Multiple Imputation for Missing Data: Making the Most of What You Know

MARK FICHMAN
Carnegie-Mellon University

JONATHON N. CUMMINGS
Massachusetts Institute of Technology

Missing data are a common problem in organizational research. Missing data can occur due to attrition in a longitudinal study or nonresponse to questionnaire items in a laboratory or field setting. Improper treatments of missing data (e.g., listwise deletion, mean imputation) can lead to biased statistical inference using complete case analysis statistical techniques. This article presents a simulation and data analysis case study using a method for dealing with missing data, multiple imputation, that allows for valid statistical inference with complete case statistical analysis. Software for implementing multiple imputation under a multivariate normal model is freely and widely available (e.g., NORM, SAS, SOLAS). It should be routinely considered for imputing missing data. The authors illustrate the application of this technique using data from the HomeNet project.

Keywords: *missing data; multivariate analysis; multiple imputation; statistical estimation; Internet use*

Any empirical social or organizational research that makes use of collected data requires the researcher to consider and treat missing data. Many research design choices are conditioned on how missing data will be treated. For example, in determining the sample size to provide adequate statistical power, the researcher will collect data from more than the required sample to adjust for the expected amount of missing

Authors' Note: We are grateful to Bob Kraut and the HomeNet Project for sharing their data and providing guidance and programming assistance to us throughout the project. Robyn Dawes, Don Rubin, and Joe Schafer provided helpful comments and suggestions on an earlier draft of the article. Thanks to Joe Schafer for advice on using NORM. The second author was supported by a Graduate Student Fellowship from Carnegie-Mellon University and the National Science Foundation (Grant No. IRI-9408271). Correspondence concerning this article should be addressed to Mark Fichman, Graduate School of Industrial Administration, Carnegie-Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213-3890; e-mail: mf4f@cmu.edu.

Organizational Research Methods, Vol. 6 No. 3, July 2003 282-308
DOI: 10.1177/1094428103255532
© 2003 Sage Publications

data. In a survey study, missing data may be due to nonresponse. Often, elaborate and expensive follow-up procedures are used to reduce missing data. In an archival study using historic records, missing data may be due to records having been lost, discarded, or erased. Even with the best efforts, we still often find significant levels of missing data in many of our studies when we have completed the data collection.

In this article, we focus on the situation in which a respondent or subject does not provide data on one or more measures out of a set of measures, for example, a survey respondent who answers several questions but not others. This is termed *item nonresponse* (Little & Rubin, 1987). We are not treating the issue in which the respondent is unwilling to answer any of our questions. This type of situation, termed *unit nonresponse*, is the subject of a large literature, exemplified by work on selection bias in econometrics (Berk, 1983; Heckman, 1976) and much work in statistics (Little & Rubin, 1987, 1989).

We then present a statistical case study dealing with some missing data problems using multiple imputation (MI), a statistically sound and disciplined approach. Imputation is a procedure in which missing data are simulated (imputed) given the available information. MI, described in more detail below, is a technique in which several imputations are generated for a given missing data point. Taken together, these several imputations provide both an imputation of the missing data point x and an estimate of the uncertainty of the imputations of x. Any single imputation will not properly capture the uncertainty associated with x. Given the availability of this method, researchers can approach study design and analysis differently, resulting in a more powerful, valid analysis. Our purpose is not to refine or develop the statistical procedure of MI but to introduce its use in the context of a realistic data analysis situation such as one a researcher might encounter in organizational and social research. In effect, we are offering a case study to illustrate the use and benefits of MI as an approach to missing data. We will emphasize two benefits of this approach. First, using an MI procedure provides a general purpose solution to statistical analysis with missing data that is tractable, relatively low cost, and readily accessible to most researchers. Second, using an MI solution can provide more valid estimates of statistical quantities (e.g., means, standard errors, regression coefficients) than are provided by current practices. This claim will be considered both statistically and through a simulation analysis.

Missing Data: The Problem

Ignorability and Missing at Random (MAR)

Whether missing data occurs as a result of unobserved social processes or is planned, the first step necessary to appropriately treat missing data is to systematically characterize it. In doing so, we build on the work of Rubin and his colleagues (Little & Rubin, 1987, 1989; Rubin, 1976, 1977; Schafer, 1997) and employ their framework. The language and framework is very useful for characterizing missing data and essential for analysis of the problem.

Consider a $n \times p$ data matrix Y with n cases for p variables.[1] We first want to characterize the nature of the missing data. Suppose there is a $n \times p$ response matrix R, where $r_{i,j}$ is a missingness indicator.

$$r_{i,j} = \begin{cases} 1 & \text{if } y_{i,j} \text{ is observed} \\ 0 & \text{if } y_{i,j} \text{ is missing} \end{cases} \quad (1)$$

Y has two parts, Y_{obs} and Y_{miss}. Y_{obs} has the elements in Y that are observed ($r_{i,j} = 1$) and Y_{miss} the missing elements in Y. The key assumption that needs to be considered is the nature of the mechanism generating R. One can assume the mechanism is ignorable or nonignorable. Ignorability in turn is a function of two assumptions.

1. The data are missing at random (MAR). MAR intuitively means that the probability that observations are missing may be dependent on Y_{obs} but not on Y_{miss}. That is,

$$p(R \mid Y_{obs}, Y_{miss}, \xi) = p(R \mid Y_{obs}, \xi), \quad (2)$$

where ξ is an unknown parameter of the missingness mechanism. MAR essentially says that the probability of $P(Y)$ being missing may be related to or depend on Y_{obs} but given Y_{obs}, $P(Y)$ is independent of Y_{miss}. Essentially, MAR allows for the possibility that the probability of missingness can be predicted from other available responses. In that sense, the assumption of MAR depends on the data and model for Y (Schafer, 1997).

2. The parameters of the model for the data in Y are distinct from the parameters of the missingness mechanism ξ. That is, ξ tells us nothing about the parameters underlying the data model for Y.

A special case of MAR is missing completely at random (MCAR). MCAR means that no information in Y can predict whether data are missing, or $P(R \mid Y) = P(R)$. MCAR would hold if each element $r_{i,j}$ were determined by a random process like a rolled die. Missing data under MCAR is a random sample of Y. This could happen if the data were missing by design. Little (1988) developed a multivariate test for whether data are MCAR, which is implemented in SYSTAT and the SPSS Missing Values Analysis module. One can compare the distribution of $Y_{obs} \mid R_{k,k \neq j} = 1$ to $Y_{obs} \mid R_{k,k \neq j} = 0$. If there are differences between the two distributions, MCAR is violated. No such tests are available for the MAR assumption.

If data are MAR and the parameters of the data model and missingness parameters are distinct, then the missing data mechanism is ignorable. If this is the case, then we can consider methods for imputing missing data using information in Y_{obs}. What does a nonignorable mechanism look like? Suppose survey respondents that are heavy computer users are more likely to answer questions about their e-mail use. Furthermore, suppose that other variables in the survey do not let us predict who is likely to be a heavy computer user. If this holds, then missing data cannot be imputed with the data in hand; the missingness mechanism is nonignorable. Methods such as selection models have been developed to handle such situations (Berk, 1983; Heckman, 1976) and are not discussed here.

What is important to note is that ignorability is an assumption made by the analyst and depends both on the nature of the missing data mechanism(s) and the data in Y. Experienced analysts in this area emphasize that it is in the data analyst's interest to build as rich a set of data as possible with which to impute missing data (King, Honaker, Joseph, & Scheve, 2001; Schafer, 1997). Whether the missing data mecha-

nism is ignorable depends both on the processes generating the missing data and the information introduced by the analyst in construction of *Y*.

If Y_{obs} contains a lot of information relevant for predicting Y_{miss} and if the data model is sufficiently complex to make use of this information, then we should expect the residual dependence of *R* upon Y_{miss}, after conditioning on Y_{obs} to be relatively minor (Schafer, 1997, p. 28).

How Prevalent Are Missing Data Problems?

