

A Two-Stage Approach to Missing Data: Theory and Application to Auxiliary Variables

Victoria Savalei

Peter M. Bentler

University of California Los Angeles

## Abstract

A popular ad-hoc approach to conducting structural equation modeling (SEM) with missing data is to obtain a saturated maximum likelihood (ML) estimate of the sample covariance matrix and then to use this estimate in the complete data ML fitting function to obtain parameter estimates. This two-stage approach is appealing because it minimizes a familiar function while being only marginally less efficient than the standard ML (“direct ML”) approach. Importantly also, the two-stage approach allows for easy incorporation of auxiliary variables, which can mitigate bias and inefficiency due to missing data. Incorporating auxiliary variables with direct ML is not straightforward and requires setting up a special model. However, standard errors and test statistics provided by the complete data routine when using the two-stage approach will not be correct. Empirical approaches to finding the right corrections have failed to provide unequivocal solutions. In this paper, we develop theoretical formulas for the correct standard errors and test statistics for the two-stage approach and its extension to include auxiliary variables. Since these accurately reflect the variability of the two-stage estimator, the actual sample size multiplier  $n$  can be used, and no adjustments are necessary. We study the performance of the two-stage test statistics and standard errors in a small simulation study. We find that not only does the new two-stage approach perform well in all conditions, but that one of its associated test statistics outperforms the direct ML statistic, deemed optimal in the missing data literature. We call for an incorporation of this new and promising missing data method into standard SEM software for further study.

## Introduction

Structural equation modeling (SEM) is a popular multivariate data analysis technique among psychology researchers, as it allows for complex theory testing (e.g., Raykov & Marcoulides, 2006). More often than not, applied researchers work with datasets that contain missing values. The topic of how to properly handle missing data in multivariate statistical analyses has received a lot of attention in recent years (e.g., Little & Rubin, 2002; Schafer, 1997). In structural equation modeling (SEM), the old ad-hoc approaches, such as listwise and pairwise deletion, hot deck imputation, and so on, are no longer deemed acceptable, while the new approaches such as ML (also called direct ML, raw ML, or FIML) and multiple imputation (MI) are increasingly recommended as most appropriate (Arbuckle, 1996; Enders & Bandalos, 2001; Schafer & Graham, 2002). The direct ML approach is implemented in many popular SEM programs, such as Amos (Arbuckle, 2007), EQS (Bentler, 2007), LISREL (Joreskog & Sorbom, 1997), and *Mplus* (Muthen & Muthen, 2006), thus making it easy for psychology researchers to handle missing data properly. The MI approach, which creates several completed datasets and then conducts inferences by averaging across them, requires a bit more thought to implement but has also been made available in programs such as LISREL and NORM (Schafer, 1999). The MI approach is slightly less efficient than the ML approach if the number of imputations is small, but it has been argued to have other advantages.

An important advantage of the MI approach is that it allows for easy incorporation of *auxiliary variables*, which are additional variables that are not in the model but that are related either to the cause of missingness or to the missing values themselves. Including such variables can increase efficiency of the estimator when data are missing at random (MAR; see Little & Rubin, 2002). Additionally, and perhaps more importantly, including auxiliary variables can

reduce bias introduced by missing data when the data are missing not at random (MNAR), because the presence of auxiliary variables can change the mechanism to MAR (Schafer and Olsen, 1998; Collins, Schafer, & Kam, 2001). The direct ML approach can be adapted to include auxiliary variables as well. Graham (2003) proposed several ways of doing so. The “saturated correlates” model, which imposes a saturated covariance structure on the auxiliary variables and their relationship to other independent variables in the model, including error variables, provided the best solution. But several complications remain. First, setting up this model is nontrivial and requires following a set of rules (Graham, 2003). Second, if the number of auxiliary variables is large, or if the residual variances associated with some of the variables are small, the saturated correlates model may run into numerical trouble. Third, the independence model produced automatically by the software and required for some fit indices will be incorrect, and a separate run is required to estimate it appropriately (Graham, 2003). Until software developers provide a way to automate the “saturated correlates” procedure, researchers using direct ML may be tempted to follow the “restrictive” (omitting auxiliary variables) rather than the “inclusive” strategy (Collins et. al., 2001). Finally, there exist situations when the “saturated correlates” model is theoretically inappropriate, as will be illustrated in the discussion, yet inclusion of auxiliary variables still makes sense.

Alternatives to the “saturated correlates” direct ML estimator exist. The so-called two-stage method, following Yuan and Bentler (2000), allows for easy incorporation of auxiliary variables and does not run into any of the problems associated with the “saturated correlates” estimator. This method first obtains the saturated ML estimate of the entire population covariance matrix (i.e., “the EM covariance matrix”<sup>1</sup>) including auxiliary variables if they are

---

<sup>1</sup> The most common way to obtain the saturated ML estimate of the population covariance matrix is via the EM algorithm—hence the term the “EM covariance matrix.” We dislike this terminology because it confuses the

present; it then uses the (sub)matrix pertaining to the variables in the model in the complete data ML fit function (Brown, 1983; Graham, 2003; Rovine, 1994). This approach, in its original form, falls somewhere in between the ad-hoc approaches such as pairwise deletion and statistically justified approaches such as direct ML. It is better than listwise and pairwise approaches because it obtains a proper ML estimate of the covariance matrix that is consistent under MCAR and MAR normal data (pairwise/listwise deletion only produces consistent estimates under MCAR data) and because it produces more efficient estimates (Brown, 1983). It is worse than direct ML because it will produce consistent parameter estimates but incorrect standard errors and test statistics (Graham, 2003; Enders & Peugh, 2004). The reasons for this are obvious: the uncertainty associated with missing data is never incorporated into the standard errors and test statistics because the input matrix is treated as if it were obtained from complete data. The goal of this paper is to introduce a new statistically justified version of the two-stage approach by providing a way to compute the correct standard errors and asymptotically chi-squared test statistics, so that valid inference used two-stage estimates can be conducted.

