Structural Equation Modeling, 15:651–675, 2008
Copyright © Taylor & Francis Group, LLC
ISSN: 1070-5511 print/1532-8007 online
DOI: 10.1080/10705510802339072

Psychology Press
Taylor & Francis Group

What Improves with Increased Missing Data Imputations?

Todd E. Bodner
Portland State University

When using multiple imputation in the analysis of incomplete data, a prominent guideline suggests that more than 10 imputed data values are seldom needed. This article calls into question the optimism of this guideline and illustrates that important quantities (e.g., p values, confidence interval half-widths, and estimated fractions of missing information) suffer from substantial imprecision with a small number of imputations. Substantively, a researcher can draw categorically different conclusions about null hypothesis rejection, estimation precision, and missing information in distinct multiple imputation runs for the same data and analysis with few imputations. This article explores the factors associated with this imprecision, demonstrates that precision improves by increasing the number of imputations, and provides practical guidelines for choosing a reasonable number of imputations to reduce imprecision for each of these quantities.

Multiple imputation is a statistical method to replace each missing value in a data set with m pseudo-random values thereby creating m "complete" data sets. Researchers analyze these m complete data sets and then combine the m results to make inferential statements about parameters of interest while incorporating additional inferential uncertainty due to the missing data. Several sources document the practical and technical issues one encounters when using the multiple imputation technique (e.g., Allison, 2002; Longford, 2005; Rubin, 1987; Schafer, 1997a). One practical decision is how many imputations are needed. Rubin (1987) illustrated that using between 2 and 10 imputations loses little estimator efficiency (relative to an infinite number of imputations) when the fraction of

Correspondence should be addressed to Todd E. Bodner, Department of Psychology, Portland State University, P.O. Box 751, Portland, OR 97207. E-mail: tbodner@pdx.edu

missing information (defined later) is modest and generates confidence intervals and hypothesis tests close to their nominal coverage and significance levels, respectively. However, when the fraction of missing information is very large, more than 10 imputations may be needed (Schafer, 1997a).[1]

Two recent contributions to the "Teacher's Corner" in this journal revisited the number of imputations question. Hershberger and Fisher (2003), using results from sampling theory where the "true" number of imputations m is a quantity to be estimated, found that the needed m can often be surprisingly large (e.g., several hundred) for a specified level of precision. In response, von Hippel (2005) challenged the sampling theory approach, clarified Rubin's earlier work on this topic, and was generally thankful for practical reasons that so many imputations were in fact not needed.

Motivated by results scattered across published work on multiple imputation that question such optimism (e.g., Royston, 2004), this article reconsiders the number of imputations question and argues that researchers should use (when practical) a larger number of imputations than the 2 to 10 Rubin suggested in certain situations. The main point made and illustrated here is that important inferential quantities such as null hypothesis significance test (NHST) p values and confidence interval (CI) half-widths can exhibit substantial variability across independent multiple imputation runs with small m for the same data and analysis. Substantively, a researcher (or two researchers) can draw categorically different conclusions about null hypothesis rejection or estimation precision in such cases. Furthermore, the estimated fraction of missing information (a quantity and concept with great practical and theoretical importance for inference with incomplete data) can also exhibit substantial variability with small m. In contrast with Rubin's (1987) seminal work justifying between 2 and 10 imputations so that point estimates are near their minimum sampling variance on repeated sampling from the population of interest, this work focuses on minimizing the variability in estimated NHST p values, CI half-widths, and fractions of missing information for a realized sample if the data were reimputed.

This article reviews the existing research on this phenomenon called *imputation variance* and provides a more extensive and focused illustration. Fortunately, prior research and this article demonstrate that increasing m reduces imputation variability. After reviewing the logic and relative merits of two existing methods for choosing m, a simulation experiment is conducted and reported to offer researchers further guidance on this decision. In the discussion of this simulation experiment, a practical method is provided to determine a reasonable number of imputations at two levels of confidence. The final section summarizes the results

[1] For example, in latent class analysis, where the indicators of class membership are completely unobserved, the fraction of missing information can exceed .95 (e.g., Loken, 2004).

and offers further guidelines and suggestions for the analysis of data sets with missing values using multiple imputation.

HYPOTHESIS TESTS AND CONFIDENCE INTERVALS

Common inferential procedures such as NHSTs and CIs seek to make probabilistic statements about population parameter values from sample statistics. With complete sample data (i.e., data sets with no missing values), inferential procedures must account for uncertainty due to the sampling variability of these statistics. With incomplete sample data (i.e., data sets with missing values), inferential procedures must also account for uncertainty due to the unknown missing data values. The following procedures for complete and incomplete data are designed to accomplish these tasks.

For both the complete and incomplete data cases, NHSTs and CIs are functions of the following three components: a parameter estimate, an estimated standard error for that parameter estimate, and a critical value on a reference distribution. To illustrate the complete data case, let $\hat{\theta}$ be an estimator of parameter θ where θ is a mean or a linear function of means (e.g., a difference in means or a regression slope). Under normal theory, an NHST that θ equals θ_0 in the population of interest can be written as

$$\frac{\hat{\theta} - \theta_0}{\widehat{SE}(\hat{\theta})} \sim t_{(1-\alpha/2, df)} \tag{1}$$

where $\widehat{SE}(\hat{\theta})$ is the estimated standard error for $\hat{\theta}$ and $t_{(1-\alpha/2, df)}$ is the reference distribution with critical value $t^*_{(1-\alpha/2, df)}$ determined by the degrees of freedom df and a decision for α. Similarly a CI for θ can be written as

$$\hat{\theta} \pm t^*_{(1-\alpha/2, df)} \widehat{SE}(\hat{\theta}), \tag{2}$$

where $h = t^*_{(1-\alpha/2, df)} \widehat{SE}(\hat{\theta})$ denotes the CI half-width. Methods to determine these three quantities are widely known, available in most statistical textbooks, and implemented in most software packages.

For inference in the face of incomplete data using multiple imputation, the quantities for equations analogous to Equations 1 and 2 are less widely known and are less available in standard statistical textbooks and software packages. To motivate these equations, let $\hat{\theta}_k$ represent the estimate of θ and $\hat{s}^2_k$ represent the estimated sampling variance of $\hat{\theta}$ from the kth data set completed by multiple imputation, where $k = 1, \ldots, m$. From these m pairs of $\hat{\theta}_k$s and $\hat{s}^2_k$s, Rubin's (1987) rules for inference under multiple imputation are used to compute the three quantities of interest.[2] The parameter estimate of θ is $\bar{\theta}_m = \sum_k \hat{\theta}_k / m$. The estimated sampling variance of $\bar{\theta}_m$ equals $\hat{T}_m = \hat{U}_m + \left(1 + m^{-1}\right)\hat{B}_m$, where $\hat{B}_m = \sum_k \left(\hat{\theta}_k - \bar{\theta}_m\right)^2 / (m-1)$ and $\hat{U}_m = \sum_k \hat{s}_k^2 / m$. The estimated standard error of $\bar{\theta}_m$ equals $\sqrt{\hat{T}_m}$. Finally, the degrees of freedom[3] of the reference t distribution for multiple imputation inference is

$$\hat{\nu}_m = (m-1)\left[\frac{\left(1 + m^{-1}\right)\hat{B}_m}{\hat{T}_m}\right]^{-2}. \tag{3}$$

Thus for inference using multiple imputation of incomplete data, Equations 1 and 2 are, respectively,

$$\frac{\bar{\theta}_m - \theta_0}{\sqrt{\hat{T}_m}} \sim t_{(1-\alpha/2, \hat{\nu}_m)} \tag{4}$$

with critical value $t^*_{(1-\alpha/2, \hat{\nu}_m)}$ and

$$\bar{\theta}_m \pm t^*_{(1-\alpha/2, \hat{\nu}_m)} \sqrt{\hat{T}_m}. \tag{5}$$

In the remainder, let $\hat{p}_m$ represent the NHST p value from Equation 4 and $\hat{h}_m$ represent the half-width of the CI from Equation 5 each based on m imputations.

