rubin (1987) recommended the use of a Student's t approximation $T^{-1/2}(\bar{Q} - Q) \sim t_\nu$, where the degrees of freedom are given by

$$\nu = (m-1)\left[1 + \frac{\overline{U}}{(1+m^{-1})B}\right]^2.$$

The degree of freedom may vary from $m - 1$ to ∞ depending on the rate of missing information. When the degrees of freedom are large, the t distribution is essentially normal, the total variance is well estimated, and there is little to be gained by increasing m. The estimated rate of missing information for Q is approximately $\tau/(\tau + 1)$, where $\tau = (1 + m^{-1})B/\bar{U}$ is the relative increase in variance due to nonresponse. Additional methods for combining multidimensional parameter estimates, likelihood-ratio test statistics, and probability values from hypothesis tests were reviewed by Schafer (1997, chapter 4).

Proper MI

The validity of MI rests on how the imputations are created and how that procedure relates to the model used to subsequently analyze the data. Creating MIs often requires special algorithms (Schafer, 1997). In general, they should be drawn from a distribution for the missing data that reflects uncertainty about the parameters of the data model. Recall that with single imputation, it is desirable to impute from the conditional distribution $P(Y_{\text{mis}}|Y_{\text{obs}}; \hat{\theta})$, where $\hat{\theta}$ is an estimate derived from the observed data. MI extends this by first simulating m independent plausible values for the parameters, $\theta^{(1)}, \ldots, \theta^{(m)}$, and then drawing the missing data $Y_{\text{mis}}^{(t)}$ from $P[Y_{\text{mis}}|Y_{\text{obs}}; \theta^{(t)}]$ for $t = 1, \ldots, m$.

Treating parameters as random rather than fixed is an essential part of MI. For this reason, it is natural (but not essential) to motivate MI from the Bayesian perspective, in which the state of knowledge about parameters is represented through a posterior distribution. Bayesian methods have become increasingly popular in recent years; a good modern overview was provided by Gelman et al. (1995). For our purposes, it is not necessary to delve into the details of the Bayesian perspective to explain how MI works; we simply note a few principles of Bayesian analysis.

First, in a Bayesian analysis, all of the data's evidence about parameters is summarized with a likelihood function. As with ML, the assumed parametric form of the model may be crucial; if the model is inaccurate, then the posterior distribution may provide an unrealistic view of the state of knowledge about θ. Second, Bayesian analysis requires a prior distribution for the unknown parameters. Critics may regard the use of a prior distribution as subjective, artificial, or unscientific. We tend to regard it as a "necessary evil." In some problems, prior distributions can be formulated to reflect a state of relative ignorance about the parameters, mitigating the effect of the subjective inputs. Finally, in a Bayesian analysis, the influence of the prior diminishes as the sample size increases. Indeed, one often finds that a range of plausible alternative priors leads to similar posterior distributions. Because MI already relies on large-sample approximations for the complete-data distribution, the prior rarely exerts a major influence on the results.

Creating MIs Under a Normal Model

To understand what is happening within an MI algorithm, consider a hypothetical data set with three variables Y_1, Y_2, and Y_3, which we assume to be jointly normally distributed. Suppose that one group of participants (Group A) has measurements for all three variables, another group (Group B) has measurements for Y_1 and Y_2 but missing values for Y_3, and a third group (Group C) has measurements for Y_3 but missing values for Y_1 and Y_2. The parameters of the trivariate normal model—three means, three variances and three correlations—are not known and should be estimated from all three groups. If the parameters were known, MIs could be drawn in the following way. Group A requires no imputation. For Group B, we would need to compute the linear regression of Y_3 on Y_1 and Y_2. Then, for each participant in Group B, we would use his or her own values of Y_1 and Y_2 to predict the unknown value of Y_3 and impute the predicted value $\hat{Y}_3$ plus random noise drawn from a normal distribution with the appropriate residual variance. For Group C, we would compute the bivariate regression of Y_1 and Y_2 on Y_3, obtain the joint prediction $(\hat{Y}_1, \hat{Y}_2)$ for each participant, and add random noise drawn from a bivariate normal distribution

with the appropriate residual variances and covariance.

A crucial feature of MI is that the missing values for each participant are predicted from his or her own observed values, with random noise added to preserve a correct amount of variability in the imputed data. Another feature is that the joint relationships among the variables Y_1, Y_2, and Y_3 must be estimated from all available data in Groups A, B, and C. ML estimates of the parameters could be computed using an EM algorithm, but proper MI requires that we reflect uncertainty about these parameters from one imputation to the next. Therefore, instead of using ML estimates, we need to draw random values of the parameters from a posterior distribution based on the observed-data likelihood and a prior. The form of this posterior distribution is not easy to describe, but it can be sampled from by a variety of techniques; data augmentation (Schafer, 1997) is straightforward, but one could also use importance resampling (King et al., 2001). The effect of drawing parameters from a posterior distribution, rather than using ML estimates, means that for Group B the regression of Y_3 on Y_1 and Y_2 will be randomly perturbed from one set of imputations to the next; similarly, in Group C the joint regression of (Y_1, Y_2) on Y_3 will also be perturbed.

