THE IMPACT OF MISSING DATA ON SAMPLE RELIABILITY ESTIMATES: IMPLICATIONS FOR RELIABILITY REPORTING PRACTICES

CRAIG K. ENDERS
University of Nebraska, Lincoln

A method for incorporating maximum likelihood (ML) estimation into reliability analyses with item-level missing data is outlined. An ML estimate of the covariance matrix is first obtained using the expectation maximization (EM) algorithm, and coefficient alpha is subsequently computed using standard formulae. A simulation study demonstrated that the EM approach yields (a) less bias in reliability estimates, (b) dramatically reduces cross-sample fluctuation of estimates, and (c) yields more accurate confidence intervals. Implications for reliability reporting practices are discussed, and the EM procedure is demonstrated using a heuristic data set.

Keywords: *missing data; reliability; EM algorithm; maximum likelihood; reliability generalization*

A great deal of recent interest has been paid to maximum likelihood (ML) missing data methods in the methodological literature (e.g., Arbuckle, 1996; Collins, Schafer, & Kam, 2001; Enders, 2001a, 2001b; Enders & Bandalos, 2001; Graham, Hofer, & MacKinnon, 1996; Schafer & Graham, 2002; Wothke, 2000). Although this literature clearly suggests that ML estimation is far superior to traditional missing data techniques (MDTs) such as listwise (LD) and pairwise deletion (PD), the vast majority of these simulation studies have examined continuously measured scale scores.

One area that received relatively little attention in the missing data literature is internal consistency reliability estimation. To date, it appears that only three studies have examined the impact of missing data in this context. Downey and King (1998) examined two forms of mean imputation (MI),

Educational and Psychological Measurement, Vol. 64 No. 3, June 2004 419-436
DOI: 10.1177/0013164403261050

whereas McDonald, Thurston, and Nelson (1996) examined LD, MI, and regression imputation. Only a single study, Enders (2003), examined the use of ML estimation, and these results suggested that computing coefficient alpha from an expectation maximization (EM) algorithm covariance matrix can be far superior to traditional methods. Although some gain in accuracy was demonstrated using this approach, the variance of reliability estimates was greatly reduced.

Reliability Reporting Practices

Because reliability is not an indelible property of a test, Thompson and Vacha-Haase (2000) argued that "scholars conducting either substantive (i.e., nonmeasurement focused) or measurement research should be expected to report explicit and direct evidence of score integrity in their sample" (p. 188). For similar reasons, Fan and Thompson (2001) argued that confidence intervals should be reported in conjunction with sample reliability estimates and outlined existing approaches for doing so.

Given the ubiquitous role that reliability estimation plays in measurement and applied research, it is somewhat surprising that so little attention has been paid to missing data in this context. Not only are there few simulation studies that might inform methodological practice, but a casual inspection of published studies reveals that authors rarely explicitly acknowledge the presence of missing data, much less the method used to deal with it. Most typically, the reader is left to infer the presence of missing data based on degrees of freedom values that are inconsistent with the number of sampled cases. The presence of missing data in reliability analyses is virtually impossible to detect unless explicitly discussed because confidence intervals for reliability estimates—and the associated degrees of freedom values used to construct the intervals—are not yet widely reported. However, if software defaults are any indication, it is a reasonable assumption that LD is the most commonly invoked MDT in the published studies; the Statistical Package for the Social Sciences (SPSS) automatically implements LD in its reliability procedure and gives the user no other options, whereas the Statistical Analysis System (SAS) offers LD and PD. Based on the results of the subsequent simulation results, I will argue that missing data should be explicitly discussed in measurement and substantive studies that report sample reliability estimates.

Purpose

The purpose of this article is threefold. First, a two-step method for conducting reliability analyses with item-level missing data using ML estimation is outlined; a heuristic analysis is also included to demonstrate the ease with which this method can be applied. Second, simulation results are presented to illustrate the deleterious impact that traditional MDTs such as LD

can have on sample reliability estimates. Finally, implications for reliability reporting practices are discussed, with particular attention being paid to practices advocated by *Educational and Psychological Measurement*, namely reliability generalization methodology (RG) and the construction of confidence intervals around reliability estimates.

Theoretical Background

Rubin's (1976) theoretical work provided a taxonomy of missing data patterns. According to Rubin, values can be missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). To illustrate, consider a multi-item questionnaire designed to assess math self-efficacy in a college student population.

Missing values would be considered MCAR if the observed values represent a random sample of the hypothetically complete data set. That is, the propensity for missing values is unrelated to one's underlying level of self-efficacy to solve a given problem, and is also unrelated to other measured variables. In this case, missing values might result from random distractions, but could also be a purposive facet of the design. For example, suppose a researcher was providing validity evidence for the scores from a short form of the efficacy scale. In such a case, it might be reasonable to design the data collection such that a random subset of individuals are administered the full-length questionnaire and the remaining subjects are administered the short form. An excellent discussion of planned missingness such as this can be found in Graham et al. (1996).