As previously mentioned, an important advantage of the two-stage approach, other than its obvious simplicity, is that it allows for easy incorporation of auxiliary variables (Graham, 2003). These variables are trivially added during the first stage, when the saturated ML estimate of the population covariance matrix is obtained, and are ignored in the second stage, when the complete data fit function is minimized. The two-stage parameter estimates are also easily obtained: programs such as SPSS and EMCOV (Graham & Hofer, 1995) compute the saturated ML estimate of the covariance matrix with missing data, and the resulting matrix or portion

---

estimator with how it was obtained: after all, no one refers to complete data ML estimates as “Newton-Raphson” estimates. The ML covariance matrix with missing data can be obtained via other optimization algorithms, and the EM algorithm is not restricted to the saturated model and can be used to optimize the structured model (Jamshidian & Bentler, 1999). However, rather than changing the accepted terminology, we employ the quotation marks.

thereof can be fed into any SEM program. The two-stage parameter estimates were studied by Graham (2003), who found that they were very similar to direct ML estimates across four different studies, and the theoretical loss of efficiency was in practice minimal: observed differences in empirical standard errors were in the third decimal place.

Graham did not, however, study estimated standard errors or test statistics, which would be incorrect as obtained directly from the software. Enders and Peugh (2004) provided such a study. They found that the standard errors associated with the two-stage approach were underestimated, but that this bias was often minimal, and coverage rarely fell below 90% at  $\alpha = .05$ . However, the ML chi-square statistic was very adversely affected. Across the studied conditions, the rejection rates ranged from 28% to 100% (instead of the nominal 5%), and were worse for higher proportions of missing data. As an example, in their Study 1, with  $n=600$  and with 15% of data missing, the rejection rate was 92% instead of 5%. The goal of Enders and Peugh's (2004) work was to also study various adjustments to the value of  $n$  that would yield more reasonable confidence intervals and rejection rates. However, no one adjustment was found to work well across all studied conditions, suggesting that the two-stage method may not be rescued via empirically obtained corrections. A suggestion to bootstrap the standard errors for the 2-stage approach has also been made (Graham et. al., 1997). A theoretical development also exists, however. Yuan and Bentler (2000) gave the correct formulae for asymptotic standard errors for the two-stage estimates and proposed scaled and residual-based statistics based on the two-stage estimator. Their results are limited to MCAR data, and while their formulae cover the general case when the data may be nonnormal, they do not cover auxiliary variables, which is when the two-stage method may be especially useful.

The present paper proceeds as follows. First, due to its omission from the relevant recent literature, we review the existing theory for the two-stage estimator (Yuan & Bentler, 2000). Second, we extend these results to normally distributed MAR data, as current statistical theory establishes consistency of the “EM covariance matrix” with both MCAR and MAR data under normality (Little & Rubin, 2002). In particular, we propose a normal theory based scaled statistic (a version of the “Satorra-Bentler” scaled chi-square) and a residual-based statistic to evaluate model fit. The scaled statistic will have its mean equal to that of a chi-square variate, while the residual-based statistic will behave asymptotically as chi-square as long as they data are normal. Third, we extend our approach to incorporate auxiliary variables. To evaluate the usefulness of these theoretical developments, we conduct a small simulation study that replicates the conditions of Enders and Peugh (2004). We summarize the results and provide a discussion of this new approach to missing data.

### **Inference based on the two-stage estimator**

First we develop the two-stage estimator theory for the standard case when there are no auxiliary variables. The appropriate way to conduct inference based on this estimator is not fully developed even in this case. Yuan and Bentler (2000) give the results for MCAR data. Here we extend them to allow for MCAR and MAR normally distributed data. Extensions to nonnormal data will be developed and studied elsewhere.

Let  $x_1, \dots, x_n$  be a random sample from a  $p$ -variate normal distribution  $N(\mu(\theta), \Sigma(\theta))$ , but possibly not fully observed. This means that for each  $i = 1, \dots, n$ , the corresponding observation vector  $x_i$  is of dimension  $p_i \times 1$ , where  $p_i \leq p$ . We are interested in the  $q \times 1$  parameter vector  $\theta$ .