A review of MI computations is beyond the scope of this article. Rubin (1987) described how to create MIs for some univariate and monotone situations shown in Figure 1, a and b. Algorithms for multivariate data with arbitrary patterns were given by Schafer (1997), along with detailed guidance and data examples. Proper application of these procedures requires some understanding of the properties of data augmentation, especially its convergence behavior. A gentle introduction and tutorial on the use of MI under a multivariate normal model was provided by Schafer and Olsen (1998; also see Graham et al., in press; Sinharay et al., 2001).

Choosing the Imputation Model

Notice that the MI procedure described above is based on a joint normality assumption for Y_1, Y_2, and Y_3. This model makes no distinctions between response (dependent) or predictor (independent) variables but treats all three as a multivariate response. The imputation model is not intended to provide a parsimonious description of the data, nor does it represent structural or causal relationships among variables. The model is merely a device to preserve important features of the joint distribution (means, variances, and correlations) in the imputed values. A procedure that preserves the joint distribution of Y_1, Y_2, and Y_3 will automatically preserve the linear regression of any of these variables on the others. Therefore, in a subsequent analysis of the imputed data, any variable could be treated as a response or as a predictor. For example, we may regress Y_2 on Y_1 in each imputed data set and combine the estimated intercepts and slopes by Rubin's (1987) rules. Distinctions between dependent and independent variables and substantive interpretation of relationships should be left to postimputation analyses.

Although it is not necessary to have a scientific theory underlying an imputation model, it is crucial for that model to be general enough to preserve effects of interest in later analyses. Suppose that we will examine differences in mean response between an experimental and control group. For differences to be preserved in the imputed values, some indicator of group membership should enter the imputation model. For example, a dummy variable (0 = control, 1 = experimental) could be included in the normal model, which will preserve the main effect of group membership on any other variable. To preserve interaction between group membership and other variables, one could split the data set and apply a separate imputation model to each group.

Real data rarely conform to normality. Some might hesitate to use a normal imputation model, fearing that it may distort the distributions of nonnormal variables. Many tricks are available to help preserve distributional shape (Schafer, 1997). Binary or ordinal variables may be imputed under a normality assumption and then rounded off to discrete values. If a variable is right skewed, it may be modeled on a logarithmic (or some other power-transformed) scale and transformed back to the original scale after imputation. Transformations and rounding help to make the imputed values aesthetically appealing; for example, a log transformation for a positive variable will guarantee that the imputed values are always positive. Depending on the analysis applied to the imputed data, however, these procedures may be superfluous. Graham and Schafer (1999) presented a simulation in which highly nonnormal variables were imputed under normality assumptions with no transformations or rounding and reported excellent performance for linear regression even with relatively small samples. Although joint normality is rarely realistic, we have found the model to be useful in a surprisingly wide variety of problems. Analyses involving normal-

based MI with real data have been published by Graham and Hofer (2000), Graham et al. (in press), and others.

Of course, there are situations in which the normal model should be avoided—to impute variables that are nominal (unordered categories), for example. Schafer (1997) presented imputation methods for multivariate categorical data and for data sets containing both continuous and categorical variables. With these models, one can specify and preserve higher order associations among the variables, provided that the data are rich enough to estimate these associations. These models assume that the rows or observational units in the data set have been independently sampled and thus do not automatically take into account longitudinal or clustered structure. Imputation models specifically designed for longitudinal and clustered data have been described by Liu, Taylor, and Belin (2000) and Schafer (2001).

For a small class of problems, it is possible to create MIs without a data model. Rubin (1987) described a method called the *approximate Bayesian bootstrap* (ABB), which involves two cycles of hot-deck imputation. Rubin's ABB applies to univariate missing data without covariates. Lavori, Dawson, and Shera (1995) generalized the ABB to include fully observed covariates as in Figure 1a. This method has been implemented in a program called SOLAS (Statistical Solutions, 1998), where it is called the *propensity-score* option. This procedure was designed to provide unbiased estimates of the distribution of a single outcome, not to preserve the relationship between the outcome and other items. Imputations created by this method may seriously distort covariance structure (Allison, 2000). An alternative procedure in SOLAS, called the *model-based* method, is described below; it is more appropriate for situations in which postimputation analyses will involve covariances and correlations.

MI Software

Many computer programs for MI are now available. NORM, a free program for Windows, creates MIs for incomplete data with arbitrary patterns of missing values under an unstructured normal model. Algorithms used in NORM were described by Schafer (1997). NORM includes utilities for automatic pre- and postimputation transformations and rounding of imputed values. A new SAS procedure, PROC MI (Y. C. Yuan, 2000), brings the method used by NORM into the popular SAS environment. Other versions of the same algorithm have also been implemented in a newly released missing-data library in S-PLUS and in LISREL. Amelia, a free program created by King et al. (2001), relies on the normal model but uses a different computational technique.

With any of the programs mentioned above, a sufficient number of cases must be available to estimate an unstructured covariance matrix for all variables. This may be problematic. For example, in psychological research it is not uncommon to collect about $p = 100$ or more items for only $N = 100$ subjects, with the intention of later averaging the items into a few scales or subscales. Without imposing additional prior information or structure, we cannot fit a normal model to all items at once. Song (1999) simplified the covariance structure by assuming a few common factors, generating MIs under this restricted model; however, a software implementation of this new method is not yet available. Longitudinal structure arising from repeated measurements over time has been implemented in the S-PLUS function PAN (Schafer, 2001; Schafer & Yucel, in press).