In contrast, missing values can be characterized as MAR when the propensity for missing values is related to other measured variables but not on the underlying level of the trait being measured. For example, suppose that individuals with low levels of motivation are more likely to have skipped items on the self-efficacy questionnaire. Furthermore, within a given group of individuals sharing a common level of motivation there is no relationship between one's self-efficacy and the propensity for missing values. That is, there is no residual relationship between missingness and self-efficacy once motivation has been controlled for. It is important to note that MAR only holds if motivation scores are incorporated into the analysis of self-efficacy scores. As discussed below, this is quite straightforward in the current context.

Finally, data are said to be MNAR when missing values are related to the underlying level of the trait being measured. For example, suppose that students with low math self-efficacy choose to skip the items they lack confidence to solve. In this case, missing values are related to one's underlying level of self-efficacy. Alternatively, an MNAR pattern could result by excluding a measured variable that is related to the missingness. For example, consider the previous MAR example whereby missing values were related to

one's level of motivation. If motivation was measured but not included in the analyst's model, the missing data mechanism would then be characterized as MNAR, not MAR.

The distinction between Rubin's (1976) missing data mechanisms has important implications for the performance of different MDTs in applied practice. For example, traditional methods such as LD and PD require the MCAR condition and may yield biased parameter estimates in cases where the missing data pattern is not MCAR. In contrast, ML estimation requires the weaker MAR assumption, and should yield unbiased parameter estimates in a wider range of circumstances. Although LD and PD should be unbiased under MCAR, ML has the additional advantage of yielding estimates with less sampling variability (i.e., greater efficiency or power) in this case. These theoretical expectations have been supported by a number of recent empirical studies (e.g., Arbuckle, 1996; Collins et al., 2001; Enders, 2001a, 2001b; Enders & Bandalos, 2001; Graham et al., 1996; Schafer & Graham, 2002; Wothke, 2000).

Using ML Estimation for Reliability Analyses

Clearly, ML estimation has strong theoretical advantages over default MDTs, and it is quite straightforward to implement ML in the context of a reliability analysis using a two-stage analytic procedure. It is well known that coefficient alpha is simply a function of the number of scale items and item variances and covariances. This being the case, coefficient alpha could be computed using standard formulae after first obtaining an ML estimate of the covariance matrix. As demonstrated later in this article, this procedure can be implemented using free software and with little additional effort on the analyst's part.

The EM algorithm is perhaps the most straightforward method for obtaining an ML estimate of the covariance matrix in the first stage of the procedure; an equivalent method would be to use the implied covariance matrix from a saturated structural equation model using the widely available full information maximum likelihood (FIML) estimator for missing data. Briefly, the EM algorithm is a two-step iterative procedure whereby missing values are estimated, and a covariance matrix and mean vector are subsequently computed. To begin, a series of regression equations are constructed for the first E, or expectation, step. These regressions are used to impute missing values, and a residual term is added to each missing value to correct for random variability lost in the imputation process. After missing values have been imputed in the E step, an updated estimate of the covariance matrix and mean vector is subsequently obtained in the M, or maximization, step. The process begins anew with a second E step that utilizes regressions based on the updated covariance matrix, and the iterative process continues until there is a minimal difference between covariance matrices in adjacent M

steps. More detailed presentations of the EM algorithm can be found in Little and Rubin (1987) and, more recently, Allison (2002).

The results of a simulation study are now presented to more concretely illustrate the impact that missing data can have on sample reliability estimates. This simulation is not meant to be comprehensive in nature—a more detailed simulation study can be found in Enders (2003). Instead, the simulation results are meant to be demonstrative and to provide a basis for critically examining current reliability reporting practices.

Method

Simulation Design

Two simulation studies were performed to demonstrate the effects of missing data under an MCAR and MAR missing data pattern, respectively. The population data for the simulations were modeled after a validity study conducted by Rangel, Enders, and Delgado (2002). In this study, validity evidence was collected for scores on a seven-item measure of stereotype threat (Steele & Aronson, 1995). The covariance structure for the current simulation was based on the seven scale items as well as three additional scale scores that measured constructs theoretically related to stereotype threat (algebra exam performance, domain identification, and math self-efficacy).

The RANNOR function in SAS IML was used to generate a population data matrix that consisted of 25,000 cases. The population data matrix was subsequently transformed to have the covariance structure of the Rangel et al. (2002) data using Cholesky factorization. The continuous distributions for the seven scale items were then categorized into five ordered categories using threshold values that maintained distribution symmetry with skewness and kurtosis values of approximately zero. The three scale scores were left in their continuous form. The correlation matrix for the population data is shown in Table 1, and the population value of Cronbach's alpha was .832.