Roth (1994) surveyed publications in the *Journal of Applied Psychology* (*JAP*) and *Personnel Psychology* (*PP*) for 1989 to 1991. It was difficult to even determine in many cases what the missing data issues might be. He found that "almost 42% of the articles in the *JAP* sample and 77% of the *PP* analyses involving a survey did not explicitly mention if there was any lack of response to individual items or methods to deal with the issue" (Roth, 1994, p. 548). In his sample of empirical articles, Roth found that only between 23% and 39% of articles did not require attention to missing data. The remaining articles either could not be coded or did require attention to missing data. A conservative estimate from Roth's results is that at least 40% of the papers coded required attention to missing data. Comparable results are found in political science journals over the period of 1993 to 1997 (King et al., 2001). Only 19% of authors explicitly discussed their treatment of missing data (which is a form of missing data in itself)! After following up with authors leading to more information, King et al. (2001) found that 94% of the survey data analyses used listwise deletion (only cases with available data on all variables are kept for analysis, which is the default procedure in most standard statistical packages and analysis), and on average, analysts lost one third of their data. Given that many papers provided insufficient information, the proportion of empirical work requiring attention to missing data may be substantially higher.

These results suggest the following conclusions. First, the problem is widespread. Ironically, the very data needed to assess the extent of the problem in any particular research report is often missing. Although recommendations for improved data reporting to suit meta-analysis have been offered and are beginning to be adopted (e.g., Johnson, Mullen, & Salas, 1995), the most basic information on whether information was available or missing for a particular study is itself missing. Second, when the problem occurs, it can have significant consequences on the cases available for analysis.[2]

Missing Data and Inference—Evaluating Solutions to Missing Data Problems

Missing data presents two kinds of problems: reduced power and threats to the validity of statistical inference. By potentially reducing the number of available observations for analysis, missing data can reduce the power of an analysis, ceteris paribus. If we had originally planned on *n* observations but now have fewer observations to work with, this will reduce the power of any test we wish to apply. If missing data are not MCAR, then statistical inferences may be biased in uncertain ways relative to an analysis when there is no missing data. We briefly discuss these issues as we present the methods most often applied to missing data problems. In our discussion, we will use multiple regression as the exemplar statistical technique, but our remarks should

hold for most other statistical techniques in wide use today (Rubin, 1987; Schafer, 1997).

Of course, it is always best to have no missing data and design research to minimize the probability of missing data. However, even with best efforts, missing data can and will occur. We also must always understand that there is no way to avoid dealing with missing data. Researchers can not ignore missing data, even if they want to ignore it. As soon as data analysis begins, missing data has to be dealt with systematically or else default actions will be taken by statistical routines (usually listwise deletion).

These methods have been reviewed by others (Little, 1992; Little & Rubin, 1989; Little & Schenker, 1995; Roth, 1994), so we only discuss them briefly with respect to the two criteria of valid inference and statistical power.

Little (1992) and Little and Rubin (1987) classified methods for dealing with missing data into several categories. Adapting their classifications, we identify these frequently used methods:

1. complete case analysis—listwise deletion;
2. available case analysis—pairwise deletion;
3. unconditional mean imputation;
4. conditional mean imputation, usually using least squares regression;
5. maximum likelihood; and
6. multiple imputation (MI).

Complete Case Analysis

The default chosen by both analysts and most statistical analysis computer packages is complete case analysis by applying listwise deletion. Only cases having nonmissing observations on all independent and dependent variables are considered for analysis. Listwise deletion when there is even modest missing data can lead to a large percentage reduction in complete cases even when there are a small number of variables in an analysis. Lost cases reduce statistical power. Listwise deletion provides for valid inference (assuming all other assumptions hold) when data are MCAR, because MCAR implies that the complete cases are a random sample from all the cases. The more general case is that listwise deletion provides for valid inference if missing data on the predictor variables does not depend on the response variable.

Available Case Analysis

Complete case analysis clearly loses information that is available in the data. To avoid this cost, one approach is pairwise deletion. This technique is often offered in statistical analysis packages and is often applied when calculating descriptive statistics. For example, one can estimate each correlation in a correlation matrix using all the available cases for each bivariate correlation. If the data are MCAR, then this technique does give more powerful unbiased estimates because all the information in the data are used (Graham, Hofer, & Piccinin, 1994). If there is multicollinearity, there is the risk that the correlation matrix will not be positive definite, which is a requirement for many analysis techniques, including multiple regression. Intuitively, pairwise deletion is troublesome because the sample for each correlation is different. One can have different portions of the correlation matrix being estimated on only partially over-

lapping samples. Little and Rubin (1987, p. 43) provided a useful illustration of the problem.

Standard error estimates from standard computer programs are not correct when pairwise deletion methods are used. One remedy that can be used when missingness is not MCAR is to include the cause(s) of missingness (if they are known) in the analysis. When this technique is applied to correlations and covariances estimated using available case methods, pairwise methods show significant bias (Graham et al., 1994, Table 9) from true values.[3]

Unconditional Mean Imputation

A simple assumption to make is that given no information, the best imputation for missing data is the mean value for that variable—unconditional mean imputation. This is a poor choice, even under MCAR assumptions. The variances are underestimated in proportion to the fraction of missing data. If the data are MCAR, the bias is proportional to $(n_{obs} - 1)/(n_{obs} + n_{miss} - 1)$. The covariance is biased by a similar factor (Little, 1992). Because the unconditional mean imputation for the missing cases has a variance of 0, the estimated variances and covariances will be underestimated. In effect, one's uncertainty in the imputations is understated. It is as though we had more confidence in our imputations for missing data than we have in the data that is not missing (King et al., 2001)! Clearly, listwise deletion, pairwise deletion, and unconditional mean imputation are not acceptable general purpose solutions for missing data problems.

Conditional Mean Imputation

Where filling in with means is unconditional mean imputation, one can generate imputations that are conditional on other information in the data. Essentially, one can construct a regression estimate for a missing data element based on all the available data. Little (1992) reviewed various methods, noting that "estimated standard errors of the regression coefficients from OLS (Ordinary Least Squares) or WLS (Weighted Least Squares) on the filled-in data will tend to be too small, because imputation error is not taken into account" (Little, 1992, p. 1232). Imputation error is the error in the estimates due to uncertainty about the imputed value. Little and Rubin (1987) suggested that a stochastic regression imputation technique is preferable. This method is an improvement by better capturing the distributional characteristics of X but still underestimates the standard errors and variability due to imputation.

The flaw with these conditional mean imputation procedures is that the analyst is confusing two issues, maximizing the accuracy of their imputation of a missing value with valid statistical inference (Rubin, 1996). Although conditional mean imputations may generate accurate values, the uncertainty or imputation error in those imputations is not properly estimated. Valid statistical inference requires unbiased estimates of quantities of interest and that confidence interval nominal levels are less than or equal to actual levels. This second condition implies that for hypothesis tests, we want nominal rejection rates to be less than or equal to actual rejection rates. Rubin (1987, 1996) termed this criterion *confidence validity.*

Maximum Likelihood (ML)

ML methods, particularly using the expectation maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977), have been proposed and implemented for many missing data problems. These methods are reviewed elsewhere (Little & Rubin, 1987, 1989; Little & Schenker, 1995) and will not be reviewed here. ML methods are preferable to conditional MI using least squares methods because the estimated parameters are consistent and efficient under MAR conditions. If the MAR assumption is met, estimates of other parameters such as standard errors are also unbiased.

Multiple Imputation (MI)

Rubin's (1987) proposal of MI was motivated by both statistical and pragmatic concerns. MI was designed to address several real, practical issues. The primary motivating factor was that many data sets in social science are public use data sets. Consequently, each investigator using a public use data set must make a decision on how to treat missing data. In some cases, users may not have access to information to which the primary investigators have access. Such private information held by primary investigators (e.g., name, address, and other personal information that are withheld from the public) may help the investigator solve some missing data problems that cannot be released to others. Primary investigators cannot anticipate all the considerations individual investigators have and consequently cannot provide exactly the right solution to each investigator's missing data issues.

Rubin (1987) proposed that a general purpose solution to this problem would be for the primary investigators to generate missing data imputations that would then be available to others who would then not have to repeatedly solve the missing data problems on their own with less than full information. These imputations would have the property of being complete case data sets that allow each investigator to proceed with analysis without having to develop a missing data imputation solution tailored to their particular analysis. A second benefit of this solution would be that analyses by different investigators would surely be based on the same data set and thus be comparable. If different investigators used different procedures, their analyses could not be readily compared. The overall efforts of the social science community would be conserved because the missing data problem would be treated once with the best available information.

One can also apply this argument to the individual investigator analyzing a data set. That investigator should have a general purpose tool that allows him or her to treat the problem of missing data once and then be able to proceed with analysis using statistical tools designed for complete case analysis. In either the public use or the individual investigator case, we also want the researcher to be able to distinguish between error due to sources such as measurement from error due to imputation. In methods such as mean imputation or regression imputation, we do not consider imputation uncertainty, whereas in MI we do. Given these objectives, MI becomes an attractive procedure for dealing with missing data imputation issues.