For nonnormal imputation models, the software choices are more limited. Methods described by Schafer (1997) for multivariate categorical data, and for mixed data sets containing both continuous and categorical variables, are commercially available in the new S-PLUS missing-data module; this supersedes the older CAT and MIX libraries for S-PLUS written by Schafer. The propensity score method of SOLAS does not assume a parametric model for the data but, as previously mentioned, can seriously distort intervariable relationships and is potentially dangerous.

Rather than describing the joint distribution of all variables by a multivariate model, some prefer to implement MI with a sequence of single-response regressions. With the monotone pattern of Figure 1b, the information about missing values can indeed be captured by a sequence of regressions for Y_j given $Y_1, \ldots, Y_{j-1}$ for $j = 1, \ldots, p$. The so-called model-based method in SOLAS (Version 2.0 or later) creates proper MIs for univariate or monotone patterns by such a regression sequence. The method can also be applied to nonmonotone patterns, but cases that do not conform to the monotone pattern are ignored in the model fitting. For data sets that deviate substantially from a monotonicity, the joint modeling procedures used in S-PLUS seem preferable.

In a large survey application, Kennickell (1991) performed approximate MI for nonmonotone data by specifying a single-response regression model for

each variable given the others and repeatedly fitting the models in an iterated sequence. A general implementation of this method using SAS macros is found in the IVEware library (Raghunathan, Solenberger, & Van Hoewyk, 2000). A similar library for S-PLUS, called MICE, was written by Van Buuren and Oudshoorn (1999). From a theoretical standpoint, this technique is problematic, because the sequence of regression models might not be consistent with a true joint distribution. Technically speaking, these iterative algorithms may never "converge" because the joint distribution to which they may converge does not exist. Nevertheless, simulation work (Brand, 1999) suggests that in some practical applications the method can indeed work well despite the theoretical problems.

Some utilities are also available to simplify the task of analyzing imputed data sets. NORM combines parameter estimates and standard errors from multiple analyses using Rubin's (1987) rules; it can also combine groups of coefficients and their covariance matrices for multiparameter inference. A new SAS procedure called PROC MIANALYZE does essentially the same thing. The new S-PLUS missing-data library includes the functions miApply and miEval, which automatically carry out a data analysis procedure (e.g., fitting a regression model) for all the imputed data sets, storing the results together in a convenient format; four additional functions are available to consolidate the results.

This list of software is not exhaustive and may be outdated when this article appears in print. For up-to-date information on MI software, readers should refer to on-line resources; a good starting point is the Web site at http://www.multiple-imputation.com. Some of the programs listed above were recently reviewed by Horton and Lipsitz (2001).

Example

Returning to the blood pressure example, we simulated one thousand samples of $N = 50$ participants and imposed missing values by the MCAR, MAR, and MNAR methods. Using NORM, we multiply imputed the missing values for each incomplete data set 20 times. We chose $m = 20$ because of the unusually high rate of missing values in this example (nearly 80%); with more moderate rates fewer would suffice. After imputation, we computed estimates and information-based standard errors for the five parameters from each imputed data set, then combined the results across each set of $m = 20$ imputations by Rubin's (1987) rules.

The results are shown in Table 5. Comparing the top panel of Table 5 with the likelihood-based results in the top panel of Table 4, we see that MI and ML estimates are very similar. NORM's imputation method is theoretically appropriate for MAR missingness, but some bias is evident under the MAR condition because of the small sample size; if the sample size is increased, the bias disappears. The behavior of

Table 5
Performance of Multiple Imputation (m = 20) Under a Normal Model for Parameter Estimates and Confidence Intervals Over 1,000 Samples (N = 50 Participants)

Parameter	MCAR	MAR	MNAR
Average parameter estimate (with RMSE in parentheses)			
$\mu_Y = 125.0$	124.9 (6.53)	125.3 (17.2)	**151.6** (26.9)
$\sigma_Y = 25.0$	25.9 (5.93)	**28.7** (8.24)	**13.6** (12.1)
$\rho = .60$	.57 (.19)	.45 (.37)	**.35** (.36)
$\beta_{Y\|X} = .60$	.61 (.27)	.59 (.52)	**.21** (.43)
$\beta_{X\|Y} = .60$	.56 (.22)	**.39** (.38)	.66 (.56)
Coverage (with average interval width in parentheses)			
μ_Y	93.5 (26.1)	94.5 (71.5)	**1.8** (19.9)
σ_Y	93.2 (22.1)	96.1 (35.1)	**18.0** (12.8)
ρ	90.8 (0.75)	**86.9** (1.35)	90.4 (1.06)
$\beta_{Y\|X}$	93.6 (1.05)	94.5 (2.18)	**39.4** (0.71)
$\beta_{X\|Y}$	94.9 (0.87)	95.0 (1.29)	91.4 (2.18)