Two-stage estimation proceeds as follows. In stage 1, we obtain the saturated ML estimate of the population covariance matrix, by maximizing the log-likelihood:

$$l(\beta | x_1, \dots, x_n) = \sum_{i=1}^n l_i(\beta) = C - \frac{1}{2} \sum_i \log |\Sigma_i| - \frac{1}{2} \sum_i (x_i - \mu_i)' \Sigma_i^{-1} (x_i - \mu_i), \quad (1)$$

where  $\beta = (\text{vech}\Sigma', \mu')'$ ,  $p_i$  is the length of  $x_i$ ,  $\mu_i$  is the relevant subvector of  $\mu$ ,  $\Sigma_i$  is the relevant  $p_i \times p_i$  submatrix of the  $p \times p$  model covariance matrix  $\Sigma$ ,  $C$  is a constant, and “vech” is the operator that picks out the nonredundant elements of a symmetric matrix columnwise (Magnus & Neudecker, 1999). Let  $\hat{\beta}$  be the saturated estimator obtained by maximizing (1). We will also write  $\hat{\mu}_\beta = \mu(\hat{\beta})$  and  $\hat{\Sigma}_\beta = \Sigma(\hat{\beta})$  to represent the saturated ML estimates of means and of covariances. These are the “EM means and covariance matrix.” In addition to SEM software that does appropriate ML estimation with missing data, these estimates can be obtained using programs such as SPSS or Emcov. A crucial assumption is that the saturated estimates  $\hat{\beta}$  are consistent for the population mean and covariance matrix. This assumption is met for normally distributed MCAR and MAR data (Little & Rubin, 2002).

In stage 2, we use the saturated ML estimates  $\hat{\beta}$  in the usual complete data log-likelihood. In other words, we minimize:

$$F_{ML}(\theta) = \text{tr}\{\hat{\Sigma}_\beta \Sigma^{-1}(\theta)\} - \log\{|\hat{\Sigma}_\beta \Sigma^{-1}(\theta)|\} + (\hat{\mu}_\beta - \mu(\theta))' \Sigma^{-1}(\theta) (\hat{\mu}_\beta - \mu(\theta)), \quad (2)$$

where the usual sample estimates  $\bar{x}$  and  $S$  have been replaced by saturated ML estimates. We denote the minimizer of (2) by  $\tilde{\theta}$ . This is the two-stage estimator. For comparison, the direct ML approach would instead proceed by maximizing (1) again for the structured model.

To conduct inference based on the two-stage estimator, we only require the asymptotic covariance matrix (ACM) of the saturated estimator  $\hat{\beta}$ ,  $\text{acov}(\sqrt{n}\hat{\beta})$ . Under the assumption of



normality, this is the inverse of the information matrix for the saturated model. We write

$\text{cov}(\sqrt{n}\hat{\beta}) = A_{\beta}^{-1}$ , where

$$A_{\beta} = -\lim_{n \rightarrow \infty} \frac{1}{n} \sum_i \frac{\partial^2 l_i(\beta)}{\partial \beta \partial \beta'} \Big|_{\beta = \beta_0}. \quad (3)$$

As with complete data, the estimated information matrix in the sample can be “expected” or “observed.” To compute the “expected” information would require knowing the limit, or the expected value, of the second derivatives of the saturated log-likelihood. This is only possible if we make the assumption of homogeneity of means and covariances (HMC) across missing data patterns, which is a consequence of the MCAR assumption (Kim & Bentler, 2002). “Observed” information, on the other hand, would be computed by dropping the limit notation and simply evaluating the matrix of second derivatives at the saturated estimates. When the data are MCAR, both “expected” and “observed” information matrices will be consistent, although “observed” information may still be preferred (Dolan & Molenaar, 1991), whereas when the data are MAR, generally only the “observed” information matrix will be consistent. The use of “observed” information with missing data was first advocated on the same theoretical grounds by Kenward and Molenberghs (1998). In practice, it is not known how much the additional computational complexity of “observed” information will affect its small sample stability.

The “expected” information is computed as (Yuan & Bentler, 2000):

$$\hat{A}_{\beta, EXP} = \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} .5\kappa_i' (\hat{\Sigma}_{\beta, i}^{-1} \otimes \hat{\Sigma}_{\beta, i}^{-1}) \kappa_i & 0 \\ 0 & \tau_i' \hat{\Sigma}_{\beta, i}^{-1} \tau_i \end{pmatrix}, \quad (4)$$

where  $\kappa_i = \frac{\partial \text{vec} \Sigma_i}{\partial \sigma'}$  and  $\tau_i = \frac{\partial \mu_i}{\partial \mu'}$  are 0-1 matrices that select only those rows/columns that are

present for a particular observation  $i^2$ . The ‘‘observed’’ information matrix is computed as:

$$\hat{A}_{\beta, OBS} = \sum_{i=1}^n \begin{pmatrix} \kappa_i' [\hat{\Sigma}_{\beta, i}^{-1} \otimes \{\hat{\Sigma}_{\beta, i}^{-1} (x_i - \mu_i)(x_i - \mu_i)' \hat{\Sigma}_{\beta, i}^{-1} - .5 \hat{\Sigma}_{\beta, i}^{-1}\}] \kappa_i & -- \\ \tau_i' [\hat{\Sigma}_{\beta, i}^{-1} \otimes (x_i - \mu_i)' \hat{\Sigma}_{\beta, i}^{-1}] \kappa_i & \tau_i' \hat{\Sigma}_{\beta, i}^{-1} \tau_i \end{pmatrix}. \quad (5)$$

The complete data formulae for the components of the observed information matrix were given in Yuan and Hayashi (2006), but the incomplete data version given by (5) has not, to our knowledge, previously appeared in the literature. Note that if we could make the HMC assumption, i.e., that  $E(x_i) = \mu_i$  and that  $E(x_i - \mu_i)(x_i - \mu_i)' = \Sigma_i$  for each  $i$ , taking expectations of (5) would reduce to (4). The formulas for standard errors and test statistics we give next can utilize either the ‘‘expected’’ or the ‘‘observed’’ information, depending on the researcher’s beliefs about the nature of the missing mechanism, and we will denote both estimators by  $\hat{A}_{\beta}$ .