FRACTIONS OF MISSING INFORMATION

An important concept in statistical research on inferential methods for incomplete data is missing information. Longford (2005) summarized this and related concepts succinctly:

> Information about a quantity is defined as the reciprocal of the MSE [mean-squared error] of its efficient estimator.... The missing information about a

[2] The discussed combination rules are for scalar quantities. The rules for combining multidimensional quantities (e.g., partial regression slopes) are slightly more complicated and not used to keep the discussion simple. See Schafer (1997a, pp. 112–114) for the multidimensional combination rules.

[3] Note that $\hat{\nu}_m$ does not depend on just the sample size and the number of parameter restrictions as in typical inferential situations. Thus, the degrees of freedom $\hat{\nu}_m$ can be small even despite a large overall sample size.

> parameter is defined as the difference of the information contained in the complete and incomplete data sets, and the fraction of missing information is the ratio of this difference and the complete-data information. (p. 55)

Less formally, Allison (2002) stated that the fraction of missing information is "how much information is lost about each coefficient because of missing data" (p. 48). Both descriptions imply that missing information is specific for each parameter of interest and need not be equal for different parameters. In the remainder, let λ represent the fraction of missing information for a given parameter in a particular data set.

In multivariate data sets, the fraction of cases with missing values is not equivalent to the fraction of missing information (Longford, 2005). Thus another approach is needed to estimate λ, as its value is generally unknown. One approach is to use estimated quantities in the multiply imputed data sets.[4] Many sources formally define and develop an expression for λ and its estimator $\hat{\lambda}_m$ based on m imputations in the context of multiple imputation (e.g., Longford, 2005; Schafer, 1997a). For brevity, we move directly to the estimator

$$\begin{aligned}\hat{\lambda}_m &= 1 - \frac{(\hat{\nu}_m + 1)\hat{U}_m}{(\hat{\nu}_m + 3)\hat{T}_m} \\ &\simeq \frac{\hat{B}_m}{\hat{U}_m + \hat{B}_m},\end{aligned} \tag{6}$$

where the approximation in Equation 6 holds with large degrees of freedom $\hat{\nu}_m$ (indicating relatively precise estimation of T). To illustrate the estimated quantity using the approximation, consider the case with no missing data and therefore no missing information. In this case $\hat{\theta}_k = \bar{\theta}_m$ for all k, which implies that $\hat{B}_m$ and $\hat{\lambda}_m$ both equal zero. However, with missing data for the variables involved in $\hat{\theta}$ and assuming no linear dependencies among those variables, $\hat{\theta}_k \neq \bar{\theta}_m$ for all k, which implies that $\hat{B}_m$ is strictly positive and $\hat{\lambda}_m > 0$.

Although in applications of multiple imputation, a researcher's attention centers primarily around inference for the parameters of interest through NHSTs and CIs, it is important to inspect and report the $\hat{\lambda}_m$s for those parameters. Consider that Allison (2002), Longford (2005), Rubin (1987), and Schafer (1997a) all reported $\hat{\lambda}_m$ for each parameter of interest in their examples using

[4] One could also use the Expectation Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) to obtain estimates of the fraction of missing information. However, the EM algorithm produces estimates of the fraction of missing information for linear combinations of parameters rather than for each distinct parameter (Schafer, 1997a).

multiple imputation. Estimates of λ are important for three related reasons. First, as noted before, larger values of λ suggest the need for more than the typical 2 to 10 imputations. Second, Schafer (1997a) highly recommended researchers calculate $\hat{\lambda}_m$ for each quantity of interest as a "useful diagnostic for assessing how the missing data contribute to inferential uncertainty" (p. 110). Clearly, larger values of $\hat{\lambda}_m$ imply greater inferential uncertainty about θ due to the unknown values of the missing data. Third, Schafer (1997a, p. 136) suggested a certain robustness of multiple imputation to the assumptions underlying the method (e.g., an ignorable response mechanism, distributional assumptions, and the functional form of the imputation model) when the amount of missing data is not large. By implication, larger values of $\hat{\lambda}_m$ suggest the need for more serious consideration and if possible exploration of the appropriateness of these assumptions through techniques like sensitivity analysis (Longford, 2005). In short, the validity of inferences using multiple imputation is more critically dependent on the appropriateness of the imputation model and its assumptions when $\hat{\lambda}_m$ is large. Therefore, because important conclusions rest on the value of $\hat{\lambda}_m$, one would want this value to reflect λ as closely as possible.

MONTE CARLO VARIABILITY

For a given sample and with complete data, the quantities $\hat{\theta}$, $\widehat{SE}(\hat{\theta})$, and $t^*_{(1-\alpha/2,df)}$ from Equations 1 and 2 are constants (i.e., there is a unique number representing each quantity for a given sample) and therefore so are the NHST p value and CI half-width. Similarly, for incomplete data completed by multiple imputation with $m = \infty$, the quantities $\bar{\theta}_\infty$, $\sqrt{\hat{T}_\infty}$, and $t^*_{(1-\alpha/2,\hat{\nu}_\infty)} = z^*_{1-\alpha/2}$ in Equations 4 and 5 are also constants along with $\hat{p}_m = p_\infty$ and $\hat{h}_m = h_\infty$.[5] However for incomplete data completed by multiple imputation with finite m, the quantities $\bar{\theta}_m$, $\sqrt{\hat{T}_m}$, and $t^*_{(1-\alpha/2,\hat{\nu}_m)}$ in Equations 4 and 5 are no longer constants and therefore neither are $\hat{p}_m$ nor $\hat{h}_m$. All of these quantities can vary across distinct runs of $m < \infty$ imputations for the same data and analysis. The same holds true for $\hat{\lambda}_m$ (i.e., $\hat{\lambda}_m = \lambda_\infty = \lambda$ only for $m = \infty$).

Recall that for fixed m these quantities are functions of $\bar{\theta}_m$, $\hat{B}_m$, and $\hat{U}_m$, which are functions of the m pairs of $\hat{\theta}_k$ and $\hat{s}^2_k$. Variability in the $\hat{\theta}_k$ and $\hat{s}^2_k$ arise from the stochastic nature of the unknown and imputed missing data values and the unknown and estimated parameters in the imputation model (i.e., the statistical model used to generate the pseudo-random imputed data values;

[5] To keep the notation simple, all quantities for $m = \infty$ are subscripted with ∞ rather than $m = \infty$.

Schafer, 1997a). One can label variability in the $\hat{\theta}_k$ and $\hat{s}_k^2$ under multiple imputation (and therefore $\bar{\theta}_m$, $\hat{B}_m$, $\hat{U}_m$, $\sqrt{\hat{T}_m}$, $t^*_{(1-\alpha/2,\hat{\nu}_m)}$, $\hat{p}_m$, $\hat{h}_m$, and $\hat{\lambda}_m$) as imputation variability or variance to distinguish it from ordinary sampling variability.

Prior research has illustrated and discussed the effects of imputation variability on estimates of h_∞ and λ (e.g., Allison, 2002; Royston, 2004; Schafer, 1997a). The results of this research generally suggest that these estimates can be noisy for small m but become more stable with larger m. Royston (2004) attributed variability in $\hat{h}_m$ to the fact that $t^*_{(1-\alpha/2,\hat{\nu}_m)}$ and $\sqrt{\hat{T}_m}$ are functions of m and concluded that values of $\hat{h}_m$ can be unreliable with small m. Schafer (1997a) defended the inferential validity but ignored the reliability of these CI widths stating that the intervals explicitly account for uncertainty due to simulation error and that such intervals should exhibit the nominal parameter coverage probability (cf. Rubin, 1987, chap. 4). Thus, prior research has provided somewhat mixed conclusions on the importance of imputation variability in $\hat{h}_m$. Regarding $\hat{\lambda}_m$, Schafer (1997a) attributed its imputation variability to estimation imprecision for B (i.e., that which $\hat{B}_m$ estimates having degrees of freedom $m - 1$) and concluded that $\hat{\lambda}_m$ should be considered as a rough guide when m is small.