Note. m is the number of imputations. Parameters: μ is the population mean; σ is the population standard deviation; ρ is the population correlation; β is the population regression slope. Coverage represents the percentage of confidence intervals that include the parameter value; values near 95 represent adequate coverage. Use of boldface type in the top panel indicates problematic levels of bias (i.e., bias whose absolute size is greater than about one half of the estimate's standard error); use of boldface in the bottom panel indicates seriously low levels of coverage (i.e., coverage that falls below 90%, which corresponds to a doubling of the nominal rate of error). MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root-mean-square error.

the MI intervals reported in the bottom panel of Table 5, however, shows some improvement over the likelihood-based intervals in the bottom panel of Table 4; in most cases the coverage is closer to 95%. Good performance of MI intervals in small samples has been previously noted (Graham & Schafer, 1999). Increasing the sample size to 250, we found that the results from MI became nearly indistinguishable from the ML results with the same sample size.

General Comments on MI

MI is a relative newcomer, but its theoretical properties are fairly well understood. Because MI relies on Bayesian arguments, its performance is similar to that of a likelihood method that makes similar assumptions. Like ML, MI does rely on large-sample approximations, but some limited experience (e.g., the simulation reported above) suggests that with small to moderate sample sizes these approximations may work better for MI than they do for ML. Of course, MI also requires assumptions about the distribution of missingness. Nearly all MI analyses to date have assumed that the missing data are MAR, but a few MNAR applications have been published (Glynn et al., 1993; Verbeke & Molenberghs, 2000). Nothing in the theory of MI requires us to keep the MAR assumption, and new methods for generating MIs under MNAR models will certainly arrive in the future.

MI uses a model at the imputation phase, so robustness to departures from the model is a concern. In many situations we expect MI to be fairly robust, because the model is effectively applied not to the entire data set but only to the conditional distribution of the missing part. It is also possible to combine a fully parametric MI procedure with postimputation analysis by a robust method, and unless the imputation model is grossly misspecified the performance should be quite good (Meng, 1999). In structural equation modeling, for example, one could multiply impute the missing values under a normality assumption and then fit the structural model to the imputed data using the robust techniques of Satorra and Bentler (1994). Although further study of this is needed, we conjecture that the properties of these hybrid procedures will be excellent, perhaps better than the normality-based likelihood approach of AMOS or Mx.

One important difference between MI and likelihood methods is that with likelihood the missing values are dealt with during the model-fitting procedure, whereas in MI they are dealt with prior to the analysis. When the same model is used for imputation and analysis, MI produces answers similar to those of a likelihood analysis under that same model. Much of the strength and flexibility (and, perhaps, the danger) of MI, however, stems from the interesting possibility of using different models for imputation and analysis. Differences in these two models do not necessarily invalidate the method but may actually strengthen it. With MI the imputer is free to make use of additional data (e.g., extra variables) that do not appear in the analysis, and if those data are useful for predicting missing values, then MI increases power. Properties of MI when the imputer's and analyst's models differ have been explored theoretically by Meng (1994) and Rubin (1996) and from a practical standpoint by Collins, Schafer, and Kam (2001).

Recent Developments

Methods Based on Weighting

The notion of reducing bias due to non-MCAR missingness by reweighting has a long history in the survey literature (Little & Rubin, 1987, chapter 4). Recently, biostatisticians have begun to apply this idea in regression modeling with incomplete covariates. Robins et al. (1994) developed weighted regression that requires an explicit model for the missingness but relaxes some of the parametric assumptions in the data model. Their method is an extension of *generalized estimating equations* (GEE), a popular technique for modeling marginal or population-averaged relationships between a response variable and predictors (Zeger et al., 1988). These models are called *semiparametric,* because they require the regression equation to have a specific form (e.g., linear or log-linear) but beyond that do not specify any particular probability distribution for the response variable itself. The same estimation procedure can be applied whether the response is continuous or discrete. Older GEE methods can accommodate missing values only if they are MCAR; newer methods allow them to be MAR or even MNAR, provided that a model for the missingness is correctly specified. Further results and extensions have been given by Robins and Rotnitzky (1995) and Robins, Rotnitzky, and Scharfstein (1998).

A primary motivation of these weighting methods is to achieve robustness, good performance over more general classes of population distributions. However, extra generality does not come for free. Semiparametric estimators can be less efficient and less powerful

than ML or Bayesian estimators under a well-specified parametric model. With missing data, Rubin's (1976) results show that ML or Bayesian methods perform uniformly well over any MAR missingness distribution, and the user does not need to specify that distribution. However, semiparametric methods that relax assumptions about the data must in turn assume a specific form for the distribution of missingness. We acknowledge that these weighting techniques may be useful in some circumstances. However, as a general principle, we also believe that a researcher's time and effort are probably better spent building an intelligent model for the data rather than building a good model for the missingness, especially if departures from MAR are not a serious concern. In one simulated example, Meng (1999) noted that, for these semiparametric methods to gain a substantial advantage over Bayesian MI, the parametric model had to be so grossly misspecified that only the most "statistically challenged" researcher would use it.