The correct asymptotic covariance matrix of the two-stage estimator  $\tilde{\theta}$  is computed as:

$$a \text{cov}(\sqrt{n} \tilde{\theta}) = A_{\tilde{\theta}}^{-1} = (\dot{\beta}' H \dot{\beta})^{-1} \dot{\beta}' H A_{\beta}^{-1} H \dot{\beta} (\dot{\beta}' H \dot{\beta})^{-1}, \quad (6)$$

where  $\dot{\beta} = \frac{\partial \beta(\theta)}{\partial \theta'}$  is the matrix of model derivatives and  $H$  is the complete data information

matrix

$$H = \begin{pmatrix} .5 D_p' (\Sigma^{-1} \otimes \Sigma^{-1}) D_p & 0 \\ 0 & \Sigma^{-1} \end{pmatrix}, \quad (7)$$

where  $D_p$  is the duplication matrix (Magnus & Neudecker, 1999). Standard errors for the two-stage estimator can thus be obtained by using sample estimates in (6), computing  $\hat{A}_{\beta}$  following

---

<sup>2</sup> A useful computational relationship between these matrices is  $\kappa_i = (\tau_i \otimes \tau_i) D_p$ , where  $D_p$  is the duplication matrix (Magnus & Neudecker, 1999).

either (4) or (5),  $\hat{H}$  by using the model implied covariance matrix  $\hat{\Sigma}$  in (7), and evaluating the matrix of model derivatives at the two-stage estimates. Because (6) gives the correct asymptotic variability of the two-stage estimator, no corrections to the sample size value are needed, and the diagonal elements of (6) should simply be divided by  $n$  before taking the square root.

It is worth noting that  $(\dot{\beta}'\hat{H}\dot{\beta})^{-1}$  is the “naïve” covariance matrix estimate, obtained via complete data routines. The standard errors obtained from the diagonals of this “naïve” estimate will be underestimates of the true standard errors given by (6). On the other hand, the asymptotic covariance matrix of the FIML estimator, let us call it  $\hat{\theta}$ , is given by  $(\dot{\beta}'A_{\beta}\dot{\beta})^{-1}$ , and we immediately have the result that the two-stage estimator is less efficient than the FIML estimator because the matrix in (6) is “larger” than this matrix (the difference is positive definite; Ferguson, 1996, Corollary to Theorem 23). Graham (2003) found that in practice the loss of efficiency of  $\tilde{\theta}$  relative to  $\hat{\theta}$  was extremely minor.

To evaluate the fit of the model to the data, we can compute a scaled test statistic (Satorra & Bentler, 1994):

$$T_{2st,NTSB} = ncF_{ML}(\tilde{\theta}), \quad (8)$$

where  $F_{ML}(\tilde{\theta})$  is the minimum of the fit function in (2), and the scaling factor is

$c = (p^* - q) / \text{tr}\{\hat{A}_{\beta}^{-1}(\hat{H} - \hat{H}\dot{\beta}(\dot{\beta}'\hat{H}\dot{\beta})^{-1}\dot{\beta}'\hat{H})\}$ . The purpose of the scaling correction is to bring

the mean of the original inflated statistic,  $nF_{ML}(\tilde{\theta})$ , closer to that of a chi-square variate. The

asymptotic distribution of (8), however, is not chi-square, so this statistic is only an

approximation. When similar scaling is used to account for nonnormality (i.e., the “Satorra-

Bentler scaled chi-square”), the resulting approximation has been found to work quite well

(Curran, West, & Finch, 1996; Hu, Bentler, & Kano, 1992). However, this is quite a different

situation, where (8) still assumes normality like the original statistic, but corrects for the fact that complete data routines were used to analyze the “EM covariance matrix.” The performance of the scaling correction in this situation needs to be studied.

A second statistic that can be used to assess the fit of the model is the residual-based statistic, originally proposed to account for nonnormality (Browne, 1984). A normal-theory version of it is given by:

$$T_{2st,NTres} = (\hat{\beta} - \beta(\tilde{\theta}))'(\hat{A}_\beta - \hat{A}_\beta \hat{\beta}(\hat{\beta}' \hat{A}_\beta \hat{\beta})^{-1} \hat{\beta}' \hat{A}_\beta)(\hat{\beta} - \beta(\tilde{\theta})) \quad (9)$$

Unlike (8), this statistic is asymptotically chi-square distributed, and thus may perform better. Nonnormal versions of the residual-based statistics were found to perform very similarly to the ADF test statistics, in that they had very inflated rejection rates at smaller sample sizes (Hu, Bentler, & Kano, 1992; Yuan & Bentler, 1998). However, we do not believe the normal-theory version of the residual-based statistic given by (9) will suffer from this limitation, because the normality assumption stabilizes the computations.

We have proposed a statistically justified way to conduct inference based on the two-staged estimator, using “robust” standard errors and scaled or residual-based test statistics. This inference will be valid with normally distributed MCAR or MAR data. We note that these developments have not been studied even in the standard case with no auxiliary variables. However, the advantage of the two-stage approach is that it allows for easy incorporation of such variables, a case we consider next.