In general, the situation facing a researcher who wants to address missing data problems is as follows: Many simple fixes reviewed above are inadequate to the task

and can produce biased and incorrect analyses. Most reviewers of the missing data literature conclude that there are two approaches to these problems that are satisfactory. One is the use of ML procedures using specialized algorithms (e.g., the EM algorithm) (e.g., Little & Schenker, 1995; Schafer, 1997) or structural equation modeling software. The other is the use of the MI simulation methods reviewed here. Clearly, neither method has become accepted practice in social and organizational research. We reviewed results earlier showing that few researchers, if any, treat missing data in a principled way now. Our view, consistent with others (Allison, 2000; King et al., 2001; Rubin, 1987; Schafer, 1997, pp. 134-145) is that MI offers some appreciable advantage for the applied data analyst.

When the analyst is going to do a series of analyses and the data analysis plan is somewhat uncertain and open-ended, MI is very useful. Although ML and MI methods are comparable, given current software implementations, MI provides a more general purpose solution to the problem of missing data (Collins, Schafer, & Kam, 2001), particularly with smaller samples (Schafer & Graham, 2002). ML methods often require specialized solutions for each model to be estimated, with higher attendant costs in both the programming and estimation of such models. For MI, there will be no need for specialized software other than the imputation software. All analyses can be done using complete case methods. This is probably the strongest argument for MI methods. The lack of implementation of ML methods speaks to the practical difficulties analysts encounter in implementing such procedures precisely because they are not general purpose. When the analyst has a particular focused question, the specialized software to do the estimation of the appropriate model, and is confident that the model assumptions are met, ML methods can be very attractive. Unfortunately, these conditions often do not hold in social and organizational research. Schafer and Graham (2002) provided a thoughtful discussion of the two approaches, concluding that for many purposes when working with data under the MAR assumption, both ML and MI are appropriate ways to deal with missing data.

Imputation methods are often robust to violations of assumptions in the underlying model. "In many realistic scenarios, multiple imputation may tend to be more robust to departures from the data model than parameter simulation" (Shafer, 1997, p. 136). Allison (2000) presented simulation results showing that MI is robust to model violations, whereas some alternatives (listwise deletion, propensity scoring) are not. King et al. (2001) reported simulation results showing that MI works well, even in situations where the assumptions of MI are violated; there is nonignorable missingness. In both the Allison (2000) and King et al. (2001) simulation results, listwise deletion sometimes does very poorly.

MI: A General Purpose Solution

In single imputation methods, we replace missing values with imputations, yielding one complete data set that can be analyzed using complete case statistical methods. With MI, a data set is composed of the observed values and MIs for the missing values generated using some procedure. This is done *m* times, with each data set having the same observed values and different imputations for the missing observations, yielding *m* complete data sets. The variability across the *m* data sets for the imputations reflects our uncertainty about the imputations.

Estimating Quantities From MIs

Rubin (1987) provided a procedure for combining the estimated quantities (e.g., regression coefficient estimates or means) from the *m* imputations. We use Schafer's (1997) notation. Consider a quantity $\hat{Q}$ with an estimated variance *U*. After analysis of the *m* data sets, there are now *m* estimates of $\hat{Q}$ and *U*. A combined estimate is

$$\overline{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i. \tag{3}$$

There are two components to the variability of $\overline{Q}$. The within imputation variance,

$$\overline{U} = \frac{1}{m}\sum_{i=1}^{m}\hat{U}_i, \tag{4}$$

and the between imputation variance,

$$B = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{Q}_i - \overline{Q}\right)^2. \tag{5}$$

T is the total variance, which is the sum *B* and $\overline{U}$ corrected for *m* being finite by the term $\frac{m+1}{m}$.

$$T = \overline{U} + \frac{m+1}{m}B. \tag{6}$$

$\frac{B}{U}$ indicates how much information is missing. It is an estimator of $\frac{\gamma}{1-\gamma}$ where γ is the fraction of information missing for *Q* due to nonresponse (Little & Rubin, 1987, p. 257). If γ is zero, then *B* goes to zero. Confidence intervals are calculated as

$$\overline{Q} \pm t_{df}\sqrt{T}, \tag{7}$$

where the *t* distribution degrees of freedom *df* are calculated as

$$df = (m-1)\left(1 + \frac{1}{r^2}\right), \tag{8}$$

with

$$r = \left(1 + \frac{1}{m}\right)\frac{B}{\overline{U}}, \tag{9}$$

which indicates the relative increase in variance due to nonresponse. Equation 8 (Rubin, 1987) shows that the degrees of freedom are influenced by both m and $\frac{\overline{U}}{B}$. As m increases, df increases. As $\frac{\overline{U}}{B}$ increases, df gets larger.

> If the computed value of df is very small—say, less than 10—it suggests that greater efficiency (i.e., more accurate estimates and narrower intervals) could be obtained by increasing the number of imputations m. If df is large, however, it suggests that little will be gained from a larger m. (Schafer & Olsen, 1998, p. 19)

The fraction of information about $\overline{Q}$ missing due to nonresponse is

$$\gamma = \frac{r + \frac{2}{df+3}}{r+1}. \tag{10}$$

One can use γ and r as diagnostic statistics to examine the impact of missing data on estimates of $\overline{Q}$. Calculation of the quantities in Equations 3 through 10 can be easily implemented as a set of macros in most statistical packages or in a spreadsheet.[4]

Determining How Many Imputations Are Needed

Working back from the computation of statistics and confidence intervals, we need to ask how many imputations are required to generate useful statistical inferences. Surprisingly, the answer is often no more than five and sometimes as little as two or three. The relative efficiency (RE) of the m imputations given a fraction of missing data (Rubin, 1987, p. 114) is computed in standard error units as

$$RE = \left(1 + \frac{\gamma}{m}\right)^{-\frac{1}{2}}. \tag{11}$$

Even when γ is .9, five imputations give a relative efficiency of .92. Consequently, one rarely has to create more than five imputations, and 10 imputations are more than suitable in almost any realistic application.

MI Using Markov Chain Monte Carlo (MCMC) Methods

Given that we know how to combine the estimates from MIs and how many imputations to estimate, what remains is to determine the proper way to generate imputations. This has been the computationally difficult problem that has prevented the more widespread diffusion of MI methods. Schafer (1997) provided an excellent exposition of these methods. Schafer also provided software for several of these methods (see Note 4).[5]

Assuming nonresponse is ignorable (the MAR assumption), Rubin (1987) developed a Bayesian method for attacking the problem. The approach is to simulate the missing data under a set of assumptions. For MI to allow valid inference, the imputations must be proper. Rubin developed the definition of proper imputation, which can be viewed from either a frequentist or Bayesian perspective, even though the method

of imputation is Bayesian. A more detailed analysis of proper imputation with proofs and definitions is offered in Rubin (1987, pp. 118-128; 1996, pp. 477-478).

To generate imputations, we impute values for Y_{miss} given Y_{obs}. Assuming missingness is MAR,

$$P(Q \mid Y_{\text{obs}}) = \int P(Q \mid Y_{\text{obs}}, Y_{\text{miss}})\, P(Y_{\text{miss}} \mid Y_{\text{obs}}) dY_{\text{miss}}. \tag{12}$$

We can use simulation to approximate this integral (Rubin, 1987). It is important to note that MI does not require an ignorability or MAR assumption. We make the ignorability assumption when it seems reasonable to simplify the problem of specifying a mechanism for nonresponse. The posterior distribution of Q is an average over the repeated draws from $P(Y_{\text{miss}} \mid Y_{\text{obs}})$, the posterior predictive distribution of the missing data given the observed data. Because the imputations do not rely on response matrix R, we are treating nonresponse as ignorable or MAR. Schafer treated these results as Bayesianly proper, defined as MIs that are independent draws from $P(Y_{\text{miss}} \mid Y_{\text{obs}})$ (Schafer, 1997, p. 105).

The MIs treated here are repeated imputations, repeated draws from the posterior predictive distribution $P(Y_{\text{miss}} \mid Y_{\text{obs}})$. For imputations to be proper, they must include all sources of modeling uncertainty, including B, between imputation variability. Schafer (1997) offered data augmentation (Tanner & Wong, 1987) as a method for generating Bayesianly proper imputations that include B. He developed several adaptations. We are interested in the adaptation for the normal model, in which we assume the data to be multivariate normal. To meet these assumptions, the analyst should check the distribution of the variables, transforming the data as needed to meet the multivariate normal assumption.