Principles of weighting have also been used to compute MI estimates for fully parametric regression models with missing covariates. Ibrahim (1990) developed a weighted estimation method for generalized linear models, a class that encompasses traditional linear regression, logistic regression, and log-linear modeling (McCullagh & Nelder, 1989). Ibrahim's method is an EM algorithm under a generic model for the joint distribution of the predictors. This method requires the predictors to be discrete and does not take relationships among them into account. With a normal response, Ibrahim's algorithm is a special case of EM for mixed continuous and categorical data considered by Little and Rubin (1987) and Schafer (1997). The method has also been used for survival analysis (Schluchter & Jackson, 1989; Lipsitz & Ibrahim, 1996, 1998). Weighting methods for ML regression with missing covariates were reviewed by Horton and Laird (1999). Although formal comparisons have not yet been made, we expect that, in many cases, these weighting techniques produce answers similar to MI under a suitable joint model for the response and covariates.

Methods That Do Not Assume MAR

Many recent publications focus on MNAR missingness. MNAR is a potentially serious concern in clinical trials, in which participants may be dropping out for reasons closely related to the response being measured. For example, Hedeker and Gibbons (1997) described an experiment in which patients were treated for depression. The response variable was a standardized depression score. Patients who were doing well (i.e., experiencing lower levels of depression) appeared to have higher rates of attrition, perhaps because they believed treatment was no longer necessary. Patients who were not improving also appeared to have higher attrition rates, perhaps because they decided to seek alternative treatment. In these situations, it seems useful to allow the probability of dropout at any occasion to depend on the participant's response at that occasion.

Without the MAR assumption, one must explicitly specify a distribution for the missingness in addition to the model for the complete data. There are two fundamentally different ways to do this: selection models and pattern-mixture models.

Selection models. Selection models were first used by econometricians to describe how the probability of response to a sensitive questionnaire item (e.g., personal income) may depend on that item (Amemiya, 1984; Heckman, 1976). In a selection model, we first specify a distribution for the complete data, then propose a manner in which the probability of missingness depends on the data. For example, we could assume that the logarithm of income is normally distributed in the population and that each individual's probability of responding is related to his or her log-income by logistic or probit regression. Mathematically, a selection model builds a joint distribution for the complete data Y_{com} and the missingness R by specifying a marginal distribution for $Y_{com} = (Y_{obs}, Y_{mis})$ and a conditional distribution for R given Y_{com},

$$P(Y_{com}, R; \theta, \xi) = P(Y_{com}; \theta)\, P(R|Y_{com}; \xi), \quad (7)$$

where θ represents unknown parameters of the complete-data population and ξ represents unknown parameters of the conditional distribution of missingness. The likelihood function is obtained by collapsing this joint distribution over the unknown Y_{mis}, as shown in Equation 3.

Selection models for longitudinal studies with dropout have been reviewed by Little (1995) and Verbeke and Molenberghs (2000). In typical applications (e.g., Diggle & Kenward, 1994), researchers assume that measurements over time $(Y_1, \ldots, Y_T)$ follow a well-recognized distribution such as a multivariate normal and allow the probability of dropout at occasion t to follow a logistic regression on the previous and current responses $(Y_1, \ldots, Y_t)$ but not on future

responses. The procedures required to obtain ML estimates are not trivial, because the likelihood (Equation 3) often cannot be computed directly and must be approximated. The likelihood for these models can be oddly shaped, with large, flat regions indicating that parameters are poorly identified.

Selection models for dropout have intuitive appeal. It is natural to think about how a participant's data values may influence his or her probability of dropping out, a notion that corresponds directly to Equation 7. Moreover, the same factorization is used in the definitions of MCAR, MAR, and MNAR. If we alter the dropout model by setting the logistic coefficients for $(Y_1, \ldots, Y_t)$ or Y_t to zero, we obtain MCAR and MAR versions as special cases. This raises the interesting possibility of "testing" the MAR hypothesis, by seeing whether the confidence interval for the coefficient of Y_t includes zero. As pointed out by Kenward (1998), Little and Rubin (1987, chapter 11), and many others (see, e.g., the discussions following Diggle & Kenward, 1994), results of such tests rest heavily on untestable assumptions about the population distribution, and minor changes in the assumed shape of this distribution may drastically alter the conclusions (Kenward, 1998). Many consider these models to be too unstable for scientific applications and to be more useful for raising questions than generating answers (Laird, 1994).

Pattern-mixture models. As an alternative to the selection model, Little (1993) described an alternative class of MNAR methods based on a pattern-mixture formulation. Pattern-mixture models do not describe individuals' propensities to respond. Rather, they classify individuals by their missingness and describe the observed data within each missingness group. A generic pattern-mixture model can be written as

$$P(Y_{com}, R; \theta, \xi) = P(R; \eta)\, P(Y_{com}|R; \nu), \quad (8)$$

where η denotes the proportions of the population falling into the various missingness groups and ν represents the parameters of the conditional distributions of the data within groups. Estimation of ν always requires some unverifiable assumptions, because a portion of Y_{com} is hidden for every group having missing data; Little (1993) called these assumptions *identifying restrictions.* For a review of pattern-mixture models for longitudinal studies with dropout, see Little (1995) or Verbeke and Molenberghs (2000).