### **Inference based on the two-stage estimator with auxiliary variables**

As before,  $x_1, \dots, x_n$  is a random sample from a p-variate normal distribution  $N(\mu(\theta), \Sigma(\theta))$ , possibly not fully observed. But now let the variables  $a_{i1}, \dots, a_{ik}$  be a set of

*auxiliary* variables observed on each individual, that is, variables that the researcher is not interested in modeling but that may relate to or explain missing data. In particular, if these additional variables are correlated either with the set of variables  $x_i$  or are predictive of their missingness, they have the potential to increase efficiency or to reduce bias, respectively. It is not necessary to assume that auxiliary variables are themselves fully observed. The two-stage approach easily allows for missing data on the main variables and on the auxiliary variables, which is probably the case with most real life datasets.

$$\text{Let } \mu_{big} = \begin{pmatrix} \mu \\ \mu_a \end{pmatrix} \text{ and } \Sigma_{big} = \begin{pmatrix} \Sigma & \Sigma_{ax}' \\ \Sigma_{ax} & \Sigma_{aa} \end{pmatrix} \text{ be the mean and covariance matrix of the joint}$$

collection of variables  $x_{i1}, \dots, x_{ip}, a_{i1}, \dots, a_{ik}$ . Let the vector of saturated model's parameters be

$\beta_{big} = (\text{vech}\Sigma_{big}', \mu_{big})'$ . To create a more useful ordering, re-arrange the elements so that the

parameter vector of interest,  $\beta = (\text{vech}\Sigma', \mu)'$ , appears first, and the nuisance parameters

involving auxiliary variables follow. Denote the rearranged version of  $\beta_{big}$  by

$$\beta_{big}^* = (\text{vech}\Sigma', \mu', \text{vech}\Sigma_{ax}', \text{vech}\Sigma_{aa}', \mu_a')' = (\beta', \gamma')'.$$

The two-stage estimation with auxiliary variables proceeds as follows. In stage 1, we obtain the saturated estimate of  $\beta_{big}^*$  call it  $\hat{\beta}_{big}^* = (\hat{\beta}', \hat{\gamma}')'$ , by maximizing the saturated log-likelihood:

$$l(\beta_{big}^*) = C - \frac{1}{2} \sum_i \log |\Sigma_{big,i}| - \frac{1}{2} \sum_i (x_i - \mu_{big,i})' \Sigma_{big,i}^{-1} (x_i - \mu_{big,i}). \quad (10)$$

This equation is identical to (1), but it uses the larger covariance matrix and vector of means that include the additional variables. Stage 2 is the same as before: we minimize the complete data

ML fitting function in (2), but with  $\hat{\Sigma}_\beta$  and  $\hat{\mu}_\beta$  being the appropriate  $p \times p$  submatrix and

$p \times 1$  subvector of the larger  $(p+k) \times (p+k)$  covariance matrix and  $(p+k) \times 1$  vector of means obtained in (10). In other words, stage 2 does not do anything with auxiliary variables. As before, denote the two-stage estimator by  $\tilde{\theta}$ .

There are two ways in which the estimator  $\hat{\beta}$  is “better” when auxiliary variables are included. First, a necessary assumption to conduct any sort of analyses is that  $\hat{\beta}$  is consistent for  $\beta$ . We can never test this assumption, but it is more likely to be met when auxiliary variables are included than when they are not. If missingness can be completely explained by the auxiliary variables, their inclusion makes the missing mechanism MAR, which insures consistency. A more realistic scenario is that auxiliary variables make the missing mechanism “more MAR,” in that they can predict which values are missing to some nontrivial degree. This will mean that  $\hat{\beta}$  obtained with auxiliary variables is inconsistent, but the bias is less than if no auxiliary variables were used. Second, assuming consistency,  $\hat{\beta}$  may be more efficient (i.e., have smaller covariance matrix) when auxiliary variables are included, because the auxiliary variables that are correlated with variables that contain missingness will be predictive of the missing values themselves, reducing uncertainty due to missing data. These improved properties of the saturated estimator  $\hat{\beta}$  will be inherited by the structured estimator  $\tilde{\theta}$ .

Under the normality assumption, the asymptotic covariance matrix of  $\hat{\beta}_{big}^*$  is given by

$a \text{cov}(\sqrt{n}\hat{\beta}_{big}^*) = A_{\beta big}^{*-1}$ , where

$$A_{\beta big}^* = -\lim_{n \rightarrow \infty} \frac{1}{n} \sum_i \frac{\partial^2 l_i(\beta_{big}^*)}{\partial \beta_{big}^* \partial \beta_{big}^*} \Big|_{\beta_{big}^* = \beta_{big,0}^*} = \begin{pmatrix} A_\beta & A_{\gamma\beta}' \\ A_{\gamma\beta} & A_\gamma \end{pmatrix}. \quad (11)$$

The asymptotic covariance matrix of the parameter vector of interest,  $\hat{\beta}$ , is given by

$a \text{cov}(\sqrt{n}\hat{\beta}) = A_{\beta big}^{*11}$ , where

$$A_{\beta big}^{*11} = (A_{\beta} - A_{\gamma\beta}' A_{\gamma}^{-1} A_{\gamma\beta})^{-1}, \quad (12)$$

which is the upper block of the inverse of (11) (e.g., Magnus & Neudecker, 1999, p. 11, eq. 6).

Note that this equation only involves inverses of  $p \times p$  and  $k \times k$  matrices.