Thus the effects of imputation variability can be severe, generating substantively discrepant values of $\hat{h}_m$ and $\hat{\lambda}_m$ across independent runs of m imputations when m is small. Given the importance of the $\hat{\lambda}_m$ value developed earlier, it is difficult to reconcile Schafer's (1997a) suggestion that $\hat{\lambda}_m$ contains useful diagnostic information and Rubin's (1987) recommendation that only 2 to 10 imputations are needed with the reality that under these conditions $\hat{\lambda}_m$ could provide values quite discrepant from λ. Clearly, a researcher should use a value of m that gives reasonable precision in the estimation of h_∞ and λ and this can take much more than 10 imputations.

EXAMPLES

The two examples presented in this section illustrate the effects of imputation variability on $\hat{p}_m$, $\hat{h}_m$, and $\hat{\lambda}_m$ and factors related to such variability. Although prior publications have already illustrated many of these effects, several considerations suggest that further illustration is warranted. First, Royston (2004) and Schafer (1997a) reported conflicting conclusions on the importance of imputation variability on $\hat{h}_m$. These conflicting conclusions should be resolved. Second, the referenced prior research did not explicitly demonstrate the effects of imputation variability on NHST p values. These effects should be demonstrated given that p values are prevalent inferential tools despite ongoing controversy surrounding their use (cf. Nickerson, 2000). Third, illustrating these effects was not the primary emphasis of the published research (e.g., introducing a variety of missing data methods or a new software program). Thus, a casual reader might miss the import of such secondary discussions. Furthermore, illustrations of these effects were sparse in detail, presentations of these effects on different quantities were presented separately, and reasons for these effects were stated but not illustrated. Note also that the results of Allison (2002) and Schafer (1997a) were coarse because only two replicate values of these quantities were investigated. Fourth, the referenced prior discussions appeared in specialized books or software journals. Thus, researchers generally may not be aware of the impact of imputation variability on these important quantities. Finally, prior research has not explored the impact of sample size on imputation variability. The two examples that follow are designed to address these considerations.

Both examples present a reanalysis of the cholesterol data set from Ryan and Joiner (1994) presented and analyzed in Schafer (1997a). This data set was chosen because Schafer (1997a) found little substantive variability in $\hat{h}_m$ even for small m. The data consist of cholesterol levels for 28 patients measured 2, 4, and 14 days after myocardial infarction. The first two cholesterol measurements were fully observed; however, cholesterol levels for 9 of the 28 (32%) patients were not observed at the third time period. For brevity, the following discussion focuses on inference for the mean cholesterol level 14 days postattack $\theta = \mu_{14}$. The discussed t tests evaluate the null hypothesis H_0: $\mu_{14} = 200$, a commonly reported criterion demarcating unhealthy from healthy cholesterol levels. All discussed t tests assume $\alpha = .05$ and CIs assume two-sided intervals with 95% confidence. In Examples 1 and 2, the number of imputations and the sample size, respectively, are varied to explore and illustrate their effects on imputation variability in the $\hat{p}_m$, $\hat{h}_m$, and $\hat{\lambda}_m$. Imputations were generated using Schafer's (1997b) NORM library within the S-Plus computing environment.

Example 1: Varying the Number of Imputations

As in Schafer (1997a), the number of imputations was varied to demonstrate its effect on the values of $\hat{h}_m$ and $\hat{\lambda}_m$. Schafer used $m = 3$, 5, 10, 20, and 100 and reported two replicate values of $\hat{h}_m$ and $\hat{\lambda}_m$ for each level of m. To overcome the coarseness of his results, to study the impact of imputation variability over a wider range of m, and to include $\hat{p}_m$, 1,000 replicate values of $\hat{p}_m$, $\hat{h}_m$, and $\hat{\lambda}_m$ were generated at each level of m (i.e., $m = 3$, 5, 10, 20, 30, 40, 50, 100, 200, 400, and 800).

Figure 1 displays boxplots for the 1,000 realized values of $\hat{h}_m$, $\hat{p}_m$, and $\hat{\lambda}_m$ for each level of m. Whereas the medians of these quantities were relatively stable across m (i.e., Mdns $\simeq$ 19, .03, and .19, respectively), realized values for these

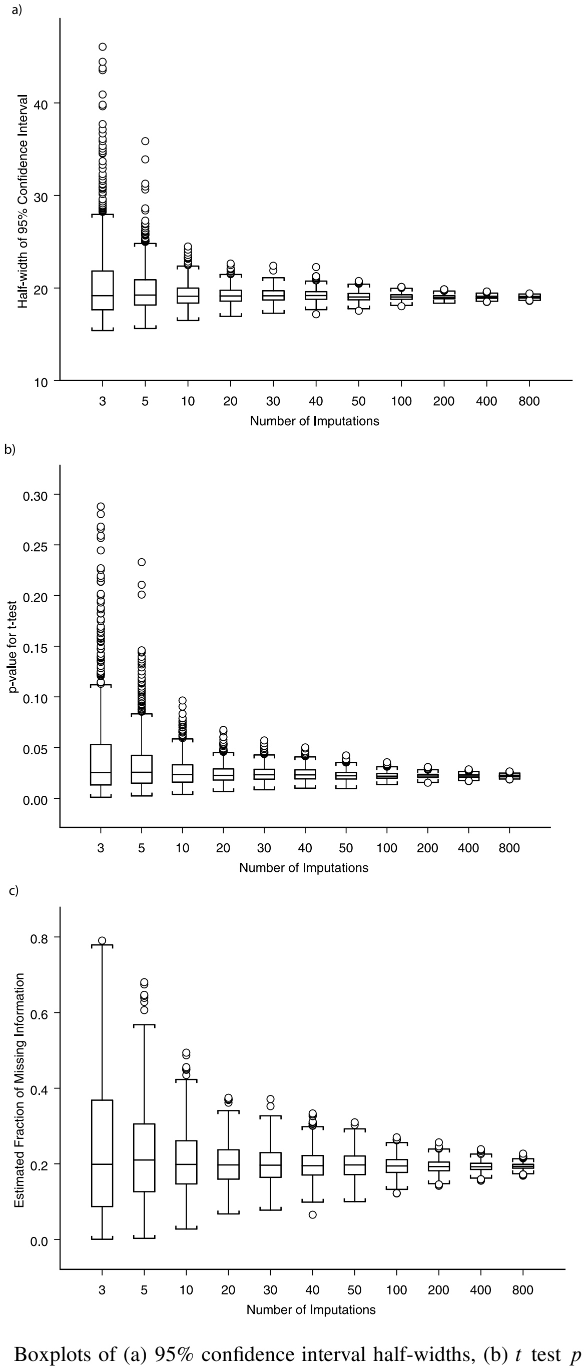

FIGURE 1 Boxplots of (a) 95% confidence interval half-widths, (b) t test p values, and (c) estimated fraction of missing information as a function of the number of imputations.

statistics were much more variable for small compared to large m. To illustrate the substantive impact of such variability in these important inferential quantities, consider the extreme yet realizable values of these statistics (i.e., the minima and maxima) in the case where $m = 3$. The maximum CI half-width ($\hat{h}_m^{max} = 46$) was over three times larger than the minimum ($\hat{h}_m^{min} = 15$); the maximum p value ($\hat{p}_m^{max} = .28$) was over 280 times larger than the minimum p value ($\hat{p}_m^{min} = .001$). In each case, these discrepancies would lead to categorically different inferential conclusions about estimation precision and null hypothesis rejection, respectively. The maximum estimated fraction of missing information ($\hat{\lambda}_m^{max} = .79$) was over 1,975 times larger than the minimum ($\hat{\lambda}_m^{min} < .0004$). This discrepancy would lead to categorically different conclusions about the extent of missing information and its potential implications for the validity of inference. In the former case, a researcher observing $\hat{\lambda}_m^{max}$ would likely want to assess the appropriateness of the imputation model and its assumptions; in the latter case, a researcher observing $\hat{\lambda}_m^{min}$ would likely ignore such additional considerations. In contrast when m was very large (e.g., 800), the maximum and minimum realized values of these three quantities were very similar, which would lead in each case to substantively identical conclusions. Thus, if one wanted to reduce the effect of imputation variability on $\hat{h}_m$, $\hat{p}_m$, and $\hat{\lambda}_m$, one should use much more than the recommended 2 to 10 imputations.