Pattern-mixture models are closely related to multiple-group procedures for missing data in structural equation modeling (Allison, 1987; Duncan & Duncan, 1994; Muthén, Kaplan, & Hollis, 1987). In the multiple-groups approach, observational units are sorted by missingness pattern and each pattern is assumed to provide information about a subset of the model parameters. A model is fit to each pattern, and parameters from different patterns with equivalent meaning are constrained to be equal. Because of the particular form of the constraints used in these published articles, the estimation procedures yielded ML parameter estimates appropriate under MAR. Other types of constraints that produce MNAR models are possible; for a simple example, see Little (1994).

By nature, pattern-mixture models do not posit strong theories about the mechanisms of missingness; rather, they describe the observed responses in each missingness group and extrapolate aspects of this behavior to unseen portions of the data. The likelihood function for a pattern-mixture model, obtained by collapsing Equation 8 over the missing data Y_{mis}, tends to be more convenient to maximize than the likelihood for a selection model. In some classes of pattern-mixture models—the so-called *random coefficients* models used by Hedeker and Gibbons (1997)—ML estimates can be computed by conventional longitudinal modeling software. One inconvenient feature of these models, however, is that the parameters appearing in the formulation (Equation 8) are rarely the parameters of scientific interest. In most cases, we want to describe some aspect of the distribution of Y_{com} (e.g., a treatment effect) in the population of all missingness groups combined. To estimate those parameters, one usually needs to compute a weighted average of group-specific estimates, with weights determined by the relative sizes of the groups in the sample. Alternatively, one may carry out this averaging through MI.

Pattern-mixture models may not suffer from the extreme sensitivity to distributional shape exhibited by selection models, but their assumptions are no less strong. Estimation of population effects is possible only through identifying restrictions, and the observed data provide no evidence whatsoever to support or contradict these assumptions. Proponents of pattern-mixture models (e.g., Little, 1993) have suggested using these methods for sensitivity analysis, varying the identifying restrictions to see how the results change. Detailed examples of pattern-mixture modeling were given by Verbeke and Molenberghs (2000).

Discussion of MNAR methods. MNAR modeling seems worthwhile for clinical studies in which reasons for dropout may be closely related to the out-

comes being measured. In other situations, psychologists should perhaps resist the urge to apply these methods routinely. Simulations by Collins et al. (2001) show that when the true cause of missingness is the response variable itself, failure to account for the MNAR aspect may indeed introduce a sizable bias into parameter estimates. In other situations in which the true cause is not the response but an unmeasured variable that is only moderately correlated with the response (with, say, a correlation of .40 or less), failure to account for the cause seems capable of introducing only minor bias. For many social science applications, we suspect that the former is the exception and the latter is the rule. For example, Graham et al. (1997) described a longitudinal study of drug and alcohol use among middle and high school students. Attrition greatly reduced the sample size over time. Common notions about substance use might cast doubt on results of an MAR-based analysis, because users could be dropping out of school at higher rates than nonusers. Subsequent debriefing of the data collectors revealed, however, that in most cases attrition could be explained by students' moving away or transferring to other schools that did not participate in the study. It was relatively infrequent that dropout could be plausibly related to substance use (Graham et al., 1997). Even if one argues that mobility and substance use are related, it stretches the imagination to believe that the correlation between them could be much greater than .40. Therefore, we are inclined to trust the results of an MAR-based analysis in this example.

When data are collected by self-report questionnaire, it is again natural to speculate whether missingness is caused by the phenomena being measured, especially if the items are of a personal or sensitive nature. If an item pertains to nonnormative behavior, some participants exhibiting that behavior may indeed leave it blank in order to mask their true values, despite repeated assurances of confidentiality. On the other hand, some who do not exhibit that behavior may also skip the item, thinking that the question cannot possibly apply to them. Reasons for nonresponse vary from one person to another. Many of these reasons could be correlated with the item itself, but probably not to the same degree, and perhaps not even in the same direction. Are we to believe that the correlation between the best aggregate measure of "cause" is correlated with our item to a degree of .40 or more, even after accounting for its relationship to other observed covariates? Some methodologists may be afraid to answer this question, choosing instead to leave it blank. Perhaps we are too bold, but in most cases we are inclined to say, "No."

If MNAR attrition is anticipated, researchers may be able to mitigate its effects by simple changes in the study design. For example, at each occasion of measurement, we could ask each participant, "How likely are you to drop out of this study before the next session?" Collecting this additional covariate and including it in analyses may effectively convert an MNAR situation to MAR.

Concluding Remarks

Although other procedures are occasionally useful, we recommend that researchers apply likelihood-based procedures, where available, or the parametric MI methods described in this article, which are appropriate under general MAR conditions. As MNAR methods are incorporated into mainstream software, they, too, become attractive in certain circumstances, particularly for sensitivity analysis. Until then, ML and MI under the MAR assumption represent the practical state of the art.

References

Allison, P. D. (1987). Estimation of linear models with incomplete data. In C. Clogg (Ed.), *Sociological Methodology 1987* (pp. 71–103). San Francisco: Jossey-Bass.

Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods and Research, 28,* 301–309.

Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics, 24,* 3–61.

Anderson, T. W. (1957). Maximum likelihood estimates for the multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association, 52,* 200–203.

Arbuckle, J. L., & Wothke, W. (1999). *AMOS 4.0 user's guide* [Computer software manual]. Chicago: Smallwaters.