Data augmentation involves two steps. The first imputation, or I-step, is

$$Y_{miss}^{(t+1)} \sim P\left(Y_{miss} \middle| Y_{obs}, Q^{(t)}, U^{(t)}\right), \tag{13}$$

where t is an indicator of the time ordering of the computational steps. The first iteration of the I-step requires a prior provided by the analyst (which can be a noninformative prior in many applications.). That is, the analyst has to give an initial estimate of $Q^{(t)}$, $U^{(t)}$. Methods for more informative priors have been developed (Schafer, 1997). After the first iteration, new values are drawn from the posterior distribution, which will have observed values and imputed values. The posterior, or P-step, is

$$Q^{(t+1)}, U^{(t+1)} \sim P(Q, U \mid Y_{obs}, Y_{miss}^{(t+1)}). \tag{14}$$

Each of these steps is a simulation of the data. It is a MCMC procedure because each step depends on the previous step, introducing dependency across the steps (hence a Markov chain), and it is a Monte Carlo procedure because we start with a random imputation of missing data. For large samples, a noninformative prior (in the normal model, the limit of a normal inverted Wishart distribution) can be used. For smaller

samples or samples with sparse observations in some portions of the data matrix, an informative prior may be introduced.

Imputations can be drawn from one Markov chain, once there has been a sufficient number of steps (termed *burn-in*) so the distribution has stabilized. The number of burn-in iterations may be larger than the iterations required to generate the imputed data sets themselves. There is disagreement in the statistical literature for MCMC about whether to use one Markov chain to draw MIs (which introduces the risk of dependency between the data sets) or running m independent chains. If one used one Markov chain, one would choose some k sufficiently large (e.g., 500) such that one would draw from the distribution only after it has stabilized and at that point, draw after every k cycles of the imputation step–posterior step (IP) procedure. One can examine diagnostic autocorrelation functions (ACF) to see if the autocorrelation across iterations is sufficiently low to treat the draws from one Markov chain are independent; m independent Markov chains are preferable because there is no autocorrelation by construction, but the cost is running $m - 1$ additional MCMC simulations using the IP algorithm. As computation costs decline, this becomes less of an issue. This trade-off is probably best addressed by running independent chains. Independent chains also should give the analyst a more reliable estimate of error due to simulation (Monte Carlo error) (Schafer, 1997). Running m chains also avoids examination of as many ACF charts because one does not have to assess autocorrelations as often.[6]

An Illustrative Application of MI-Internet Use in HomeNet[7]

Overview of the Study

In this section, we want to illustrate the application of MI in a set of data. This data has many of the kinds of missing data problems one encounters in a field setting. In this study, Internet use was investigated by Kraut et al. (1998). They reported that people were using the Internet at home to communicate with family and friends but, at the same time, were experiencing a reduction in social involvement and psychological well-being—the Internet paradox (Kraut et al., 1998). Using a panel research design, 363 individuals in 93 families were given Internet access for the first time and tracked over the course of 1 or 2 years. Greater use of the Internet was associated with a decrease in household communication and size of social circle as well as with an increase in depression and loneliness (Kraut et al., 1998).

Sample

Participants were selected for the HomeNet project through two means at two points in time (both excluding households or individuals with active Internet use). First, families with a teenager in journalism class in four area high schools were recruited to participate. Journalism students would have good reason to use the Internet, including searching the Web for article research and using e-mail to communicate with other staff members and newspaper editors at other schools. Second, and 1 year later, families with an adult on the board of directors of four community development organizations were asked to be involved. Adults active in the community could also use the Internet to search Web sites for information about their cause and use e-mail

to communicate with other staff members and community leaders of other organizations. Greater than 90% of the families in both groups agreed to take part in HomeNet.

Family members were counted as participants if they gave informed consent, had at least one login to the Internet, and were older than the age of 10 (90% of the family members participated). Families were given a Macintosh computer with software (ClarisWorks Office, Netscape Navigator, MacMail II-Carnegie Mellon University proprietary software), a free extra telephone line, and free Internet access (with an e-mail account for each family member). Training was also available to the families, because this was the first experience with a powerful home computer for many of them. At least two family members went to a training session on the use of the Web, e-mail, and computers in general. Furthermore, a help desk phone number was available to them for any problems they may have encountered.

Data were collected from participants in two ways. First, periodic mail questionnaires were sent to each family member participating in the study (including a pretest questionnaire before actual Internet use). Second, custom-designed computer logging programs automatically tracked Internet use (including connection hours, number of Web sites visited, and e-mail messages sent and received). Because families with an adult on the board of directors started the HomeNet project 1 year after families with a teenager in journalism class, Internet use for them included an additional 52 weeks. After 1 and 2 years, respectively, all participants filled out a posttest questionnaire.

Sample and Sample Size

Of the 363 study participants, 256 respondents from 93 families (a) completed the initial questionnaire, (b) were older than age 10, and (c) logged onto the Internet at least once; 169 respondents from 73 households who met these criteria also completed the follow-up questionnaire. Because we have some information other than survey data for all 363 individuals, we can use all 363 study participants, even those who did not fill out a survey.

We have information on family membership and Internet use. These are the samples and their sizes for the analyses reported in Kraut et al. (1998). To obtain 169 respondents with complete data, Kraut et al. did do some imputation (discussed below). With no imputation and applying complete case analysis procedures, the regression results reported in Kraut et al. (1998) would have sample sizes in the range of 55! In this reanalysis with MI, we use all 363 available respondents.[8]

Measures

Two underlying constructs were measured in this study: social involvement and psychological well-being. Social involvement consisted of (a) family communication,[9] (b) local social circle, (c) distant social circle, and (d) social support (SUPPORT1). Psychological well-being consisted of (a) loneliness, (b) stress (STRESS1), and (c) depression (DEPRESS1).

Family communication was measured as the average number of minutes family members communicated per day with other members of the family. Local social circle was estimated as the average number of people in their local social circle (i.e., "the number of people in the Pittsburgh area whom you socialize with at least once a month"). Distant social circle was estimated as the average number of people in their

distant social circle (i.e., "the number of people outside of the Pittsburgh area whom you seek out to talk with or visit at least once a year"). Finally, social support was measured using the Interpersonal Support Evaluation Checklist (Cohen, Mermelstein, Kamarck, & Hoberman, 1984), a 16-item self-report inventory on how easy it is to find tangible help, advice, emotional support, companionship, and sense of belonging.

Loneliness was measured using the UCLA Loneliness Scale (Russell, Peplau, & Cutrona, 1980), which contained three items used to assess feelings of connectedness to others. Stress was measured using the Hassles Scale (Kanner, Coyne, Schaefer, & Lazarus, 1981), which was a checklist of 49 possible daily life stressors. Depression was measured using the Center for Epidemiologic Studies Depression (CES-D) Scale (Radloff, 1977), which assesses mild depression in the general population with 15 items.

Control measures were age (TEEN), race (WHITE), gender (FEMALE), and household income (INCOME) for this analysis. Age was collapsed to two categories, teen and adult. Race was collapsed to two categories (White and minority). In our imputation analyses reported below, imputations using these measures used the five race categories and the original measure of age. When we then replicate the regression analyses reported in Kraut et al. (1998), we apply the same collapsing method they applied, so that our analyses make the best use of the available information in the imputation. We apply exactly the same regression model and data treatment to ensure comparability between our analyses and the Kraut et al. (1998) analyses. In all cases in which there were ordered categorical variables such as income (five categories) or binary variables such as gender, we did the imputation as though the scores were on a continuous scale. We then rounded the imputed score to the nearest integer value. To meet the multivariate normal requirements underlying the model, we transformed all variables so they would be approximately normal.[10] Log transforms were taken for INTERNET to meet assumptions of normality in both the imputations and regressions.

Analysis Methods

Kraut et al. (1998) recognized that the missing data problems in their data were severe and had to be addressed. In this analysis, as in any analysis in which missing data are an issue, the first step to take is to assess the pattern and severity of missing data problems. If there are no missing data or minimal (less than 5%) missing data, one might consider not imputing data or using the EM algorithm. SAS (Proc MI), BMDP (AM program), SPSS (MVA), and Systat (Correlation procedure) provide facilities for assessing the amount and pattern of missing data. We used MVA in SPSS. Examination of these tables indicated extensive amounts of missing data.