Recall that $\hat{h}_m$, $\hat{p}_m$, and $\hat{\lambda}_m$ are all functions of $\hat{U}_m$, $\hat{B}_m$, and m. To explore how variability in $\hat{B}_m$ and $\hat{U}_m$ relates to variability in $\hat{h}_m$, $\hat{p}_m$, and $\hat{\lambda}_m$, correlations between the former and latter quantities were computed for fixed m. The patterns in these correlations were substantively similar across m and therefore only the results for $m = 3$ are presented. $\hat{B}_m$ was strongly correlated with $\hat{h}_m$ ($r = .98$), $\hat{p}_m$ ($r = .84$), and $\hat{\lambda}_m$ ($r = .93$). In contrast, $\hat{U}_m$ was at most moderately correlated with $\hat{h}_m$ ($r = .30$), $\hat{p}_m$ ($r = .36$), and $\hat{\lambda}_m$ ($r = .08$). Thus, the variability in $\hat{h}_m$, $\hat{p}_m$, and $\hat{\lambda}_m$ values appears more largely attributable to variability in $\hat{B}_m$ rather than $\hat{U}_m$. Schafer (1997a) stated that variability in $\hat{\lambda}_m$ is likely due to imprecise estimation of $\hat{B}_m$ with degrees of freedom $m - 1$. These results illustrate this explanation and extend it to variability in $\hat{h}_m$ and $\hat{p}_m$ as well.

Example 2: Varying the Sample Size

Given the known effects of sample size on the $\hat{s}_k^2$ and variability in the $\hat{\theta}_k$ (i.e., critical components in the construction of $\hat{U}_m$ and $\hat{B}_m$), one might be concerned that the imputation variability observed in $\hat{h}_m$, $\hat{p}_m$, and $\hat{\lambda}_m$ might be an artifact due to the small sample size (i.e., $N = 28$). To explore the effect of sample size on these latter quantities, the cholesterol data were reanalyzed varying the sample size. This was achieved by concatenating the original data set (i.e., repeating the data set multiple times) to yield data sets of size $N = 28$, 280, 2,800, and 28,000. These data sets have the same mean and covariance structure and the same percentage of missing values for cholesterol measurements 14 days after myocardial infarction. The following exploration focused on the case of $m = 3$ where variability in $\hat{h}_m$, $\hat{p}_m$, and $\hat{\lambda}_m$ was maximal in Example 1. As before, each data set was analyzed 1,000 times, yielding 1,000 values of $\hat{h}_m$, $\hat{p}_m$, and $\hat{\lambda}_m$ for each sample size.

Table 1 displays five-number summaries of the $\hat{h}_m$ and $\hat{\lambda}_m$ values as a function of sample size. Regarding $\hat{h}_m$, values for $\hat{h}_m$ decreased substantially with increasing N as one would expect. However, for all sample sizes the maximum realized value of $\hat{h}_m$ was consistently about three times the minimum realized value. Regarding $\hat{\lambda}_m$, median values did not substantively change as a function of N. Furthermore, variability in $\hat{\lambda}_m$ values did not substantially change with increasing N. Regarding $\hat{p}_m$, the extremely small p values observed for the larger sample sizes preclude concise tabular or graphical presentation. Therefore these results are only presented textually. As one would expect, values for $\hat{p}_m$ decreased substantially with increasing N, with median values ranging from .03 for $N = 28$ to 1.2×10^{-64} for $N = 28{,}000$. However, the values of $\hat{p}_m$ continued to demonstrate considerable imputation variability with the ratios of maximum to minimum realized p values increasing from 260 for $N = 28$ to essentially infinity for $N = 28{,}000$. Although in this case, the p values for $N \geq 280$ were uniformly below $\alpha = .05$, data sets with large sample sizes and a smaller effect size than in this example could generate p values that straddle α as illustrated ear-

TABLE 1
Five-Number Summaries for the Estimated 95% Confidence Interval Half-Widths and Fractions of Missing Information as a Function of Sample Size

Quantity	*N*	*Min*	Q_1	*Mdn*	Q_3	*Max*
$\hat{h}_3$	28	14.94	17.77	19.37	21.83	53.00
	280	5.04	5.45	5.83	6.54	16.47
	2,800	1.63	1.71	1.83	2.05	5.00
	28,000	0.52	0.54	0.58	0.65	1.65
$\hat{\lambda}_3$	28	0.00	0.09	0.20	0.36	0.79
	280	0.00	0.09	0.20	0.35	0.81
	2,800	0.00	0.08	0.20	0.34	0.80
	28,000	0.00	0.09	0.20	0.35	0.82

Note. $\hat{h}_3$ = estimated confidence interval half-width based on $m = 3$; $\hat{\lambda}_3$ = estimated fraction of missing information based on $m = 3$; N is the sample size; Q_1 and Q_3 are the first and third quartiles, respectively.

lier for $N = 28$. In summary, the issue of imputation variability in $\hat{h}_m$, $\hat{p}_m$, and $\hat{\lambda}_m$ values is not just a small-sample problem, but can also occur in large samples.

HOW MANY IMPUTATIONS?

Given the potential imputation variability in important inferential (e.g., CI widths and NHST p values) and descriptive (e.g., $\hat{\lambda}_m$) quantities under the multiple imputation of incomplete data with small m, the question again becomes how large should m be? This section reviews two methods to determine m and discusses their relative merits. To overcome the stated limitations of these methods, a simulation experiment is conducted in the next section to help guide decisions for choosing m.

Rubin's Method

Rubin's (1987) suggestion that few imputations are needed was based on the efficiency of point estimates. The relative efficiency in using m rather than an infinite number of imputations is approximately $(1 + \lambda/m)^{-1}$. With $\lambda = .19$, estimation of θ is 94% as efficient using $m = 3$ compared to $m = \infty$. In this case, the standard error of $\bar{\theta}_m$ (i.e., $\sqrt{\hat{T}_m}$) will on average be $\sqrt{1 + .19/3} = 1.03$ times larger using $m = 3$ compared to $m = \infty$. To illustrate using the Example 1 data, consider a comparison of the mean values of $\sqrt{\hat{T}}$ for ms $= 3$ and 800 (i.e., 9.97 and 9.70, respectively). Their ratio equals 1.03, which is the same (to two decimal places) as the theoretical value because $m = 800$ returns a relative efficiency value close to that as if $m = \infty$. With $m = 10$, $\sqrt{\hat{T}_m}$ will tend to only be on average 1.01 times larger than for $m = \infty$. These small differences in estimator efficiency (i.e., both those theoretically derived and empirically estimated) support and illustrate the theoretical rationale behind Rubin's recommendation.

However, two issues merit further consideration. First, use of Rubin's equation requires knowledge of λ, which is typically unknown. Furthermore, the results of the two examples suggest that estimation of λ can be imprecise with few imputations. Second, discussions of estimator efficiency involve the expected variance of the sample statistic. However, any realized value of $\hat{T}_m$, which is a function of variable $\hat{B}_m$ and $\hat{U}_m$ values, can deviate substantially from its average value, especially with small m. This deviation, along with the corresponding variability in the critical value of the reference distribution, generates the imputation variability $\hat{h}_m$ and $\hat{p}_m$ for independent runs of m stochastic imputations. von Hippel (2007) provided an expression demonstrating that the variance of

$\hat{T}_m$ decreases with m. Furthermore as illustrated, imputation variability in $\hat{h}_m$ and $\hat{p}_m$ stabilizes as m increases. However, Rubin's method does not address imputation variability and therefore is not useful to determine a value of m to minimize its effects. Note that Schafer (1997a) defended the inferential validity of $\hat{h}_m$ on similar grounds, referring to the expected probability coverage of CIs. This is likely true as $\hat{h}_m$ will on average neither tend to over- or underestimate the CI half-width on repeated sampling. However, for a given sample with missing data, imputation variability does lead to imprecise estimation of $\hat{h}_m$ that can have substantive inferential impact. Thus, Royston's (2004) assessment that $\hat{h}_m$s can be noisy and therefore problematic under multiple imputation with small m appears the more appropriate conclusion.