Bentler, P. M. (in press). *EQS structural equations program manual* [Computer software manual]. Encino, CA: Multivariate Software.

BMDP Statistical Software. (1992). BMDP statistical software manual. Los Angeles: University of California Press.

Brand, J. P. L. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets.* Unpublished

doctoral dissertation, Erasmus University, Rotterdam, The Netherlands.

Brown, R. L. (1994). Efficacy of the indirect approach for estimating structural equation models with missing data: A comparison of five methods. *Structural Equation Modeling, 1,* 287–316.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1996). *Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs.* Chicago: Scientific Software International.

Clogg, C. C., & Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association, 79,* 762–771.

Collins, L. M., Flaherty, B. P., Hyatt, S. L., & Schafer, J. L. (1999). *WinLTA user's guide* (Version 2.0) [Computer software manual]. University Park: The Pennsylvania State University, The Methodology Center.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods, 6,* 330–351.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics.* London: Chapman & Hall.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39,* 1–38.

Diggle, P. J., & Kenward, M. G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics, 43,* 49–73.

Duncan, S. C., & Duncan, T. E. (1994). Modeling incomplete longitudinal substance use data using latent variable growth curve methodology. *Multivariate Behavioral Research, 29,* 313–338.

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods, 6,* 352–370.

Gelman, A., Rubin, D. B., Carlin, J., & Stern, H. (1995). *Bayesian data analysis.* London: Chapman & Hall.

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with followups. *Journal of American Statistical Association, 88,* 984–993.

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (in press). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Eds.), *Comprehensive handbook of psychology: Vol. 2. Research methods in psychology.* New York: Wiley.

Graham, J. W., & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of followup data. *Journal of Applied Psychology, 78,* 119–128.

Graham, J. W., & Hofer, S. M. (1991). *EMCOV.EXE users' guide* [Computer software manual]. Unpublished manuscript, University of Southern California, Los Angeles.

Graham, J. W., & Hofer, S. M. (2000). Multiple imputation in multivariate research. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multiple-group data: Practical issues, applied approaches, and specific examples* (pp. 201–218). Hillsdale, NJ: Erlbaum.

Graham, J. W., Hofer, S. M., Donaldson, S. I., MacKinnon, D. P., & Schafer, J. L. (1997). Analysis with missing data in prevention research. In K. Bryant, M. Windle, & S. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 325–366). Washington, DC: American Psychological Association.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31,* 197–218.

Graham, J. W., Hofer, S. M., & Piccinin, A. M. (1994). Analysis with missing data in drug prevention research. In L. M. Collins & L. Seitz (Eds.), *Advances in data analysis for prevention intervention research* (pp. 13–63). Washington, DC: National Institute on Drug Abuse.

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 1–29). Thousand Oaks, CA: Sage.

Graham, J. W., Taylor, B. J., & Cumsille, P. E. (2001). Planned missing data designs in the analysis of change. In L. M. Collins & A. Sayer (Eds.), *New methods for the analysis of change* (pp. 335–353). Washington, DC: American Psychological Association.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Annals of Economic and Social Measurement, 5,* 475–492.

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods, 2,* 64–78.

Heitjan, D. F., & Rubin, D. B. (1991). Ignorability and coarse data. *Annals of Statistics, 19,* 2244–2253.

Hogan, J. W., & Laird, N. M. (1997). Model-based approaches to analyzing incomplete longitudinal and failure time data. *Statistics in Medicine, 16,* 259–272.

Hogg, R. V., & Tanis, E. A. (1997). *Probability and statistical inference* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Horton, N. J., & Laird, N. M. (1999). Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research, 8,* 37–50.

Horton, N. J., & Lipsitz, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *American Statistician, 55,* 244–254.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association, 85,* 765–769.

Insightful. (2001). S-PLUS (Version 6) [Computer software]. Seattle, WA: Insightful.

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics, 38,* 967–974.

Jöreskog, K. G., & Sörbom, D. (2001). LISREL (Version 8.5) [Computer software]. Chicago: Scientific Software International.

Kennickell, A. B. (1991). Imputation of the 1989 Survey of Consumer Finances: Stochastic relaxation and multiple imputation. *Proceedings of the Survey Research Methods Section of the American Statistical Association,* 1–10.

Kenward, M. G. (1998). Selection models for repeated measurements for nonrandom dropout: An illustration of sensitivity. *Statistics in Medicine, 17,* 2723–2732.

Kenward, M. G., & Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science, 13,* 236–247.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review, 95,* 49–69.

Laird, N. M. (1994). Discussion of "Informative drop-out in longitudinal data analysis" by P. J. Diggle and M. G. Kenward. *Applied Statistics, 43,* 84.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38,* 963–974.

Lavori, P. W., Dawson, R., & Shera, D. (1995). A multiple imputation strategy for clinical trials with truncation of patient data. *Statistics in Medicine, 14,* 1913–1925.

Lindstrom, M. J., & Bates, D. M. (1988). Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association, 83,* 1014–1022.

Lipsitz, S. R., & Ibrahim, J. G. (1996). Using the EM algorithm for survival data with incomplete categorical covariates. *Lifetime Data Analysis, 2,* 5–14.