After assessing the pattern of missing data, we assessed whether the data were MCAR using a test statistic developed by Little (1988). For the entire sample of 363 cases, the missing data were not MCAR ($\chi^2 = 1032.52$, $df = 863$, $p < .001$). Because the missing data are not MCAR, a complete case analysis may not provide unbiased estimates of regression coefficients or other statistics in this sample.

Kraut et al. (1998) chose the following strategy to deal with missing data. They chose to only impute data where it was missing for the control variables. This means that those who did not complete any of the independent or dependent variables at either time period would be excluded from the analysis. Only a missing independent variable

datum that was classified as a control variable was considered for imputation. They first used deductive reasoning, field staff knowledge, and their own judgment to provide plausible values for missing data. This may sometimes be appropriate, particularly if the investigators have very good information with which to make such deductions. However, great care should be used in applying such a procedure, as proper accounting for the uncertainty of the investigators deductions is not likely.[11] For data that was still missing in the control variable set, they then applied the EM algorithm to impute control variable missing data.[12] Failure to complete the posttest questionnaire given their procedures automatically led to exclusion from the study as well. Of 363 participants, this led to analysis on 169 (53.4% missing) cases, a substantial loss of data. This is comparable to the type of attrition and loss one might find in a multiyear panel study. For these 169 cases used in the reported analysis, the data were still not MCAR using the Little (1988) test ($\chi^2 = 604.416$, $df = 531$, $p = .015$).

The results reported in Kraut et al. (1998) reproduced in Table 1 (in the 1998 Results columns) are based on the data after their imputation method was applied.[13] This method will likely yield standard errors that are too small because there is not a proper accounting for imputation uncertainty.

We applied MI using NORM, a computer program for MI (Schafer & Olsen, 1998). We decided to generate 10 imputations. If $\gamma \approx .5$ where γ is the fraction of missing data (see Equation 10), this gives us an efficiency of .98. Even for $\gamma \approx .9$, relative efficiency is .96; γ in this data set averages about .50 for each parameter, so relative efficiency is probably at .98. For the imputations, we had 32 variables and 363 cases.

We selected all the analysis variables and several additional variables. We considered several issues in selecting variables for inclusion. There is no necessary reason why only the variables for a particular analysis should be included in the imputation step. However, it is necessary that all the variables used in the *m* complete case data analyses be in the imputation. Otherwise, one introduces the risk of omitted variable bias to the analysis. Under some conditions, even with improper imputation due to omissions, one can still obtain confidence-valid inferences from repeated imputations (Rubin, 1996, p. 479). As a general rule, one should include all variables to be used in analysis in the imputation. However, there is no reason to exclude other variables in the imputation that are not used for analysis.

Rubin (1996) developed the concept of superefficiency. By superefficiency, he meant that if the imputation model to estimate $\bar{Q}$ uses information that is not used in the analysis, then the estimate of some quantity $\bar{Q}$ from the imputation may be more efficient than an estimate using the analyst's model alone as the basis for imputation. In general, adding information to the imputation model will increase the efficiency of estimation of $\bar{Q}$. We added mean family scores for the 93 families for the seven dependent variables Kraut et al. (1998) analyzed to the imputation model. This should not be taken to suggest that we can generate more efficient estimates than with the complete data set, only that additional information will make the imputations relatively more efficient than imputations based on the analysis model alone.

In this case, we introduced information about the family that we would not use in the analysis itself. We included them as family means rather than dummy variables to facilitate the simulation; 93 families would have meant 92 additional regressors, which would have made the MI simulation infeasible. Adding the family means for the seven dependent variables added only seven regressors, making the simulation feasible without changing the information in the data. The individual family intercepts capture the

Table 1
Effects of the Internet on Depression

	Internet Hours			*Depression*		
Independent Variable	*Full Sample*	*1998 Results*	*Imputed 1998 Sample*	*Full Sample*	*1998 Results*	*Imputed 1998 Sample*
INTERCEPT	.00	–.01	.00	.00	.03	.00
	(1.00)		(1.00)	(1.00)		(1.00)
INCOME	–.004	–.01	–.01	–.03	–.04	–.06
	(.54)		(.80)	(.42)		(.60)
WHITE	.16***	.12	.21***	–.03	–.21**	–.03
	(.06)		(.08)	(.47)		(.65)
TEEN	.16***	.15	.26****	.02	.09	.05
	(.03)		(.01)	(.64)		(.59)
FEMALE	–.11	–.02	–.08	.04	–.03	.03
	(.08)		(.31)	(.48)		(.71)
SUPPORT1	.11	.01	.04	–.02	–.03	–.06
	(.12)		(.68)	(.70)		(.55)
STRESS1	–.04	–.13	–.13	.07	.17*	.16**
	(.43)		(.12)	(.54)		(.08)
DEPRESS1	.09	.06	.10	.16	.22**	.20****
	(.35)		(.33)	(.20)		(.05)
INTERNET				.01	.19**	.14
				(.56)		(.11)
R^2	.10	.07	.16	.08	.19	.15
n	363	155	169	363	150	169

Source. The table is adapted from Kraut et al. (1998). Kraut et al. did not report standard errors. Full sample results use multiple imputation; 1998 results are from Kraut et al. Imputed 1998 sample results are for the Kraut et al. respondents using multiple imputations.
Note. Entries are standardized ordinary least squares coefficients. Standard errors are in parentheses.
*$p < .10$. **$p < .05$. ***$p < .01$. ****$p < .001$.

same information as the individual family dummies. In doing this, we made a decision to use an alternative imputation method, conditional mean imputation, for estimating family means. This is a pragmatic choice that in our judgment adds substantial information without violation of the assumptions of the modeling.

What did we do? For each family, we calculated the mean score from the available cases for each of the seven dependent variables. If no score was available, the family mean was coded as missing and imputed by using NORM. An alternative would be to not use the dependent variable means, impute missing data in NORM, and forego using the additional family-level information. We judged that there is benefit (possible superefficiency) in using these means for imputation rather than giving up the additional information. We know from theory that there should be substantial between-family differences on any of these dependent variables, so the information will sharpen our estimates when we analyze the data. The key point here is that all the analysis variables should be included and adding additional variables can be beneficial by increasing efficiency (Meng, 1994; Rubin, 1987, 1996). By including more variables in the imputation model, we ease the analysis task, because we can introduce and remove any variable that was included in the original imputation model. One could use a specific

model allowing for a hierarchical linear model (Schafer, 1997), but we chose to impute using a multivariate normal implementation both for ease of computation and comparability with the original 1998 analysis.

In conducting this case study, as in any data analysis, there are data analysis judgments. To provide a realistic case study, we have explicitly chosen to replicate the Kraut et al. (1998) results to facilitate ready comparison between those results and those developed using MI methods. If we chose a different set of models (which one could argue for), the reader would be unable to readily compare the MI results with those reported using more standard approaches to missing data. Because our primary purpose in this case study is to help readers judge the relative value of traditional and MI methods, we have hewed very closely wherever possible to the decisions made by the original investigators.[14]

The multiple imputations were done using scale scores. For example, the Social Support Scale has 16 items. Use of multi-item scales is very common in social research. If we did imputation at the item rather than the scale level, this would create a tremendous computational burden and would be infeasible in this study and many others. We used scale means (implicitly doing conditional mean imputation, because scale means are calculated in SAS and other packages by using only available items for calculation of scale means for each individual). There are two lines of argument that we offer to defend this choice. First, within scales, if a person responds, there is almost no missing data for that respondent. For example, in the 16-item Social Support Scale, we classified respondents into three categories. They could respond to all the items, none of the items, or some of the items; 163 respondents (44.9%) responded to all items, 196 respondents (54%) responded to no items, and 4 (1.1%) responded to some items. Consequently, there should be little or no empirical consequence of using means. Essentially, missing data occurs for this analysis when someone skips an entire section, not items within scales. Second, a Monte Carlo study of missing data in multiple-item scales (Roth, Switzer, & Switzer, 1999) found that under a broad range of conditions, using the average of nonmissing items to impute missing scale items led to extremely low levels of bias (e.g., correlations changed in the third decimal place in conditions similar to the ones we use). This result is very encouraging.