Royston's Method

Royston (2004) provided an iterative method for choosing m based on a scaled coefficient of variation (CV $= SD/M \times 100$) of $\hat{h}_m$. To use the method, one first generates w replicate sets of m imputations and then computes $\hat{h}_m$ for each set; next one computes the CV for these w values of $\hat{h}_m$. If the CV is greater than some criterion (Royston suggested CV $= 5\%$), one increases m and repeats the process until the CV is less than the criterion. The number of imputations m is the value of m when this criterion is met.

Although this method is useful for reducing imputation variability in $\hat{h}_m$, two issues merit consideration. First, Royston offered no guidance in the specification of w, an initial value of m, or the increase in m given failure to reach criterion, and does not discuss what to do if NHST p values or $\hat{\lambda}_m$s are of interest. Regarding values of w, one would expect substantial variability in the realized CV values for small w, which could lead to gross over- or underestimation of the actual CV. Second, Royston's iterative method is odd for choosing m because it requires imputing the data substantially more than m times. Consider that if the process is repeated J times, the researcher will have generated $w \times \sum_{j=1}^{J} m_j$ complete data sets by multiple imputation. Starting with a sufficiently large m would save time and reduce the complexity of the process. Thus, Royston's choice of m is a by-product of a small CV for $\hat{h}_m$ rather than vice versa.

SIMULATION EXPERIMENTS

A series of simulation experiments were conducted with the goal of providing researchers guidance for choosing a number of imputations to reduce to reasonable levels the effects of imputation variability on estimated 95% CI half-widths, NHST p values, and fractions of missing information. These experiments occurred in two phases. In the first phase, the number of imputations and the true fraction of missing information were varied systematically to explore the impact of these factors on these estimated statistics. In the second phase, criteria were sought representing reasonable levels of imputation variance for these statistics along with the minimum values of m necessary to achieve these criteria at two levels of confidence.

Next, the method for the first phase is outlined. Changes in methodology for the second phase are noted in the Results section. Throughout for simplicity, the simulated data Y were univariate and the missing data were assumed MCAR. The inferential estimand of interest was the mean of Y, μ_Y, based on incomplete data and completed by multiple imputation.

Method

A sample size of $n = 100$ was chosen to reflect the median sample size reported in a recent survey of psychological research (Bodner et al., 2004) where n_1 and n_0 ($n_1 + n_0 = n$) values for Y are observed and missing, respectively. Two factors were varied, the number of imputations (i.e., ms = 3, 5, 10, 20, 30, 40, 50, 75, and 100) and the fraction of missing information (i.e., λs = .05, .1, .2, .3, .5, .7, and .9). Manipulation of λ in this univariate setting was achieved by setting $n_1 = n(1 - \lambda)$. Thus, there were 63 conditions in the simulation experiment. Within each condition, the mean and variance of the n_1 observed values of Y were fixed arbitrarily at $\bar{y}_1 = .2$ and $s_1^2 = 1$ given that the variable Y (and any covariates X) are considered fixed with multiple imputation.[6] Descriptively, these values, $(\bar{y}_1 - 0)/s_1 = .2$, represent a point estimate for a small standardized mean difference in the social and behavioral sciences (Cohen, 1988).

Multiple imputation was used to generate plausible values for the n_0 missing data values using the following three-step process (Rubin, 1987, p. 167). First, a candidate value for σ_Y^2 was drawn such that $\sigma_{Y*}^2 = s_1^2(n_1 - 1)/g$ where g is a random draw from a $\chi_{n_1-1}^2$ distribution. Second, a candidate value for μ_Y was drawn such that $\mu_{Y*} = \bar{y}_1 + z\sqrt{\sigma_{Y*}^2/n_1}$ where z is a random draw from an $N(0, 1)$ distribution. Third, candidate values for the n_0 missing values of Y were generated as independent random draws from an $N(\mu_{Y*}, \sigma_{Y*}^2)$ distribution. This process was repeated m times to generate m completed data sets. Rubin's (1987) standard formulas were used to combined the results of the m data sets to compute $\hat{h}_m$, $\hat{p}_m$, and $\hat{\lambda}_m$ where the NHST p value evaluated the null hypothesis that H_0: $\mu_Y = 0$. This process was repeated 5,000 times yielding 5,000 values of $\hat{h}_m$, $\hat{p}_m$, and $\hat{\lambda}_m$ for each of the 63 experimental conditions.

[6] Rubin (1987) noted that the missing values of Y are considered random conditional on Y, X, and the outcome of the participant response and sampling process.

The primary outcome of interest was how the variability in these 5,000 values changed as a function of m and λ. To quantify variability, the 95% interpercentile range (IPR) was used, defined as the difference in values at the 97.5th and 2.5th percentiles of the 5,000 values. The 95% IPR was chosen to represent a conservative range such that only 5% of the realized values fall outside of the boundary values that define this IPR.

Results

Phase I. Without missing data and assuming that $\bar{y}_1 = \bar{y}$ and $s_1^2 = s^2$, note that $h = .20$ and $p = .05$ with IPRs of zero. With missing data, h_∞ and p_∞ can be computed analytically given the constraints and assumptions of the simulation (e.g., fixed values of $\bar{y}$ and s^2, univariate data, and the missing completely at random [MCAR] assumption) and given that the reference t distribution is standard normal when $m = \infty$. Figure 2 displays these values and illustrates that h_∞ and p_∞ both increase with λ. These increases are expected as multiple imputation incorporates additional inferential uncertainty due to nonresponse. Clearly a large fraction of missing information can degrade statistical power and estimation precision substantially even when an infinite number of imputations is used. Note that $\lambda_\infty = \lambda$, so these values are not graphed. Based on the

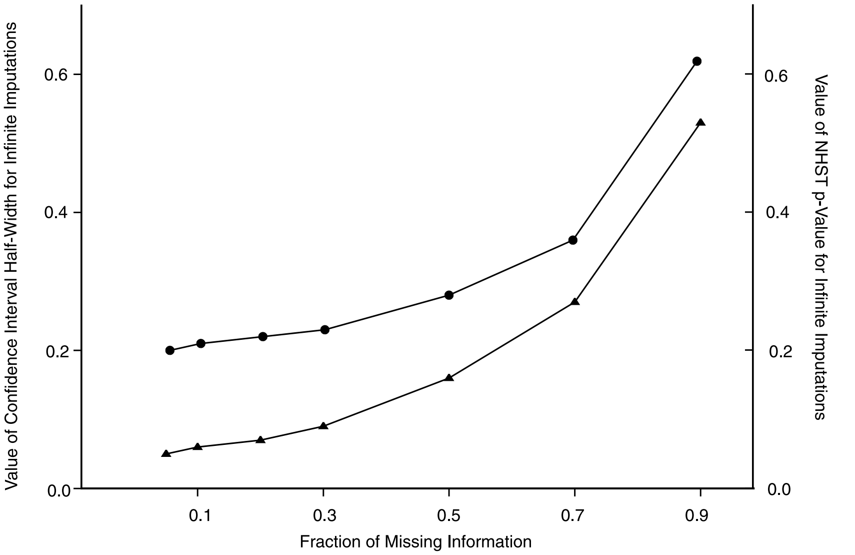

FIGURE 2 Values of h_∞ (circles) and p_∞ (triangles) as a function of the fraction of missing information.

simulation results for finite m, mean values for $\hat{h}_m$, $\hat{p}_m$, and $\hat{\lambda}_m$ (i.e., $\bar{h}_m$, $\bar{p}_m$, and $\bar{\lambda}_m$, respectively) were close to their theoretical values for $m \geq 10$ at the levels of λ studied.