Lipsitz, S. R., & Ibrahim, J. G. (1998). Estimating equations with incomplete categorical covariates in the Cox model. *Biometrics, 54,* 1002–1013.

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models.* Cary, NC: SAS Institute.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association, 88,* 125–134.

Little, R. J. A. (1994). A class of pattern-mixture models for normal missing data. *Biometrika, 81,* 471–483.

Little, R. J. A. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association, 90,* 1112–1121.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Liu, M., Taylor, J. M. G., & Belin, T. R. (2000). Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Biometrics, 56,* 1157–1163.

Madow, W. G., Nisselson, J., & Olkin, I. (Eds.). (1983). *Incomplete data in sample surveys: Vol. 1. Report and case studies.* New York: Academic Press.

McArdle, J. J., & Hamagami, F. (1991). Modeling incomplete longitudinal and cross-sectional data using latent growth structural models. In L. M. Collins & J. C. Horn (Eds.), *Best methods for the analysis of change* (pp. 276–304). Washington, DC: American Psychological Association.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models.* London: Chapman & Hall.

McLachlan, G. J., & Krishnan, T. (1996). *The EM algorithm and extensions.* New York: Wiley,

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models.* New York: Wiley.

Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science, 10,* 538–573.

Meng, X. L. (1999, October). *A congenial overview and investigation of imputation inferences under uncongeniality.* Paper presented at International Conference on Survey Nonresponse, Portland, OR.

Multilevel Models Project. (1996). *Multilevel modeling applications—A guide for users of MLn.* [Computer software manual]. London: University of London, Institute of Education.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 55,* 107–122.

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's*

guide [Computer software manual]. Los Angeles: Muthén & Muthén.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (1999). Mx: Statistical modeling (5th ed.) [Computer software]. Richmond: Virginia Commonwealth University, Department of Psychiatry.

Nesselroade, J. R., & Baltes, P. B. (1979). *Longitudinal research in the study of behavior and development.* New York: Academic Press.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London, Series A, 236,* 333–380.

Neyman, J., & Pearson, E. S. (1933). On the problem of most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A, 231,* 289–337.

Raghunathan, T. E., Solenberger, P. W., & Van Hoewyk, J. (2000). *IVEware: Imputation and variance estimation software.* Ann Arbor: University of Michigan, Institute for Social Research, Survey Research Center.

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association, 90,* 122–129.

Robins, J. M., Rotnitzky, A., & Scharfstein, D.O. (1998). Semiparametric regression for repeated outcomes with non-ignorable non-response. *Journal of the American Statistical Association, 93,* 1321–1339.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association, 89,* 846–866.

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

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91,* 473–489.

Rubin, D. B., & Thayer, D. (1983). More on EM for ML factor analysis. *Psychometrika, 48,* 69–76.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* London: Chapman & Hall.

Schafer, J. L. (1999a). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8,* 3–15.

Schafer, J. L. (1999b). NORM: Multiple imputation of incomplete multivariate data under a normal model [Computer software]. University Park: Pennsylvania State University, Department of Statistics.

Schafer, J. L. (2001). Multiple imputation with PAN. In A. G. Sayer & L. M. Collins (Eds.), *New methods for the analysis of change* (pp. 355–377). Washington, DC: American Psychological Association.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33,* 545–571.

Schafer, J. L., & Schenker, N. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association, 95,* 144–154.

Schafer, J. L., & Yucel, R. M. (in press). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics.*

Schimert, J., Schafer, J. L., Hesterberg, T., Fraley, C., & Clarkson, D. (2001). *Analyzing missing values in S-PLUS.* Seattle, WA: Insightful.

Schluchter, M. D., & Jackson, K. L. (1989). Log-linear analysis of censored survival data with partially observed covariates. *Journal of the American Statistical Association, 84,* 42–52.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods, 6,* 317–329.

Song, J. (1999). *Analysis of incomplete high-dimensional normal data using a common factor model.* Unpublished doctoral dissertation, University of California, Los Angeles.

Stata. (2001). *Stata user's guide* [Computer software manual]. College Station, TX: Author.

Statistical Solutions. (1998). *SOLAS for missing data analysis* (Version 1). Cork, Ireland: Author.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions.* New York: Wiley.

Van Buuren, S., & Oudshoorn, C. G. M. (1999). Flexible multivariate imputation by MICE. Leiden, The Netherlands: TNO Prevention Center.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* New York: Springer-Verlag.

Vermunt, J. K. (1997). LEM: A general program for the analysis of categorical data [Computer software]. The Netherlands: Tilburg University, Department of Methodology.

Wothke, W. (2000). Longitudinal and multi-group modeling with missing data. In T. D. Little, K. U. Schnabel, &

J. Baumert (Eds.), *Modeling longitudinal and multiple-group data: Practical issues, applied approaches, and specific examples* (pp. 219–240). Hillsdale, NJ: Erlbaum.

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 30,* 165–200.

Yuan, Y. C. (2000). Multiple imputation for missing data: Concepts and new development. In *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference* (Paper No. 267). Cary, NC: SAS Institute.

Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics, 44,* 1049–1060.

Received July 23, 2001
Revision received January 15, 2002
Accepted January16, 2002 ■