The simulation study was conducted by drawing 1,000 random samples (with replacement) of $N = 200$ from the population data matrix. The same set of 1,000 replicates was used to simulate two different missing data mechanisms: MCAR and MAR. For each missing data mechanism, a 20% missing data was imposed on four of the seven scale items (referred to herein as the target items), and no missing values were imposed on the three additional scale scores. Considering only the 200 by 7 data matrix of scale items, the missing values composed approximately 11% of the total number of data points.

The MCAR pattern was created by randomly deleting the desired percentage of observations from each of the four target items. This was accomplished by pairing the column vector of each target variable with a corre-

Table 1
Population Correlations Among Simulation Variables

	Variable	1	2	3	4	5	6	7	8	9	10
1.	Q1	1.01									
2.	Q2	0.31	1.02								
3.	Q3	0.42	0.41	1.01							
4.	Q4	0.42	0.36	0.57	1.01						
5.	Q5	0.35	0.40	0.57	0.66	1.02					
6.	Q6	0.27	0.25	0.43	0.58	0.44	1.02				
7.	Q7	0.12	0.32	0.49	0.52	0.44	0.40	1.01			
8.	Algebra	0.12	0.14	0.24	0.24	0.10	0.20	0.10	0.99		
9.	Domain	0.15	0.11	0.30	0.35	0.23	0.28	0.26	0.40	1.00	
10.	Efficacy	0.05	0.06	0.15	0.17	0.13	0.24	0.18	0.47	0.46	1.00

Note. Variable standard deviations are listed on the diagonal.

sponding column vector of uniform random numbers ranging between zero and unity. If the uniform random number u_{ik} was less than .20 (the desired missing data rate), the corresponding data point x_{ik} was removed.

An MAR pattern is one where missingness on a variable x is related to other observed variables but not on the values of x itself. To model this situation, a continuous selection variable, S, was generated as the average of the three additional scale scores. Again, each of the four target items was yoked to a vector of uniform random numbers. The MAR deletion process began by selecting the case i whose data vector contained the lowest value of S. Beginning with the first target scale item, the ith data point was deleted if $u_{ik} < .60$ (this deletion probability was chosen somewhat arbitrarily). The deletion process continued in ascending order until a 20% missing data rate was imposed, and the same procedure was repeated for the remaining three target items. This deletion process resulted in a situation where cases with low values on S (i.e., low math performance, domain identification, and self-efficacy) were more likely to have missing values on the target items than cases with moderate or high S values. In other contexts, this might mimic a scenario where individuals with little formal education or poor reading skills have a higher propensity to skip certain scale items.

MDTs

Within each of the 2,000 replications (1,000 samples by two missing data mechanisms), coefficient alpha was estimated using four MDTs: EM, LD, PD, and MI. For all MDTs, a covariance matrix was computed for each sample replicate, and coefficient alpha was computed from these matrices. The

EM algorithm was discussed previously, so no further details are needed. LD was carried out by removing all cases with missing values and computing the covariance matrix using the remaining complete cases. In contrast, each element of the PD covariance matrix was computed using all available cases for a given variable (i.e., variance) or variable pair (i.e., covariance). Finally, MI was implemented by substituting missing values on x_k with the arithmetic mean of x_k; the covariance matrix was subsequently computed using the imputed data.

Dependent Measures

As explained by Schafer and Graham (2002), the goal of a statistical procedure is to make unbiased and efficient inferences about the population parameter of interest, regardless of whether the data set has missing values. This means that a missing data procedure should be judged on its ability to accurately estimate the population parameter, not on its ability to estimate the hypothetically complete data set or reproduce the parameter estimates that would have been obtained had the data been complete. The outcome measures described below assess these criteria.

Bias. Bias was measured as the difference between the mean reliability estimate for the *j*th MDT and the corresponding population reliability, α.

Root mean square error (RMSE). RMSE is the average difference between $\hat{a}_j$ (the reliability estimate associated with the *j*th MDT) and α, and is given by

$$RMSE = \sqrt{\frac{\sqrt{\sum(\hat{\alpha}_j - \alpha)^2}}{1000}}.$$

In cases where $\hat{\alpha}_j$ is unbiased, RMSE quantifies the sampling variability (i.e., the standard deviation) of reliability estimates. For biased outcomes, the RMSE can be interpreted as a measure of overall accuracy that combines bias and variability into a single value.

Confidence interval coverage. Feldt, Woodruff, Salih, and Srichai (1986) gave upper and lower confidence interval limits for a sample estimate of coefficient alpha. Following their methodology, the 95% confidence interval around each reliability estimate was constructed, and confidence interval coverage was computed as the percentage of the 1,000 sample replications in which the 95% confidence interval contained the true population reliability. This process was repeated for each MDT.