Table 2 displays the 95% IPRs for $\hat{h}_m$, $\hat{p}_m$, and $\hat{\lambda}_m$ given m and λ. Variability in $\hat{h}_m$ and $\hat{p}_m$ increased with λ and decreased with m. Consistent with analytical

TABLE 2
Values of 95% Interpercentile Ranges of 5,000 Simulated 95% Confidence Interval Half-Widths, Null Hypothesis Significance Test *p* Values, and Fractions of Missing Information as a Function of the True Fraction of Missing Information and the Number of Imputations

		λ						
	m	*.05*	*.10*	*.20*	*.30*	*.50*	*.70*	*.90*
$\hat{h}_m$	3	.04	.08	.17	.30	.62	1.24	3.12
	5	.02	.04	.09	.15	.30	.57	1.48
	10	.01	.02	.05	.08	.17	.31	.82
	20	.01	.02	.03	.05	.10	.19	.53
	30	.01	.01	.03	.04	.08	.15	.43
	40	.01	.01	.02	.04	.07	.13	.36
	50	.01	.01	.02	.03	.06	.12	.33
	75	.00	.01	.02	.03	.05	.09	.27
	100	.00	.01	.01	.02	.04	.08	.23
	∞	.00	.00	.00	.00	.00	.00	.00
$\hat{p}_m$	3	.08	.15	.26	.39	.62	.89	.96
	5	.06	.10	.18	.28	.49	.74	.87
	10	.04	.06	.12	.18	.34	.55	.75
	20	.03	.04	.08	.13	.24	.40	.62
	30	.02	.03	.06	.10	.19	.32	.50
	40	.02	.03	.06	.08	.16	.29	.44
	50	.02	.03	.05	.07	.15	.25	.38
	75	.01	.02	.04	.06	.12	.21	.32
	100	.01	.02	.03	.05	.11	.18	.28
	∞	.00	.00	.00	.00	.00	.00	.00
$\hat{\lambda}_m$	3	.24	.42	.64	.76	.86	.88	.72
	5	.15	.28	.46	.58	.69	.65	.35
	10	.10	.17	.32	.40	.48	.41	.20
	20	.06	.12	.21	.28	.32	.28	.13
	30	.05	.10	.17	.22	.27	.22	.11
	40	.04	.08	.15	.19	.23	.19	.09
	50	.04	.07	.13	.17	.20	.17	.08
	75	.03	.06	.11	.14	.16	.14	.07
	100	.03	.05	.09	.12	.14	.12	.06
	∞	.00	.00	.00	.00	.00	.00	.00

Note. The 95% interpercentile range (IPR) is the difference in outcome values at the 97.5th and 2.5th percentiles.

results (cf. Harel, 2006), variability in $\hat{\lambda}_m$ was maximal at $\lambda = .5$ when $m \geq 5$ but decreased with increasing m.[7] These results suggest that larger λs generate greater imputation variability in $\hat{h}_m$, $\hat{p}_m$, and $\hat{\lambda}_m$ (i.e., up to $\lambda = .5$ for $\hat{\lambda}_m$) and that increasing m can counteract this effect.

Phase II. The goals of the second-phase simulation experiments were to construct a reasonable criterion of precision for each statistic and to provide guidelines for the number of imputations to meet those criteria for differing λ. Due to the mean shifts in these statistics across λ values and the potential variability restriction of these statistics near their boundaries, attention was paid in developing such criteria when substantively important.

Exploratory simulation experiments revealed that variability in $\hat{h}_m$ values could be made approximately proportional to its mean value $\bar{h}_m$ through an appropriate choice of m for the levels of λ studied. Thus to capitalize on this behavior and to specify a reasonable degree of imputation variability, a criterion was specified such that $\hat{h}_m$ should be within 10% of its mean value $\bar{h}_m$ (i.e., $\hat{h}_m \in \left[\bar{h}_m \pm .1\bar{h}_m\right]$). For example, if $\bar{h}_m = 1$, then $\hat{h}_m$ should fall between .9 and 1.1 to meet this criterion. Note that this criterion accounts for increases in variation in $\hat{h}_m$ as $\bar{h}_m$ increases away from zero.

The asymptotic results of Harel (2006) demonstrate that variability of $\hat{\lambda}_m$ around λ is maximal for $\lambda = .5$ and approaches zero for λ near 0 and 1. One can easily manipulate Harel's (2006) formulas to find the value of m to estimate $\lambda = .05$ and $\lambda = .5$ with equal precision; in brief, a much larger m is needed for fixed precision to estimate the latter λ compared to the former λ. However, for typical research purposes, it is not important to estimate λ with equal precision across the [0,1] interval. Thus, a criterion was specified such that $\hat{\lambda}_m$ should be within ± .1 of its mean value $\bar{\lambda}_m$ (i.e., $\hat{\lambda}_m \in \left[\bar{\lambda}_m \pm .1\right]$). For example, if $\bar{\lambda}_m = .2$, then $\hat{\lambda}_m$ should fall between .1 and .3 to meet this criterion. Although this criterion does not account explicitly for the known decrease in variability in $\hat{\lambda}_m$ as λ approaches zero and unity, I reasoned that a difference of .1 between $\hat{\lambda}_m$ and λ was sufficiently precise irrespective of λ and that greater precision is not typically of substantive interest.

[7] Harel (2006) noted that asymptotically

$$\frac{\sqrt{m}\left(\hat{\lambda}_m - \lambda\right)}{\sqrt{2}\lambda\,(1-\lambda)} \rightarrow N\,(0,1) \tag{7}$$

for values of λ not near zero or unity and further reported that $\sqrt{m}\left(Logit(\hat{\lambda}_m) - Logit(\lambda)\right) \rightarrow N\,(0,2)$ offers a better approximation for λ near zero or unity. These formulas imply that there is less variance in $\hat{\lambda}_m$ for λ near zero and maximal variance at $\lambda = .5$.

TABLE 3
Minimum Number of Imputations Needed for Estimated 95% Confidence Interval Half-Widths and Fractions of Missing Information to Achieve Specified Precision at Two Levels of Confidence

	$\hat{h}_m \in [\bar{h}_m \pm .1\bar{h}_m]$		$\hat{\lambda}_m \in [\bar{\lambda}_m \pm .1]$	
λ	*80% Confidence*	*95% Confidence*	*80% Confidence*	*95% Confidence*
.05	2	3	2	4
.10	3	6	5	9
.20	7	12	11	23
.30	12	24	18	36
.50	27	59	23	50
.70	50	114	16	36
.90	108	258	4	10

At each level of λ studied, an iterative search was conducted varying m to find the minimum value of m to meet these criteria for 80% and 95% of $\hat{h}_m$ and $\hat{\lambda}_m$ values. Specifying two levels of coverage permits researchers to dictate more (i.e., 95%) or less (i.e., 80%) control over imputation variability through the choice of m. Table 3 displays the results of these searches. Two results are apparent. First, as one would expect, a larger number of imputations is needed to contain a larger proportion of the $\hat{h}_m$ and $\hat{\lambda}_m$ values in the specified intervals (i.e., moving from 80% to 95% containment). Second, for $\hat{h}_m$ the minimum m increased approximately at an exponential rate with increasing λ; for $\hat{\lambda}_m$ the minimum m was maximal for $\lambda = .5$ and decreased as λ differed from .5.[8]

In contrast with the case of $\hat{h}_m$, the variability of $\hat{p}_m$ around $\bar{p}_m$ could not be adequately stabilized across differing $\bar{p}_m$ to generate minimum ms with any hope of generality. Smaller values for $\bar{p}_m$ resulted in less variability and common transformations for data values on the [0,1] interval (e.g., the logit and arcsine transformations) were not sufficient to counter this effect. Therefore, a simulation experiment was conducted where $\bar{p}_m$ was varied explicitly along with n and λ. The sample sizes n considered were 100 and 300. For each level of λ studied, $\bar{y}_1$ was manipulated to generate mean p values $\bar{p}_m$ of .50, .10, .05, .01, and .001. Of interest were the 80% and 95% IPRs and their boundary values for $\hat{p}_m$ around these $\bar{p}_m$s. (Note that the 80% IPR is difference in values at the 10th and 90th percentiles of the data distribution.) The values of m used were those appearing

[8] To explore the generality of these results, further simulations were conducted increasing $\bar{y}_1$ to .80 and n to 1,000. The change in $\bar{y}_1$ did not affect the minimum ms for either statistic. Furthermore, the change to $n = 1{,}000$ did not affect the minimum ms for $\hat{\lambda}_m$. However for $\hat{h}_m$, the change from $n = 100$ to $n = 1{,}000$ resulted in minimum ms about 10% smaller, which suggests that $\hat{h}_m$s are slightly less variable with larger n_1.