To obtain computational formulas for “expected” and “observed” information, we note that the

computational formulas for the estimator of  $A_{\beta big} = -\lim_{n \rightarrow \infty} \frac{1}{n} \sum_i \frac{\partial^2 l_i(\beta_{big})}{\partial \beta_{big} \partial \beta_{big}'}$  are identical to those

given by (4) and (5), except that all matrices and vectors involved are now of larger  $(p + k)$

dimensions, incorporating auxiliary variables. We thus obtain  $\hat{A}_{\beta big}$ , either “expected” or

“observed.” We then obtain  $\hat{A}_{\beta big}^*$  by rearranging the rows and columns of  $\hat{A}_{\beta big}$  to correspond to

the rearrangement given by  $\beta_{big}^*$ . Finally, we partition as in (11) and apply the formula in (12) to

obtain  $\hat{A}_{\beta big}^{*11}$ , again either “expected” or “observed.”

The asymptotic covariance matrix of the two-stage estimator  $\tilde{\theta}$  is computed as

$$a \text{cov}(\sqrt{n}\tilde{\theta}) = (\dot{\beta}' H \dot{\beta})^{-1} \dot{\beta}' H A_{\beta big}^{*11} H \dot{\beta} (\dot{\beta}' H \dot{\beta})^{-1}, \quad (13)$$

where  $H$  and  $\dot{\beta}$  are defined as before, since the second stage has remained the same. Standard

errors can be obtained by using sample estimates in (13) and taking roots of the diagonal

elements. Similarly, the scaled test statistic is computed as in (8) but with

$c = (p^* - q) / \text{tr}\{\hat{A}_{\beta big}^{*11} (\hat{H} - \hat{H} \dot{\beta} (\dot{\beta}' \hat{H} \dot{\beta})^{-1} \dot{\beta}' \hat{H})\}$ . The mean of this statistic is that of a chi-square

variate with  $df = p^* - q$ . To obtain the residual-based statistic, we first define

$$A_{\beta \cdot \gamma} = A_{\beta, aux} - A_{\gamma\beta}' A_a^{-1} A_{\gamma\beta},$$

which is just the inverse of (12). The residual based statistic is then computed as:

$$T_{2st, NTres} = (\hat{\beta} - \beta(\tilde{\theta}))'(A_{\beta \cdot a} - A_{\beta \cdot a} \dot{\beta} (\dot{\beta}' A_{\beta \cdot a} \dot{\beta})^{-1} \dot{\beta}' A_{\beta \cdot a}) (\hat{\beta} - \beta(\tilde{\theta})) \quad (14)$$

This statistic is asymptotically chi-square distributed as long as the data are normal and at least MAR.

### Simulation Study

We investigated the performance of the modified two-stage method in a small simulation study. We included both the case of no auxiliary variables and the case when such variables are present. Of particular interest was the comparison to the ad-hoc two-stage approach which would just compute standard errors and test statistics as if the saturated ML covariance matrix were the regular sample covariance matrix obtained from complete data. That is, ad-hoc standard errors are obtained from the roots of the diagonal of

$$(\dot{\beta}' \hat{H} \dot{\beta})^{-1} / n, \quad (15)$$

and the ad-hoc test statistic is computed as:

$$T_{2st} = nF_{ML}(\tilde{\theta}). \quad (16)$$

In contrast, the appropriate standard errors are estimates of (13) and its special case (6). The appropriate test statistics are either the scaled or the residual-based test statistics.

To provide continuity with existing literature, we chose to replicate several of the conditions employed by Enders and Peugh (2004). These authors studied the ad-hoc standard errors and test statistic given by equations (15) and (16) both with and without auxiliary variables (their Studies 1 and 2, respectively). They employed sample sizes of 200, 400, and 600, and set proportion of missing of data to be 5%, 15%, or 25%. They also studied various ad-hoc



adjustments to (15) and (16), but we do not focus on those here. In study 1, data were generated from a 2-factor CFA model with 9 indicators per factor, yielding a total of 18 observed variables in the main model of interest. Loadings were set to .6, .65, and .70 (three of each) for each factor. The correlation between the factors was set to .4, and the observed variables' variances were set to 1. In Study 2, 6 auxiliary variables were appended to the model, set to correlate with the latent factors at .1 and .3, yielding a total of 24 observed variables. Enders and Peugh found that the 95% confidence interval based on (15) was always too narrow, but the coverage rarely fell below 90%, except when the proportion of missing data was highest. The standard errors in (15) are therefore downward biased, but the bias may often be negligible. In contrast, Enders and Peugh found that the rejection rates of the statistic in (16) were very inflated, across all proportions of missing data and all sample sizes. Notably, the statistic was seriously inflated even when 5% of the data were missing, casting doubt on the popular rule of thumb that 5% missing data can safely be ignored.

We replicated selected conditions from Studies 1 and 2 as follows. We set the proportion of missing data to 15%, and used sample sizes of 200, 400, and 600. For the condition with no auxiliary variables, we used the same model as in Study 1 of Enders and Peugh. For the condition with auxiliary variables, we modified their Study 2 slightly, by setting the correlations between auxiliary variables and observed variables (as opposed to latent factors) to .10 and .30 instead. The reason is that setting correlations with latent factors to those values would set the correlations between observed variables lower yet, essentially making auxiliary variables uncorrelated with the main variables, which may make it hard to detect any effects. The chosen correlations of .1 and .3 are still very low, but on the other hand values in this range are probably realistic for most applications.