Table 2
Mean Reliability Estimates and Bias for Each Missing Data Technique (MDT)

MDT	Mean	Bias	Min	Max
Complete data				
	.830	–.002	.767	.877
MCAR				
EM	.830	–.002	.762	.882
LD	.829	–.004	.704	.915
PD	.830	–.002	.758	.891
MI	.796	–.037	.720	.863
MAR				
EM	.832	.000	.757	.884
LD	.816	–.017	.695	.883
PD	.818	–.014	.727	.881
MI	.792	–.040	.700	.859

Note. MCAR = missing completely at random; EM = expectation maximization; LD = listwise deletion; PD = pairwise deletion; MI = mean imputation; MAR = missing at random. Population reliability value was .8324.

Results

Bias

Table 2 gives the mean reliability estimate and bias for each MDT. In the MCAR simulation, three MDTs (EM, LD, and PD) were essentially unbiased, and MI resulted in negatively biased reliability estimates. In the MAR simulation, only reliability estimates computed from the EM algorithm covariance matrix were unbiased. These findings are consistent with theoretical expectations and are not surprising.

RMSE

The RMSE values for each MDT are shown in Table 3. For comparative purposes, RMSE values for the 1,000 complete data matrices are also given in the table. Because MCAR bias was nil for EM, LD, and PD, the RMSE values for these MDTs essentially quantify the standard deviation of the reliability estimates. With the exception of PD, the EM estimates had much less sampling variation than the remaining MDTs.

To better illustrate these results, Figure 1 contains kernel density estimates of the empirical sampling distribution of reliability estimates for EM, LD, and MI. From the upper two panels of Figure 1, it is clear that EM and LD estimates are unbiased, as the graphs are centered approximately at the population reliability value, .832. In contrast, the MI distribution is obviously

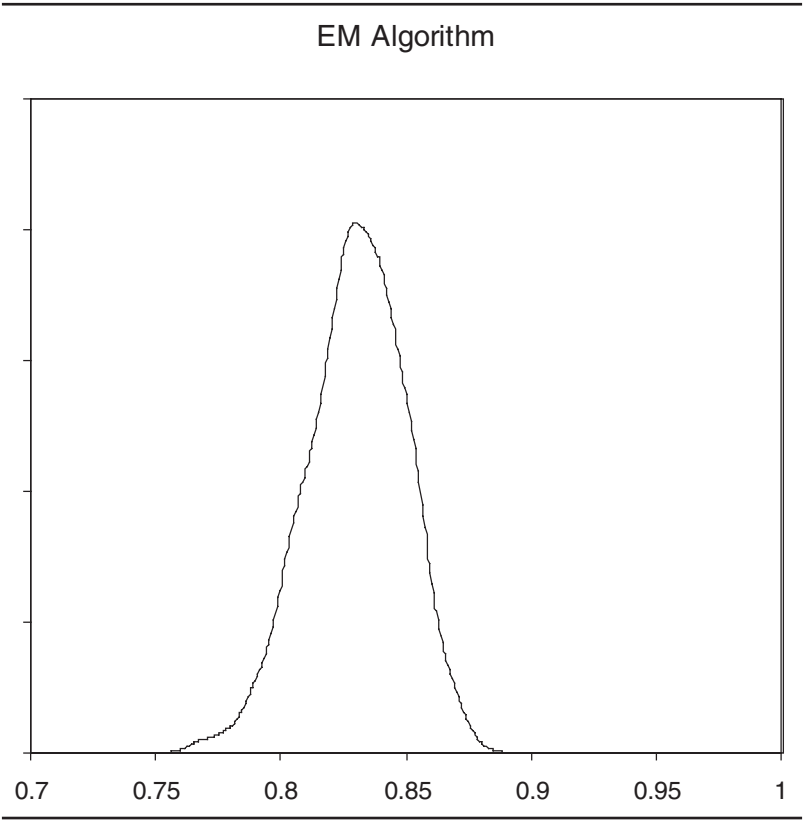

Figure 1. Kernel density estimates of the empirical sampling distribution of reliability estimates based on the expectation maximization (EM) algorithm, listwise deletion, and mean imputation.

Note. From the figure, it is clear that reliability estimates from the EM algorithm are much closer, on average, to the population reliability (.832), but also have much less sampling fluctuation than listwise estimates. In contrast, mean imputation estimates are fairly precise but negatively biased.

biased downward. Although EM and LD were unbiased, differences in the cross-sample variation of estimates were fairly dramatic. Clearly, the EM reliability estimates were much closer, on average, to the true population reliability. The ratio of the standard deviations for LD and EM ($\sqrt{.001/.00038} = \sqrt{2.63} = 1.62$) corresponds closely to the figure, which shows that LD reliability estimates are more than half again as spread out as EM estimates.