TABLE 4
Interpercentile Range and Boundary Values for Simulated Null Hypothesis Significance Test *p* Values Using Values of *m for Confidence Interval Half-Widths From Table 3*

	Boundary of 80% and 95% IPRs		
$\bar{p}_m$	*Lower Bound*	*Upper Bound*	*IPR*
.50	[.35	.67]	.32
.10	[.05	.16]	.11
.05	[.02	.09]	.07
.01	[.001	.023]	.022
.001	[.0002	.0032]	.0030

Note. Choice of *m* for confidence interval half-widths from Table 3 determines the percentage of $\hat{p}_m$ values contained in the interval. IPR = interpercentile range.

in Table 3 for $\hat{h}_m$. In particular, for varying levels of λ, *m*s for 80% confidence in $\hat{h}_m$ were used to assess 80% IPRs for $\hat{p}_m$ and *m*s for 95% confidence in $\hat{h}_m$ were used to assess 95% IPRs for $\hat{p}_m$. These values of *m* were chosen because they were successful for meeting the criterion of a fixed percentage of $\hat{h}_m$s within 10% of $\bar{h}_m$ the specified levels of λ. It was also thought that this stability might transfer in part to $\hat{p}_m$ conditional on $\bar{p}_m$.

Table 4 presents the results of the simulation where the choice of *m* for confidence level for $\hat{h}_m$ from Table 3 dictates the percentage coverage in the IPRs for $\hat{p}_m$. Note that Table 4 does not provide boundary values and IPRs for $\hat{p}_m$ around $\bar{p}_m$ for distinct λ, *n*, and *m* combinations. Two results justify this economy of presentation. First, for given values of $\bar{p}_m$ and *m*, the IPRs for $\hat{p}_m$ were substantively similar across $\lambda < .5$ and *n*.[9] Thus, the values of *m* that stabilize variability in $\hat{h}_m$ around $\bar{h}_m$ from Table 3 also appear to make variability in $\hat{p}_m$ around $\bar{p}_m$ relatively insensitive to differences in λ and *n*. Second, conditional on $\bar{p}_m$, the *m*s used from Table 3 to stabilize $\hat{h}_m$ to a fixed criterion interval also generated a criterion interval (approximately) common to both the 80% and 95% IPRs for $\hat{p}_m$. Thus for brevity these IPRs are discussed without distinction between their 80% and 95% coverage.

As presented in Table 4 and as expected, the IPRs for the $\hat{p}_m$ values decreased in size as $\bar{p}_m$ approached zero. Thus, when *m* is chosen to minimize variability in $\hat{h}_m$ from Table 3, variability in $\hat{p}_m$ is reasonably controlled for large and small $\bar{p}_m$ (e.g., $\bar{p}_m \leq .01$ and $\bar{p}_m \geq .10$) assuming $\alpha = .05$. In these cases, variability

[9] For $\lambda \geq .5$, the IPRs for $\hat{p}_m$ tended to decrease in size as λ increased.

in $\hat{p}_m$ with the specified confidence level occurs over a range far from α leading to the same decision regarding null hypothesis rejection. However, when $\bar{p}_m$ is at or near α (e.g., the conventional $\alpha = .05$), variability in $\hat{p}_m$ can occur over a range that straddles α and therefore can lead to differing decisions regarding null hypothesis rejection in subsequent runs of multiple imputation. In such cases (i.e., when an obtained $\hat{p}_m$ is near α), it would be prudent to increase m beyond the values presented in Table 3 for $\hat{h}_m$.

Discussion and Recommendations

For relatively accurate estimates of CI half-widths and NHST p values, the needed number of imputations increases with λ. For relatively accurate estimates of the fraction of missing information, the needed number of imputations increases as λ approaches .50. If λ was known, researchers could use the results presented in Table 3 to choose a reasonable value for m for relatively accurate estimates of CI half-widths and fractions of missing information. Researchers focusing on NHST p values can reduce substantively important imputation variability in $\hat{p}_m$ with respect to null hypothesis rejection when $\bar{p}_m$ is not near α by choosing a value of m from Table 3 that controls the amount of imputation variability in $\hat{h}_m$ at a specified level of confidence.

However, because λ is seldom known with incomplete multivariate data, one must estimate its value to use this table. As these results suggest, such an estimate will suffer from marked imprecision for small m for λ not near zero or unity. The minimum ms for λ in Table 3 use the criterion that $\hat{\lambda}_m$ is within ±.1 of λ and therefore do not offer sufficiently precise λ estimates to determine a unique value of m for $\hat{h}_m$. Indeed the minimum m for $\hat{h}_m$ can differ by a factor of two or more given that $\hat{\lambda}_m$ is within ±.1 of λ. Naturally, one can increase the estimation precision for λ by increasing the number of imputations beyond those appearing in Table 3. However, the minimum ms to achieve greater precision can be prohibitive (e.g., for 95% confidence that $\hat{\lambda}_m$ is within .02 of λ, $m = 156$ for $\lambda = .1$ and $m = 847$ for $\lambda = .5$). Thus, maximizing estimation precision for λ may not be the most efficient method given that the ms needed to reduce imputation variability in $\hat{h}_m$ to criterion are far less than these numbers.

In lieu of a very precise estimate of λ with which one can use Table 3 to choose m to bring imputation variability in $\hat{h}_m$ and $\hat{p}_m$ to reasonable levels, the following conservative approach is suggested. First, compute a conservative estimate of λ based on one minus the ratio of the number of cases available for the estimation of a given parameter using listwise deletion (e.g., n_L) to the total sample size (i.e., $\hat{\lambda}_L = 1 - n_L/n$). This method is conservative in the sense that λ typically will be less than $\hat{\lambda}_L$ with the difference increasing as variables with missing values are more strongly related to the other available variables used in the imputation model. Next, consult Table 3 to choose m to minimize imputation variability in $\hat{h}_m$ to the desired confidence level using linear interpolation of the needed m between λ values as necessary and rounding to an integer value. The analysis is then conducted with the computed number of imputations.

To illustrate the method, consider the data used in Example 1 where $n_L = 19$ of the $n = 28$ patients had observed cholesterol measurements 14 days after myocardial infarction. Therefore, $\hat{\lambda}_L = 1 - 19/28 = .32$ for inference for μ_{14}. Assuming the desired confidence is 95% that $\hat{h}_m$ is within 10% of $\bar{h}_m$, then m should be between 24 (for $\lambda = .30$) and 59 (for $\lambda = .50$) from Table 3. Linear interpolation of these ms yields

$$m = 24 + (.32 - .30) \times \frac{59 - 24}{.50 - .30} = 27.5 \rightarrow 28$$

rounded to the nearest integer value.[10] Using 28 imputations in an independent analysis of the Example 1 data resulted in $\hat{h}_m = 19.42$, $\hat{p}_m = .03$, and $\hat{\lambda}_m = .20$. These estimates, in this case, are very close to their median values reported in Example 1 and one may affirm on inspection of Figure 1 for $m = 30$ that the variability in these statistics around their median values is relatively small, especially for $\hat{h}_m$ and $\hat{p}_m$. Note that the obtained $\hat{p}_m$ is close to $\alpha = .05$; as a result, the cautious researcher may wish to increase m (perhaps substantially) to minimize the chance of erroneous null hypothesis rejection or nonrejection due to imputation variability. Finally, note that λ is actually closer to .19 (estimated from $m = 800$ in Example 1) than $\hat{\lambda}_L = .32$, which illustrates the conservative nature of this approach. The sizable difference in estimates of λ is attributable in part to the large correlations among the variables (median $r = .67$). Data sets with smaller correlations among variables should expect smaller differences between $\hat{\lambda}_L$ and $\hat{\lambda}_m$.