Monte Carlo Simulation

To ensure that our imputation model outlined above was appropriate, we conducted a simulation using the 169 cases from Kraut et al. (1998). Following previous missing data studies employing simulation, we varied three factors: (a) method for handling missing data (MI vs. listwise deletion), (b) data missingness mechanism (MCAR, MAR, and nonignorable), and (c) percentage of data missing (10%, 30%, and 50%). These past studies have shown that on average, MI outperforms listwise deletion (Allison, 2002; King et al., 2001); the advantage is particularly strong when the missingness mechanism is MAR rather than MCAR or nonignorable (Allison, 2000; Russell, Stern, & Sinharay, 2002), and estimates from all methods for handling missing data get worse as the percentage of missing data increases (Collins et al., 2001; Graham & Schafer, 1999). For brevity, we only report on the simulation showing the relationship between Internet use and depression, with missingness dependent on social support under the MAR condition (see design below).

To run the simulation, we generated multivariate normal data from the variance/covariance matrix of the 169 cases (recall that EM imputation on the control variables was already done by Kraut et al., 1998, so for comparability, we use these same 169 cases as the basis for our simulation). We used the same superefficiency technique discussed above of adding family means on several dependent variables in this simulation. A total of 1,000 samples were generated for each cell in the 2 (method) × 3 (missingness mechanism) × 3 (percentage missing) design. MI was conducted on the simulated datasets using SAS (proc MIANALYZE), and the regression analyses for both MI and listwise deletion were identical to those reported in Kraut et al. (1998). For the MCAR missingness mechanism, data were deleted randomly from depression and social support; for MAR, missingness on depression depended on social support; and for nonignorable, missingness on depression was dependent on the value of depression itself such that the lowest (.1, .3, or .5) fraction of the sample was set as missing. The percentage of missing data was removed according to missingness mechanism and was either 10%, 30%, or 50% of the dataset.

Means, standard deviations, standardized bias, root mean square error (RMSE), and confidence interval coverage were computed for the estimates of the regression analyses (see Table 2). Overall, we see patterns similar to results reported in earlier simulations. For the missingness mechanism MAR, MI is clearly better than listwise deletion, and MI and listwise are comparable for the MCAR condition (with MI's lower RMSE implying better efficiency). For the nonignorable condition, the results are similar. Clearly for our data, MI is a better choice.

Results

Table 1 reports our results and the Kraut et al. (1998) regression model results (1998 Results column). We compare the 1998 results with two sets of imputation results. First, we report the same model estimated with the full data set of 363 observations (Full Sample column). In comparing the full sample of 363 with the 1998 results, we are confounding differences due to imputation with differences due to different samples. For this reason, we also report the results for the subsample used to estimate the 1998 results based on our imputation (Imputed 1998 Sample column). The imputed 1998 sample is those respondents used in the Kraut et al. (1998) analysis with missing data simulated using MI. This allows a clean contrast of the Kraut et al. results with our analysis using MI. The results in Table 1 are representative of the entire set of results.

Depression

Table 1 presents the most provocative results from Kraut et al. (1998), suggesting that increased Internet use at Time 1 (T1) was associated with increased depression scores at Time 2. In the 1998 Results column, there is an estimated $\beta = .19$, $p < .05$, suggesting that increased Internet use was associated with higher depression scores. This is of great theoretical and policy interest and has provoked significant reactions and controversy (Kiesler & Kraut, 1999). The full sample estimate is $\beta = .01$, *ns*, and the imputed 1998 sample is $\beta = .14$, *ns*. This suggests that when one considers the entire HomeNet sample, there is no clear evidence of Internet use affecting depression levels.[15] The other substantive change in the regression models predicting depression is that in the 1998 results, there is a race effect ($\beta = .21$ for the WHITE variable), whereas

Table 2
Effects of Missingness Mechanism, Rate of Missingness, and Using Listwise Deletion or Multiple Imputation (MI) on Mean, Standard Deviation, Standardized Bias, Root Mean Square Error (RMSE), and Coverage of Simulated Estimates of Internet Hour Effects on Depression

	10% Missing		*30% Missing*		*50% Missing*	
	Method		*Method*		*Method*	
	Listwise	*MI*	*Listwise*	*MI*	*Listwise*	*MI*
Missing completely at random						
Mean	0.08	0.08	0.08	0.08	0.08	0.08
Standard deviation	0.03	0.03	0.04	0.03	0.06	0.04
Standardized bias	0.88	–0.02	1.97	1.44	1.68	1.84
RMSE	0.03	0.03	0.04	0.03	0.06	0.04
Coverage	0.95	0.95	0.95	0.94	0.96	0.93
Missing at random						
Mean	0.07	0.08	0.05	0.08	0.04	0.08
Standard deviation	0.03	0.03	0.03	0.03	0.03	0.03
Standardized bias	–54.39	–0.39	–101.90	0.79	–126.84	–0.29
RMSE	0.03	0.03	0.04	0.03	0.05	0.09
Coverage	0.90	0.94	0.81	0.87	0.76	0.95
Nonignorable						
Mean	0.08	0.08	0.08	0.08	0.08	0.08
Standard deviation	0.03	0.03	0.03	0.03	0.04	0.03
Standardized bias	0.51	–2.86	3.72	–6.99	3.71	–8.89
RMSE	0.03	0.03	0.03	0.03	0.04	0.03
Coverage	0.95	0.95	0.95	0.94	0.96	0.94

Note. For the single complete case analysis, the estimated coefficient is 0.08. Standard deviation is the standardized error of the estimate. Following Collins, Schafer, and Kam (2001), standardized bias is "100 × (average estimate – parameter)/SD", RMSE is the square root of the mean of the squared difference between the estimate and the parameter, and coverage is the length of the actual confidence interval relative to the expected confidence interval. With the 95% confidence intervals used here, coverage should be .95. Collins et al. (2001) suggested that coverage less than .90 is problematic.

that effect goes to essentially zero in both the full sample and the imputed 1998 sample. The models for Internet use also are a bit different, with stronger and more precisely estimated effects for AGE, WHITE, FEMALE, and DEPRESS1. The greater precision arises from the larger sample size. The full imputation results suggest that Internet use may not influence depression scores.

A useful set of statistics to examine are γ, the amount of missing information (Equation 10) and *r*, the increase invariance due to nonresponse (Rubin, 1987, p. 91, Equation 3.3.12) $r_{\mathrm{m}} = (1 + m^{-1}B_{\mathrm{m}}/\overline{U}_m)$, which can be estimated from the imputation results. For the results in Table 1, we get the following values:

Variable	γ	*r*
INCOME	.52	.41
WHITE	.36	.32
TEEN	.42	.36
FEMALE	.53	.42

SUPPORT	.39	.34
STRESS1	.36	.32
DEPRESS1	.38	.34
INTERNET	.27	.25

INTERNET has the lowest fraction of missing information (not surprising given that this variable was measured automatically every time any user signed on to their personal computer), and consequently, the relative increase in variance due to nonresponse is smallest (.25). The psychological variables all have nonresponse effects of similar magnitude. Interestingly, INCOME and FEMALE both have high γs and correspondingly high rs. In the case of INCOME, this is because of a low response rate, but FEMALE has a high response rate. Although FEMALE has a high response rate, it is contributing very little information to our estimation of $\overline{Q}$, in this case, depression. One can use these statistics both to get a better sense of where the information is lost due to nonresponse and as an estimate of how many imputations m to use in the complete case analyses that are the focus of an investigation.

We would not claim at this point that Internet use does or does not have the effects suggested by Kraut et al. (1998). Rather, we would argue that the more appropriate way to approach the analysis of this problem is by use of MI to address missing data. Our simulation results support this as well. At this point, the substantive analysis would have to consider whether the relationship between Internet use and depression is appropriately specified and measured, to what degree selection and self-selection effects might be influencing the pattern of results, and whether a hierarchical linear models approach might be a more appropriate way to analyze this data. In any case, however this interesting and important problem is addressed, MI procedures such as those demonstrated here would greatly facilitate the analysis. One could think of an MI approach as a way of assessing the implications of different assumptions about missingness for the results.

Discussion

Clearly, the use of MI changes the conclusions we might draw in the HomeNet study, that is, that the use of the Internet is related to changes in level of depression. Nevertheless, neither the original analysis nor our reanalysis should be considered as anything more than the first (not last) word on this issue. The concerns about the consequences of Internet use on psychological and social well-being are important and need continued research attention. We hope the use of MI to deal with one perennial problem in such research, missing data, will help researchers to conduct studies that provide them with valid and useful guidance on both the policy questions raised by the Kraut et al. (1998) results and a better empirical basis for drawing statistically and conceptually informative conclusions.