The MAR RMSE values are highly consistent with those from the MCAR simulation. However, because reliability estimates were biased for all MDTs except the EM algorithm, the RMSE values now reflect a combination of bias and sampling variation. Nevertheless, the conclusion is the same: EM reli-

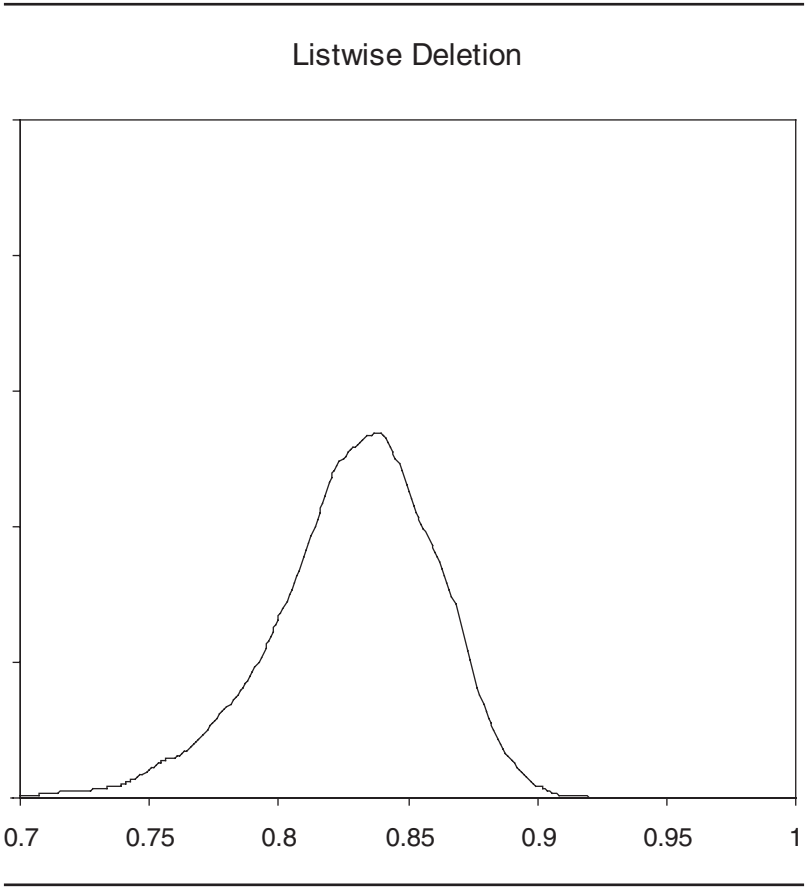

Figure 1 (Continued)

ability estimates had less cross-sample fluctuation and were closer, on average, to the true population reliability.

Confidence Interval Coverage

Table 3 gives 95% confidence interval coverage values for the reliability estimates of each MDT. As seen in the table, the EM algorithm yielded the most accurate coverage values across both simulations; coverage values were close to the advertised 95% rate. The superiority of EM estimates was particularly evident in the MAR simulation. Although coverage rates for LD and PD were reasonably accurate under an MCAR missing data pattern, they