Clearly, more systematic research is needed to evaluate and refine this approach. However, until such methods are available and evaluated, it is hoped that the simplicity of this method enables researchers to reduce imputation variability in $\hat{h}_m$ and $\hat{p}_m$ to reasonable levels. Furthermore, adopting guidelines based on results from a simulation and specific data sets must be made with some caution. In the simulations, missing data occurred on only one variable and the estimand of interest was a univariate population mean. Thus, more research is needed to explore the generality of these results and the sensibility of the presented recommendations for more complex situations (e.g., missing data on several variables and multivariate estimators such as multiple regression

[10] Let λ_0 be the smaller λ and m_0 be its corresponding m and let λ_1 be the larger λ and m_1 be its corresponding m. Then the linearly interpolated value of m for $\hat{\lambda}_L$ equals $m = m_0 + \left(\hat{\lambda}_L - \lambda_0\right) \frac{m_1 - m_0}{\lambda_1 - \lambda_0}$, which is rounded to the nearest integer.

slopes). However, there is reason expect some degree of generality for these results. Note that $B_\infty = U_\infty\lambda/(1-\lambda)$ both for univariate and multivariate estimators and note that only U_∞ was fixed in the simulations (i.e., λ was varied systematically and B_∞ is a function of λ and U_∞). Because a reasonably large sample size (or degrees of freedom) dominates the value of U_∞, these results and the recommendations that follow from them are anticipated to generalize to multivariate estimators.

CONCLUDING REMARKS

A current guideline suggests that using more than 10 imputations is rarely worth the additional effort for the analysis of incomplete data with multiple imputation (Rubin, 1987; Schafer, 1997a). This article questioned the optimism of this guideline and illustrated the problems that can surface if one follows it without considering the fraction of missing information λ. When λ is nontrivial and m is small (e.g., $\lambda > .2$ and $m \leq 10$), important quantities of interest (e.g., $\hat{\lambda}_m$, $\hat{h}_m$, and $\hat{p}_m$) can suffer from considerable imputation variability. As illustrated in Example 1, such variability within the same data set and for the same analysis can lead a researcher to categorically different conclusions about missing information, estimation precision, and null hypothesis rejection, respectively. This article demonstrated that more precise estimates of λ, CI widths, and NHST p values are obtained simply by increasing the number of imputations. Researchers can use Table 3 to choose the minimum m needed for specified precision at two levels of confidence using $\hat{\lambda}_L$, the conservative estimate of λ presented in the Discussion to the simulation experiments.

In many cases increasing m beyond 10 does not take much additional computer time or researcher effort given the prevalence of high-speed computing and automated routines for combining results for multiple imputation inference (e.g., PROC MIANALYZE in SAS). Researchers using more complex models where using a large number of imputations (e.g., $m > 10$) is not feasible should realize, as Schafer (1997a) noted, that $\hat{\lambda}_m$ is a rough guide that might be quite different than the actual λ and furthermore that CI widths and NHST p values may be substantially imprecise. The inferential impact of this imprecision will vary depending on the the sample size, the effect size in question, and the fraction of missing information. As illustrated with the larger sample size conditions of Example 2, imputation variability in $\hat{h}_m$ and $\hat{p}_m$ still exists when m is small with very large samples, nonzero effect sizes, and modest fractions of missing information. However the magnitudes of these variabilities might occur over a range far away from critical values (e.g., α for p values) and thus might have little substantive impact on conclusions regarding null hypothesis rejection and estimation precision. For typical sample sizes and effect sizes in social and behavioral research, often based on power analysis, the variability in $\hat{h}_m$ and $\hat{p}_m$ will tend to occur in a range nearer to these critical values. Therefore the impact of imputation variability on these values will need to be minimized by increasing the number of imputations.

In addition to increasing m, researchers can attempt to minimize the impact of imputation variability in other ways because the magnitude of this variability also depends on the number of cases with missing values and the conditional variance for each variable with missing values in the imputation model. Logically, $\hat{\theta}_k$ and $\hat{s}_k^2$ will exhibit less variability across imputations when fewer cases have missing values and when variables with missing data are strongly related to variables with observed data in the imputation model. However, minimizing imputation variability through these factors requires researcher foresight on the variables likely to suffer from missing values. Prior to data collection, researchers should plan to measure variables that are not likely to be missing and are strongly related to the variables expected to exhibit missing data. Doing so will reduce the variability in the pseudo-random imputed data values from the imputation model. During data collection, researchers should consider methods to minimize the amount of missing data (e.g., due to participant attrition, fatigue, confusion, or carelessness).

In conclusion, some brief comments on reporting standards, statistical training in substantive disciplines, and power and precision analysis are provided. Regarding reporting standards, there are no specific guidelines for stating research results based on data sets with missing values in the American Psychological Association's (2001) publication manual. Given that complete data sets are not common in my experience, such reporting standards are needed. Bodner (2006) surveyed the prevalence of missing data and manner of missing data treatment in psychological research and offered some guidance on reporting standards. The American Psychological Association's Task Force on Statistical Inference (Wilkinson & The Task Force on Statistical Inference, 1999) also provided guidance regarding missing data treatment. To these suggestions, one might add that researchers provide $\hat{\lambda}_m$ for each parameter of interest when using multiple imputation given the importance of this quantity.

Regarding statistical training, popular graduate-level statistics textbooks in psychology that address missing data techniques like multiple imputation (e.g., Tabachnick & Fidell, 2007) do not discuss the missing information concept in much detail, do not state the implications of varying degrees of missing information, and do not discuss imputation variability in important quantities (e.g., $\hat{\lambda}_m$, $\hat{h}_m$, and $\hat{p}_m$). Given that many researchers will learn to use missing data techniques from such sources without further study, textbook writers should consider these concepts in greater detail.

The final comment pertains to power and precision analysis. Although this article focused on variability in important inferential and descriptive quantities, the simulation experiment also demonstrated the effects of missing information on NHST *p* values and CI half-widths. These effects translate at times to substantial degradation in statistical power and estimation precision that is not resolved with even an infinite number of imputations. Thus, researchers who anticipate missing data and seek adequate statistical power and precision should factor these effects into their calculations. Research is needed to offer specific guidance on how this might be accomplished.

ACKNOWLEDGMENTS

I thank Mo Wang, the editor, and three anonymous reviewers for helpful comments on earlier versions of this article.

REFERENCES

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

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Bodner, T. (2006). Missing data prevalence and reporting practices. *Psychological Reports, 99,* 675–680.

Bodner, T., Derry, A., Haugen, J., McDonald, M., McLeod, A., & Wille, D. (2004, May). *A random sample survey of empirical research articles.* Poster presented at the annual conference of the American Psychological Society, Chicago.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, B, 39,* 349–374.

Harel, O. (2006, August). *Inferences on missing information under multiple imputation.* Paper presented at the Joint Statistical Meetings of the American Statistical Association. Seattle, WA.

Hershberger, S., & Fisher, D. (2003). A note on determining the number of imputations for missing data. *Structural Equation Modeling, 10,* 648–650.

Loken, E. (2004). Using latent class analysis to model temperament types. *Multivariate Behavioral Research, 39,* 625–652.

Longford, N. (2005). *Missing data and small-area estimation: Modern analytical equipment for the survey statistician.* New York: Springer.

Nickerson, R. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.

Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal, 4,* 227–241.

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

Ryan, B., & Joiner, B. (1994). *Minitab handbook* (3rd ed.). Belmont, CA: Wadsworth.

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

Schafer, J. (1997b). NORM Library for S-PLUS [Computer software]. Retrieved August 23, 2001, from http://www.stat.psu.edu/∼jls/misoftwa.html

Tabachnick, B., & Fidell, L. (2007). *Using multivariate statistics* (5th ed.). New York: Allyn & Bacon.

von Hippel, P. (2005). How many imputations are needed? A comment on Hershberger and Fisher (2003). *Structural Equation Modeling, 12,* 334–335.

von Hippel, P. (2007). Regression with missing y's: An improved method for analyzing multiply imputed data. *Sociological Methodology, 37,* 83–117.

Wilkinson, L., & The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.