Our procedure was as follows. Normally distributed data on 24 variables were generated in EQS6.1, where the first 18 variables conformed to the 2-factor structure described above, and the last 6 had the prespecified correlations. Missing data were generated on the first 18 variables using random deletion of 15% of the data on each variable. One thousand replications were created at each sample size. A *saturated* model was then fit to the first 18 variables (no auxiliary variables condition), or to all 24 variables (auxiliary variables condition), and the saturated estimate of the ML covariance matrix was saved for each dataset. These data files were then read into Matlab, where all the remaining computations were done. That is, the 18 x 18 (submatrix of) the saturated ML covariance matrix was used as  $S$  in the complete data ML fitting function of (2), which was then minimized in Matlab to obtain the two-stage estimates. We then computed 95% coverage rates based on “robust” standard errors for the selected parameters and the rejection rates at  $\alpha = .05$  of the scaled chi-square statistic and the residual-based test statistic. The “expected” information matrix was used in all computations because data were known to be MCAR.

Additionally, we computed two comparison statistics: the original ad-hoc chi-square value of (16) and the direct ML chi-square based on direct ML estimates. The ad-hoc chi-square was computed to make sure our results were accurately replicating those of Enders and Peugh for the same conditions. The direct ML chi-square was computed to establish a baseline for optimal performance at a given sample size, as it is the best known test statistic for incomplete data. To incorporate auxiliary variables, the “saturated correlates” model was estimated in EQS. This model is implemented according to the following rules (Graham, 2003): 1) auxiliary variables must be correlated with the exogenous manifest variables, 2) auxiliary variables must be correlated with residuals of all predicted manifest variables (regardless of whether these residuals

are latent or observed), and 3) auxiliary variables must be correlated with one another. Because we did not consider approximate fit indices, modification of the independence model was not necessary.

Enders and Peugh's results for the selected conditions of their Study 1 are reprinted in Table 1. They provide the rejection rates of the direct ML statistic for comparison, the ad-hoc statistic of (16), and 95% coverage rates for selected factor loadings and the factor correlations, but only for sample sizes of 200 and 600. They do not provide results for Study 2 (except for coverage rates at  $N=600$ ) but state that the rejection rates were similar. Since our auxiliary variables model was slightly different, we do not reprint the provided coverage rates.

Our results for the same conditions are given in Table 2. First, we note that our results for the direct ML and the ad-hoc chi-squares replicate Enders and Peugh's results reprinted in Table 1 very closely. The direct ML test statistic rejects about 15% of models at the smallest sample size of  $N=200$ , and this overrejection replicates the best currently known performance. At  $N=400$  and 600, the direct ML statistic does reasonably well. The ad-hoc chi-square performs terribly across all sample sizes, and should not be used. The scaled chi-square represents a considerable improvement over the ad-hoc chi-square, but does worse than the direct ML chi-square, especially in smaller sample sizes. The residual based statistic, however, exhibits surprisingly good performance across all conditions. We expected this statistic to do reasonably well given its asymptotic chi-squaredness, and to thus be a good choice when the two-stage estimator is used instead. However, we did not expect that it would improve on the performance of the direct ML chi-square. In smaller samples, this improvement is considerable: for example, at  $N=200$  and with no auxiliary variables, the rejection rates for the direct ML and the residual based chi-squares are 15.6% and 10.6%, respectively. This suggests that the two-stage estimator paired

with this statistic may be preferred in smaller samples to the highly popular direct ML estimator and test statistic. We also note that the performance of these statistics is generally worse for the auxiliary variables conditions, because adding variables necessarily increases model size, and thus requires larger samples to achieve asymptotic distributions. Whether introduction of auxiliary variables is thus warranted or not will depend on the anticipated gain in efficiency and on whether researcher believes these variables can alleviate bias. Regardless, one probably should not fit a 24-variable model to a sample of 200. Finally, we note that coverages based on correct standard errors are reasonable throughout (see Table 2) but because the coverages based on the naïve standard errors (see Table 1) were not too far off, this improvement is less noticeable. However, the standard error results are likely a function of the strength of the correlations among the variables, and could have been more dramatic in other situation. We do not recommend relying on naïve standard errors for this reason.

### **Discussion**

In this paper we have shown how to properly conduct inference based on two-stage parameter estimates, that is, those estimates obtained by using the “EM covariance matrix” and the “EM means” in the complete data ML function. This problem is well known, but appropriate solutions are not obvious. As Graham and Schafer (1999) state, “It is not clear what sample size should be ascribed to a covariance matrix estimated by EM” (p. 3). However, because of the intuitive appeal of this method and the ease with which it incorporates auxiliary variables, such solutions continue to be sought (e.g., Enders, 2004; Enders & Peugh, 2004) without much success. Enders and Peugh concluded that, “Unfortunately, there is no single value of N that is appropriate when using an EM covariance matrix as input into an SEM analysis” (p. 18). We believe the theoretical approach to the problem provided here is far more appropriate, as it uses

the actual number of cases ( $n$ ) in the computation of standard errors and test statistics. That is, we have obtained accurate estimates of variability of the two-stage estimates. Our simulation has shown our results to be valid. Applications to other statistics, such as two-stage reliability estimates (Enders, 2004), can also be developed from our results.