In the analyses we report, the results certainly suggest that those who responded are different from those who did not respond. Kraut et al. (1998) had noted this.

Our reanalysis suggests that when one considers the entire sample, in addition to being wealthier, $t(361) = 3.53$, $p < .001$, and older, $t(361) = 4.73$, $p < .001$, the respondents were more extraverted, $t(361) = 2.05$, $p < .05$; had a larger distant social circle at T1, $t(361) = 2.18$, $p < .05$; and used the Internet significantly more, $t(361) = 2.98$, $p < .01$,

than nonrespondents. This suggests that the original HomeNet findings were based on a more active (socially and technologically) sample. Consistent with these nonrespondent analyses, a follow-up study by the same HomeNet researchers found that consequences of Internet use were generally more positive for extraverts and those high in social support compared to introverts and those low in social support (Kraut, Kiesler et al., 2002). Our reanalysis suggests that when one considers the entire sample, the effects are far weaker for many of the estimates in the 1998 study, including the provocative depression results. In fact, our results suggest that higher depression scores predict future Internet use (see Table 1 in which the Internet use model shows depression significantly predicting Internet use, whereas Internet use does not predict future depression score). This is not due to attenuated covariance estimates arising from MI. When we look at the imputed 1998 sample, both the 1998 data and our imputed data for that imputed 1998 sample show similar patterns of results, which are different from the analysis on the entire sample.

It is not a goal in this article to try to understand what mechanism(s) might have led to nonresponse; our assumption of MAR assumes the causes are ignorable. Should one consider modeling the decision to answer a particular question? There are several observations one can make. First, such models often are not robust (Stolzenberg & Relles, 1990). Second, there is some evidence that MI can still yield good results with plausible nonignorable mechanisms (King et al., 2001). Given these two observations, if there is a nonignorable mechanism, it may not be critical for the data analyst to model that mechanism. However, researchers certainly should consider modeling such mechanisms.

Interestingly, the imputation efforts of the HomeNet project are quite similar to our imputations, suggesting that we were able to accurately represent missing data in a manner consistent with their expert judgment and field knowledge. This suggests that our MI captures enough information in the model and data to be usefully substituted for more costly efforts to impute data from expert field knowledge. That is a very encouraging demonstration of the robustness of this procedure. When considered with our simulation results, MI does certainly warrant consideration for such an analysis.

One might argue based on considerations of cost and relative efficiency (see Equation 11 and Table 1) that collecting fewer observations and using imputation to compensate for the reduced number of observations is a strategy worth using. We agree up to a point. Certainly, if one has more questions that need to be asked than there is time and respondent energy, a design using separate forms, each of which asks a subset of questions with a common core (e.g., Graham et al., 1994), is a strategy well worth considering. However, this is not to suggest that one can merely collect a small fraction of responses and impute the remaining answers to questions. The quality of imputations is constrained by the quality and stability of observed responses, both of which should be greater the higher the fraction of data collected. We would not consider MI as an alternative to collecting data but as a useful strategy to consider when data collection is difficult or infeasible (e.g., the case of historical records that have disappeared).

An objection that comes to mind when presented with these results is that these results are based on "made-up" data. The imputations of Y_{miss} do not have the same standing as the Y_{obs}. All the data we work with has some degree of uncertainty. If we collect some y_i at two points in time, estimate a test-retest reliability, and determine that the two estimates of y_i are different even though our model or theory of the world

says they are the same, we have uncertainty in our estimate of y_i. Any score we assign someone for y_i is in some sense made up, such as when we take the average of the two scores or the sum of the two scores as our best estimate of y_i. When we use such an estimate from y_i to predict standing on some other variable z_i, the predicted score is made up, that is, computed from observations on y_i as might be done in a regression equation. We could observe z_i in the future, but it is unobserved now. What is important is to represent properly the degree of uncertainty in our estimate of z_i.

In MI, we are doing the same thing. We have some observations and use the information in those observations to impute observations that could have been made but were not made, much as in the prediction example just offered. Instead of the missing data being unobserved because we need to wait for some future time, it is unobserved for some reason, which may or may not be ignorable. However, we are doing essentially the same thing, generating our best imputation of what the score would be with proper consideration for the uncertainty of the estimate. We are not making up the data but expressing our knowledge about the missing information using the data in hand y_{obs} and a model of the both Y_{obs} and Y_{miss}. In the HomeNet example, we feel some reassurance that we are not just making up the data when we compare the results reported in Kraut et al. (1998) with the same models using the same cases but with our imputations and with our simulation results. All suggest the imputation procedure properly captured the information in the data rather than having made up the data. It is precisely this result that gives us confidence in the MI for the full sample in the HomeNet study.

Missing Data: What to Do

We have set out to demonstrate the value and utility of MI using data simulation methods to help researchers address missing data problems. These problems are ignored at our collective peril. The standard procedure in complete case analysis, listwise deletion, is certainly problematic under some conditions. One estimate is that use of listwise deletion in political science leads to "point estimate(s) . . . about a standard error farther away from the truth. . . . This is half of the distance from no effect to what we often refer to as 'statistically significant'" (King et al., 2001, p. 52). This is an unacceptably high level of error to tolerate. There is little reason to think that in field-based studies, particularly with surveys, the results in psychology and sociology would be very different. Other data sources with high missingness rates (e.g., archival research, particularly with historical records) may also generate comparable missingness rates. Researchers cannot ignore this problem; one must address this directly. No analysis will be unaffected by missing data.

The first thing researchers must do is plan their research design and data collection methods to minimize the fraction of data lost due to nonresponse. Techniques to increase response rates (e.g., Dillman, 1991) are of great value and should be applied. Training interviewers, providing motivating materials, and engaging research instruments are essential. Inventive methods to increase access to respondents, such as use of the telephone and now the Internet, have increased ease of access in the survey environment, but response rates are still only at the 70% to 80% levels. Techniques such as randomized response (Fox & Tracy, 1986) to increase response rates to sensitive questions are also essential.

Knowledge of missing data mechanisms should either be considered in the design or investigated if they are not known (Little & Rubin, 1987). We should try to understand the mechanisms leading to nonresponse to determine what they are and, in particular, if they are ignorable. In survey contexts, following up subsets of nonrespondents to learn why they did not respond can help researchers learn if the missing data mechanism is ignorable. Such investigations are very worthwhile.

For many situations, even after these efforts, we will still confront missing data. We are recommending the application of MI using some variant of the algorithms developed by Schafer (1997) or King et al. (2001). They provide a disciplined way to attack this problem and make full use of the information available in the data and the statistical tools for complete case analysis. Allison (2000) reported a simulation study that shows Schafer's (1997) MI algorithm simulating missing data with little bias.

How Might MI Change Research Practice?

Setting the sample size higher in anticipation of missing data. The researcher can now look at the cost of additional data collection against the loss of information given MI procedures. There will be cases in which MI will provide sufficiently good precision at lower cost than collecting additional data. This would be particularly true in situations in which the marginal costs of getting additional observations are not lower but rather higher for missing observations. For example, follow-ups, particularly for difficult-to-find respondents, can be very costly. In doing archival research, tracking down missing information on firms or people can be very costly. MI may allow us to do nearly as well or as well when the imputation model allows for superefficiency, as following up repeatedly or tracing missing respondents. For market surveys or political attitude polls that are time sensitive, there may not be time available to conduct follow-up surveys with respondents.

Keeping the length of the data collection instrument short so one does not lose respondents. We sometimes need to know more about our respondents than our time (or more likely, our respondents' time) allows. One technique is to design several questionnaires, each of which has a common core of questions and a set of questions that are only asked of a fraction of the respondents. Graham et al. (1994) reported such a design in which there were three forms, and only a third of the respondents answered some questions. This is missing data by design, and the mechanism (random assignment by the investigator) is ignorable. MI allows the investigator to handle this situation readily. For a detailed example, see Schafer and Olsen (1998).

When means are calculated for multi-item scales, they are often calculated on the basis of available data. We found this not to be a problem in the HomeNet analysis, but it can be in other cases. Such mean analysis is a form of pairwise deletion but is hidden from the analyst. MI prior to scale construction, where feasible, is advisable. Roth et al.'s (1999) results showing relatively small effects of such improper methods are of real value in deciding whether to apply MI at the item level or scale level.