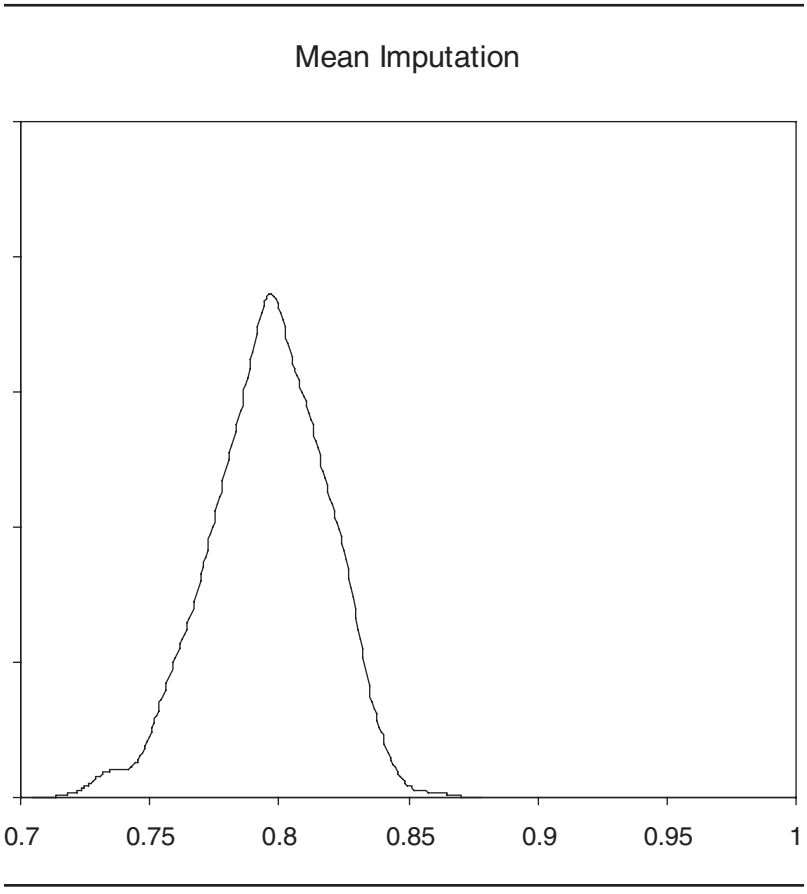

Figure 1 (Continued)

were unacceptably low in the MAR simulation (89.8% and 88.2%, respectively). In all cases, MI coverage rates were grossly inaccurate.

Although the EM algorithm coverage rates were clearly superior to those of other MDTs, it is important to note that these confidence intervals were based on $N = 200$. Due to missing data, elements within the EM covariance matrix are clearly estimated with different levels of precision. So although it is necessary to specify a single value of N to compute confidence intervals using the method outlined by Feldt et al. (1986), the correct choice of N is unclear. In any case, specifying N as the total number of sampled cases is probably not ideal, as it presumes better precision than was actually obtained—which explains why EM coverage values were slightly lower than the nominal 95% rate.

Table 3
Root Mean Square Error (RMSE) and Confidence Interval (CI) Coverage Values

	RMSE		95% CI Coverage	
MDT	MCAR	MAR	MCAR	MAR
Complete	.019	.019	94.9%	94.9%
EM	.020	.020	93.9%	93.1%
LD	.032	.032	93.4%	89.8%
PD	.021	.026	91.8%	88.2%
MI	.043	.046	61.5%	55.8%

Note. MDT = missing data technique; MCAR = missing completely at random; MAR = missing at random; EM = expectation maximization; LD = listwise deletion; PD = pairwise deletion; MI = mean imputation.

Table 4
Expectation Maximization (EM) Algorithm Confidence Interval Coverage for Different Values of N

Sample Size	MCAR	MAR
N	.939	.931
Min PV	.965	.961
Mean PV	.956	.952
Min PC	.983	.972
Mean PC	.966	.961

Note. MCAR = missing completely at random; MAR = missing at random; Min PV = minimum *N* per variance; Mean PV = mean *N* per variance; Min PC = minimum *N* per covariance; Mean PC = mean *N* per covariance. *N* = 200.

To further examine this issue, the 95% confidence intervals for the EM reliability estimates were recomputed using four different choices of sample size: the minimum *N* per variance, the harmonic mean *N* per variance, the minimum *N* per covariance term, and the harmonic mean *N* per covariance. Note that, in this case, the maximum *N* per variance and covariance is simply the full sample size of *N* = 200. Confidence interval coverage rates for each choice of *N* are displayed in Table 4. Although further studies need to clarify this issue, these results suggest that the accuracy of EM confidence intervals can be improved using the harmonic mean of the number of complete cases associated with each variable; this choice of *N* yielded coverage rates of 95.6% and 95.2% in the MCAR and MAR simulations, respectively.

Discussion and Implications

The primary goal of this article was to demonstrate the impact that missing data can have on sample reliability estimates and to outline a two-stage

estimation procedure that utilizes modern ML estimation. In the first stage, an ML estimate of the covariance matrix and mean vector are obtained (e.g., using the EM algorithm). The resulting covariance matrix is subsequently used as input into a reliability estimation program, and coefficient alpha is computed using standard formulae. The simulation results clearly support the use of a two-stage reliability estimation over traditional methods, particularly LD (perhaps the most commonly used MDT) and MI, and are consistent with those reported by Enders (2003).

The superiority of the EM approach has implications for reliability reporting practices in general but also relates to reporting practices advocated by *Educational and Psychological Measurement*, in particular RG methodology and reporting of confidence intervals around sample reliability estimates. These implications will now be discussed.