With the availability of these techniques and useful software, MI using MCMC methods can become part of mainstream research practice. Data analysis by simulation using MCMC methods will profoundly change how we do data analysis in the next decade. If used properly, simulation methods such as data augmentation can

improve the validity and quality of our results. Researchers would be well advised to monitor developments in this area, as new work on this problem is appearing (e.g., Enders, 2001; Schafer & Graham, 2001; Sinharay, Stern, & Russell, 2001; West, 2001). In the case of missing data, MI methods can potentially reduce the costs of data collection if smaller samples are warranted given the availability of such techniques, which allow us to make the most of what we know.

Notes

1. We follow Schafer (1997) in this portion of the exposition, including his notation.

2. Simple calculation suggests this problem will often be substantial. If people randomly fail to provide data on 5% of our measures (missing completely at random), then if we are using five measures, we will only have the fraction $.95^5 = .77$ available cases.

3. In this case and many evaluations of missing data imputation procedures, analysts assess the quality of these procedures by Monte Carlo simulation or by taking complete data and simulating a missingness mechanism to see if the imputation technique given some missingness mechanism provides imputations in an unbiased and efficient way when there is no missingness. In the Graham, Hofer, and Piccinin (1994) analysis, their conclusion is based on simulating a missingness mechanism.

4. For example, SAS enables a user to create a file with parameter estimates such as regression coefficients and standard errors in many of its procedures, enabling easy compilation of the necessary data for calculating the overall value of a given $\overline{Q}$ or $\overline{U}$. Schafer's package NORM (available at no cost from Schafer at http://www.stat.psu.edu/~jls) calculates several of these statistics in a straightforward, accessible way.

5. Software for multiple imputation is also freely available from King (http://gking.harvard.edu/stats.shtml). SAS 8.1 now includes experimental modules for multiple imputation (PROC MI) and analysis (PROC MIANALYZE). SOLAS, published by Statistical Solutions, also offers software for missing data estimation, although there have been some negative reviews of earlier versions of this software (Allison, 2000).

6. In the most recent version of the NORM, the software shows the worst autocorrelation functions (ACF) plot. If that plot is satisfactory, then one can avoid looking at any ACF plots. This nicely addresses one of the objections to using ACF functions with many variables.

7. We do not report all features of the study because it is not our purpose to assess in detail the issues of Internet use and its impact. For details on sample characteristics and other aspects of the study, the reader should consult Kraut et al. (1998).

8. This sample includes 84 participants who completed the pretest too late for inclusion in the study, who did not complete the pretest but completed the posttest, who signed up to participate in the study but did not use the Internet service (a type of missing observation, because we do not know if they used another Internet service), and who were family members dropped or added during the study.

9. All variable names in tables are capitalized, and their name is given when they are first introduced.

10. Kraut et al. (1998) did the same transformations to meet the requirements for their regression analysis.

11. Thanks to one of the referees for calling this point to our attention.

12. This procedure is not the usual and appropriate use for the expectation maximization (EM) algorithm. The EM algorithm is used to estimate parameters of the data, not the values of the data themselves.

13. We have recomputed the other regression results in Kraut et al. (1998). Those results are available from the first author upon request.

14. Pragmatically, most investigators would be unable to currently use multiple imputation with mixed or random effects models. Software for such procedures is not readily available. Schafer's (1997) PAN program is designed to do this but is written only for some versions of S+.

15. In Table 1, there are several cases in which the standard errors for β coefficients are substantially larger in the full sample compared to the 1998 article subsample. For INTERNET and STRESS1, the *SE*s increase from .11 to .56 for INTERNET and from .08 to .54 for STRESS1. This is due to a substantial increase in between-imputation variance for these two coefficient estimates (as in Equation 5).

References

Allison, P. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods and Research*, *28*(3), 301-309.

Allison, P. (2002). *Missing data*. Thousand Oaks, CA: Sage.

Berk, R. A. (1983). An introduction to sample selection bias in sociological analysis. *American Sociological Review*, *48*, 386-398.

Cohen, S., Mermelstein, R., Kamarck, T., & Hoberman, H. (1984). Measuring the functional components of social support. In I. G. Sarason (Ed.), *Social support: Theory, research and applications* (pp. 73-94). The Hague, Netherlands: Martinus Nijhoff.

Collins, L. M., Schafer, J. L., & Kam, C. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330-351.

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

Dillman, D. A. (1991). The design and administration of mail surveys. *Annual Review of Sociology*, *17*, 225-249.

Enders, C. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, *8*(1), 128-141.

Fox, J. A., & Tracy, P. E. (1986). *Randomized response: A method for sensitive surveys*. Beverly Hills, CA: Sage.

Graham, J. W., Hofer, S. M., & Piccinin, A. M. (1994). Analysis of missing data in drug prevention research. In L. M. Collins & L. A. Seitz (Eds.), *Advances in data analysis for prevention intervention research* (Vol. 142). Rockville, MD: National Institute on Drug Abuse.

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

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.

Johnson, B. T., Mullen, B., & Salas, E. (1995). Comparison of three major meta-analytic approaches. *Journal of Applied Psychology*, *80*(1), 94-106.

Kanner, A. D., Coyne, J. C., Schaefer, C., & Lazarus, R. S. (1981). Comparisons of two modes of stress measurement: Daily hassles and uplifts versus major life events. *Journal of Behavioral Medicine*, *4*, 1-39.

Kiesler, S., & Kraut, R. (1999). Internet use and ties that bind. *American Psychologist*, *54*(9), 783-784.

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*(1), 49-69.

Kraut, R., Kiesler, S., Boneva, B., Cummings, J., Helgeson, V., & Crawford, A. (2002). The Internet paradox revisited. *Journal of Social Issues*, *58*(1), 49-74.

Kraut, R., Patterson, M. , Lundmark, V., Kiesler, S., Mukhopadhyay, T., & Scherlis, W. (1998). Internet paradox: A social technology that reduces social involvement and psychological well-being? *American Psychologist*, *53*(9), 1017-1031.

Little, R. J. A. (1992). Regression with missing *X*s: A review. *Journal of the American Statistical Association*, *87*(420), 1227-1237.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198-1202.

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

Little, R. J. A., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods and Research*, *18*(2,3), 292-326.

Little, R. J. A., & Schenker, N. (1995). Missing data. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 39-75). New York: Plenum.

Meng, X. L. (1994). Multiple imputation with uncongenial sources of input (with discussion). *Statistical Science*, *9*, 538-574.

Radloff, L. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385-401.

Roth, P. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, *47*, 537-560.

Roth, P. L., Switzer, F. S., III, & Switzer, D. (1999). Missing data in multiple item scales: A Monte Carlo analysis of missing data techniques. *Organizational Research Methods*, *2*, 211-232.

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

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, *72*, 538-543.

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

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473-489.

Russell, D., Peplau, L., & Cutrona, C. (1980). The Revised UCLA Loneliness Scale: Concurrent and discriminant validity evidence. *Journal of Personality and Social Psychology*, *39*, 472-480.

Russell, D., Stern, H., & Sinharay, S. (2002). *An evaluation of multiple imputation as an approach to missing data*. Unpublished manuscript, University of Iowa.

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

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147-177.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analysts perspective. *Multivariate Behavioral Research*, *33*(4), 545-571.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, *6*(4), 317-329.

Stolzenberg, R. M., & Relles, D. A. (1990). Theory testing in a world of constrained research design: The significance of Heckman's censored sampling bias correction for nonexperimental research. *Sociological Methods & Research*, *18*(4), 395-415.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, *82*, 528-550.

West, S. (2001). New approaches to missing data in psychological research: Introduction to the special section. *Psychological Methods*, *6*(4), 315-316.

Mark Fichman is an associate professor of organizational behavior at Carnegie Mellon's Graduate School of Industrial Administration. He received a Ph.D. in organizational psychology from the University of Mich-

igan in 1982. He has done research on absenteeism, interorganizational relations, dynamic decision making, and methods.

Jonathon N. Cummings is an assistant professor of management at the Sloan School of Management, Massachusetts Institute of Technology. He completed his doctoral dissertation on "Work Groups and Knowledge Sharing in a Global Organization" at Carnegie Mellon University, where he was a postdoctoral fellow at the Human-Computer Interaction Institute. He is interested in the social impact of geographic dispersion and communication technology on interpersonal relationships and work groups.