RG

In a relatively short period of time, Vacha-Haase's (1998) RG methodology has proven to be a popular meta-analytic technique for studying factors that influence cross-sample fluctuation of reliability estimates; there appear to be at least 28 RG studies that have been published or are currently in press (for an RG bibliography, see http://www.coe.tamu.edu/bthompson/religenr.htm). Unlike the characteristics typically examined by RG studies (e.g., gender, language), missing data cannot be viewed as a facet of substantive interest. Rather, they should be viewed as an unpredictable statistical nuisance, the effects of which should be minimized whenever possible. Nevertheless, the results shown in Table 3 clearly demonstrate the influence missing data can exert on the cross-sample fluctuation of reliability estimates. Because the presence of missing data is rarely discussed in published research studies—nor is the analyst's choice of MDT—it is impossible to quantify the impact of missing data on the fluctuation of reliability estimates in a given RG study. However, if one accepts that (a) missing data are a common problem, and (b) LD is probably the most widespread approach for handling the problem in this context, it is almost certain that missing data do contribute to the fluctuation of published reliability estimates.

Of course, the ideal outcome is to avoid the complication of missing data altogether. In that sense, one way to assess the performance of the two-stage EM approach is to examine how much additional sampling variability was incurred due to the missing data. From the MCAR results given in Table 3, a relative comparison of EM and complete data RMSE values, respectively, suggests that only a minor increase in sampling variation is accrued: $\sqrt{MSE_{EM} / MSE_{COMP}} = \sqrt{.00038/.00035} = 1.04$. The lack of practical significance in this case is bolstered by a relative comparison of LD and complete data RMSE values: $\sqrt{.001/.00035} = 1.69$. Although more research is clearly

needed, it would appear that RG studies—and the published reliability estimates they use as fodder—could essentially be purged of nuisance variation due to missing data if researchers begin to implement two-stage EM reliability estimation.

Confidence Intervals Around Reliability Estimates

A recent *Educational and Psychological Measurement* editorial by Fan and Thompson (2001) called for confidence intervals around sample reliability estimates and outlined existing approaches for doing so. However, the current results suggest that one's confidence in these intervals may, in fact, be low—in some cases very low—if the sample data contain missing values.

For LD and PD, the accuracy of confidence intervals was dependent on the mechanism that caused the missing data; coverage values were acceptable under an MCAR pattern but became unacceptably low when data were MAR. Although it has been argued that an MCAR pattern is unrealistic in applied studies (e.g., Graham et al., 1996; Muthn, Kaplan, & Hollis, 1987), researchers will rarely be privy to the mechanisms that caused the missing data. In any given study, missing values may result from a combination of MCAR, MAR, and MNAR mechanisms. Clearly, an MDT that is robust to the missing data mechanism is desirable, and the EM approach does appear to yield honest confidence intervals under a wider set of circumstances. Note that similar findings were observed by Enders (2003) under MNAR patterns as well. Although these results should be viewed as somewhat tentative, it also appears that the accuracy of EM confidence intervals can be further improved by using the harmonic mean of the number of complete cases associated with each variable when following Feldt et al.'s (1986) approach. Clearly, future research should clarify this issue.

Heuristic Illustration

Finally, it is important to underscore that reliability analyses using the EM approach do not require a great deal of additional time and effort, nor do such analyses require the purchase of specialized software. To illustrate the ease with which the procedure can be implemented, a reliability analysis was performed using a single sample replicate extracted from the MCAR simulation. The raw data are available upon request to readers who are interested in reproducing these results. Although an EM covariance matrix can be obtained using commercial software (e.g., SPSS Missing Values Analysis), I performed the initial missing data analysis using EMCOV (Graham & Hofer, 1993), a DOS-based freeware program that produces an ML covariance

matrix and mean vector using the EM algorithm. The EM covariance matrix could also be obtained from NORM (Schafer, 1999), a Windows-based freeware package for performing multiple imputation, a related procedure. Both EMCOV and NORM are freeware programs that can be downloaded from http://methcenter.psu.edu/mde.html.

In substantive and measurement studies alike, it is probably typical that researchers have item-level data for several questionnaires as well as continuously scored data from other variables of interest. Previous research has suggested that the precision of ML estimates may be enhanced by the inclusion of additional variables that are extraneous to the model being estimated (Collins et al., 2001). Although this issue has not been investigated in the context of reliability estimation, it seems reasonable to perform the initial EM analysis using a superset of variables that will ultimately be used in the reliability analysis. Not only might this improve precision, but it carries the additional benefit of making the MAR assumption (required by the ML estimator) more tenable. Consistent with this suggestion, the EM covariance matrix for the heuristic data was obtained using the entire set of 10 variables that included seven questionnaire items and three continuous scale scores.

The ML covariance matrix and mean vector were subsequently used as input into the SPSS reliability procedure, the syntax for which is given in the appendix. The resulting reliability estimate for the heuristic data was α = .8436. Following recommendations of Fan and Thompson (2001), the 95% confidence interval around the sample reliability estimate was also obtained. Based on the current simulation results, this interval was constructed using the harmonic mean sample size per variance. In this case, the *n*s associated with each variable were n = [200, 200, 200, 160, 166, 159, 155], resulting in a harmonic mean of approximately 175. The upper and lower confidence interval limits for this sample were .8766 and .8055, respectively, and were obtained from the SPSS reliability syntax given in the appendix.

Summary

In sum, the current results suggest that missing data can have a deleterious impact on the accuracy and efficiency of sample reliability estimates, as well as the confidence intervals constructed around those estimates. Researchers are encouraged to (a) explicitly discuss the number of missing values in the sample data matrix, and (b) abandon traditional MDTs in favor of ML estimation. These simulation results suggest that the use of a two-stage reliability estimation procedure incorporating the EM algorithm may, for practical purposes, negate the impact of missing data on sample reliability estimates. Although these findings are certainly promising, future studies should seek to expand the present findings, investigating a wider variety of conditions.

Appendix
SPSS Syntax for Conducting Reliability Analyses From an Expectation Maximization (EM) Covariance Matrix

```
*EM COVARIANCE MATRIX AND MEAN VECTOR FROM EMCOV PROGRAM.
matrix data var = q1 q2 q3 q4 q5 q6 q7 algebra domain efficacy
    /format = full
    /contents = mean cov n.
begin data
3.8050000 3.8350000 3.8350000 3.9509474 3.8001644 3.9000478 3.8375333 .0301049 .1351093 –.0326647
1.0924372 .3947990 .4601256 .4681271 .3854709 .2928376 .1459645 .1415792 .1164666 .0711347
.3947990 1.0329397 .4399749 .3357625 .3605959 .2788659 .2119496 .0536419 .0940783 .0528460
.4601256 .4399749 1.1334422 .6360364 .6691675 .5502479 .5675685 .1690422 .3018479 .1154337
.4681271 .3357625 .6360364 .9856013 .7048876 .5697274 .5314632 .1648437 .3842124 .1809840
.3854709 .3605959 .6691675 .7048876 1.0912329 .5092328 .4809614 .0561087 .2235549 .1754665
.2928376 .2788659 .5502479 .5697274 .5092328 .9104715 .3896696 .1657792 .2322104 .1999089
.1459645 .2119496 .5675685 .5314632 .4809614 .3896696 .9394434 .0942663 .3379829 .1718234
.1415792 .0536419 .1690422 .1648437 .0561087 .1657792 .0942663 .9749671 .3723062 .3640281
.1164666 .0940783 .3018479 .3842124 .2235549 .2322104 .3379829 .3723062 1.0828703 .3474897
.0711347 .0528460 .1154337 .1809840 .1754665 .1999089 .1718234 .3640281 .3474897 .7511411
175 175 175 175 175 175 175 175 175 175
end data.
*CONVERT EM COVARIANCE MATRIX TO CORRELATION MATRIX FOR SPSS RELIABILITY.
mconvert in(*) out(*).
run SPSS reliability procedure
reliability var q1 to q7
/scale(sts) = q1 to q7
/matrix = in(*)
/statistics = corr cov
/summary = means var total
/icc = model(random) type(consistency) cin = 95
/model = alpha.
```

References

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

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243-277). Mahwah, NJ: Lawrence Erlbaum.

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

Downey, R. G., & King, C. V. (1998). Missing data in Likert ratings: A comparison of replacement methods. *Journal of General Psychology*, *125*, 175-191.

Enders, C. K. (2003). Using the EM algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychological Methods, 8,* 322-337.

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

Enders, C. K. (2001b). The performance of the full information maximum Likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement*, *61*, 713-740.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*, 430-457.

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An *EPM* guidelines editorial. *Educational and Psychological Measurement*, *61*, 517-531.

Feldt, L. S., Woodruff, D. J., Salih, F. A., & Srichai, M. (1986). *Statistical tests and confidence intervals for Cronbachs coefficient alpha*. Iowa Testing Programs Occasional Papers Number 33. (ERIC Document Reproduction Service No. ED 291 755)

Graham, J. W., & Hofer, S. M. (1993). *EMCOV reference manual* [Computer software]. Los Angeles: University of Southern California, Institute for Prevention Research.

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

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

McDonald, R. A., Thurston, P., & Nelson, M. (1996). *A Monte Carlo study of missing item procedures*. Paper presented at the Southern Academy of Management meetings, New Orleans, LA.

Muthn, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*, 431-462.

Rangel, A., Enders, C. K., & Delgado, R. (2002, April). *The development of the Stereotype Threat Scale: Measuring and testing the theoretical construct of stereotype threat*. Poster session presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

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

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

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

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African-Americans. *Journal of Personality and Social Psychology*, 69, 797-811.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*, 174-195.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6-20.

Wothke, W. (2000). Longitudinal and multigroup modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multiple group data: Practical issues, applied approaches and specific examples* (pp. 219-240). Mahwah, NJ: Lawrence Erlbaum.