We would like to point out the versatility of the proposed modified two-stage approach. With complete data, many fitting functions are popular and available; for example, ULS function, GLS (normal theory generalized least squares) function, and RLS (iteratively reweighted least squares). See Bollen (1989) and Bentler (2007). These methods, however, have become unavailable with missing data because of the dominance of the direct ML approach, which is likelihood based and does not easily allow to minimize a different kind of fit function<sup>3</sup>. The two-stage theory presented here, however, easily allows for other types of fit functions in Stage 2. Estimates remain consistent for as long as the saturated estimator in Stage 1 is consistent. The computations that follow would only require modification of the  $H$  matrix; e.g., it would be set to identity in case of ULS estimation, and it would be obtained by evaluating (7) at the saturated “EM covariance matrix” in case of GLS estimation. It is impossible to predict how these methods will compare with direct ML, but our results show that they may improve on the performance of the direct ML test statistic in small samples. Additionally, ULS estimation is often useful when convergence cannot be achieved using ML methods, and our approach gives the correct standard errors and test statistic.

In addition to its surprising superior performance relative to the direct ML method when it comes to rejection rates, the two-stage approach may have other advantages over the “saturated correlates” model. An awkward feature of the “saturated correlates” model is that it does not

---

<sup>3</sup> A modification of the direct ML approach to allow for this is possible, but this line of research will be developed elsewhere.

always represent a logical construction. For example, consider a small model with six observed variables, two underlying factors that correlate with each other at .4 and have three indicators each, with loadings of .7. Suppose there is an additional observed variable, which is auxiliary and correlates with all six variables at .3. This seems reasonable enough; however, this simple model presents problems for the interpretation of the saturated correlates solution. The population covariance matrix for all seven variables is

$$\Sigma = \begin{pmatrix} 1 & & & & & & \\ .49 & 1 & & & & & \\ .49 & .49 & 1 & & & & \\ .196 & .196 & .196 & 1 & & & \\ .196 & .196 & .196 & .49 & 1 & & \\ .196 & .196 & .196 & .49 & .49 & 1 & \\ .3 & .3 & .3 & .3 & .3 & .3 & 1 \end{pmatrix}, \quad (17)$$

which is positive definite. However, the “saturated correlates” approach requires decomposing it according to the following model structure:

$$\Sigma = \begin{pmatrix} \Lambda\Phi\Lambda' & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \Psi & a \\ a' & 1 \end{pmatrix}, \quad (18)$$

where  $a' = (.3 \ .3 \ .3 \ .3 \ .3 \ .3)$ , and  $\Psi$  is a diagonal matrix with diagonal elements equal to .51. The model for the first six variables is valid. The model for all seven variables, however, is no longer valid because the new  $7 \times 7$  error variance matrix is not positive definite, i.e., it has a negative eigenvalue. In this case, the problem is primarily interpretational, and simply makes the “saturated correlates” model less elegant, unless the software of choice insures positive-definiteness of the estimated matrices. The problem becomes more real if the goal is to generate data with this covariance structure. Programs can generate simulated data either directly from the covariance matrix in (17) or from the underlying structure in (18). Often the second approach is

easier, but in this case the program will quit with an error message, because it cannot generate underlying latent variables with a negative definite covariance matrix.

Another problem may arise from the fact that the rules used to set up the “saturated correlates” model (see previous section) require the auxiliary variables to be correlated with residual variables in the model. This may become problematic and lead to unstable inference when the residual variances are very small, and it may simply be impossible if the residual variances tend to go negative and are held at zero. This happens quite frequently in real life data (Anderson & Gerbing, 1984; Chen et. al., 2001; van Driel, 1978). Consider the same model, but suppose that one of the indicators perfectly predicts the factor<sup>4</sup>. In this case, the population covariance matrix is still positive definite, and the correlation of the auxiliary variable with the perfect indicator can help if there is missing data on that indicator. However, the covariance matrix of residual error variances now has zero on the main diagonal. The fact that the auxiliary variable is only allowed to influence the perfect indicator via the residual variance means that the auxiliary variable will not be able to exert much influence in the “saturated correlates” setup. Different programs will do different things here, possibly leading to numerical problems or boundary solutions. Finally, the sheer size of the “saturated correlates” model, which always works with all  $p + k$  variables, may in some cases make it prohibitive.

Of course, the “saturated correlates” approach also has advantages over the proposed two-stage approach. The most obvious one is that the direct ML estimates are asymptotically more efficient than the two-stage estimates. It is not clear when this efficiency gain is most pronounced, but we hypothesize that because the complete data information matrix, the  $H$  matrix in (7), is block-diagonal, the two-stage approach will lose most efficiency when the true

---

<sup>4</sup> We are considering a zero variance in the population, which will lead to zero or negative sample variances in about half the samples from this population. But the same problem will arise when the population value is a small positive value, but the sample value happened to be zero/negative.

incomplete data information matrix, the one in (3), is *not* block-diagonal, which will happen with MAR data. However, in a simulation study of relative efficiency of the two approaches, Graham (2003) found that the maximum difference in efficiency between the direct ML and the two-stage methods was only 0.0008, which did happen in the condition of MAR data and low factor loadings. Thus, limited available evidence suggests efficiency may not be a problem for the two-stage method, but more research is clearly needed.

A more comprehensive study that will explore the relative efficiency of the two-stage and direct ML parameter estimates, the impact of using “observed” rather than “expected” information matrix with incomplete data, and the performance of the new test statistics under different complete data estimation methods and conditions is a direction of future research. To facilitate this research, it would be desirable to implement the auxiliary variables methods in mainstream computer programs. Given the complexity of the proposed computations, we could only carry out a limited simulation study using our own Matlab code. As Collins et. al. (2001) noted, ML estimation and testing methods as available in existing software tend to favor the “restrictive” strategy of omitting auxiliary variables, but expressed hope that “it would be possible to revise these programs to make it easy for users to add auxiliary variables” (p. 349). So far, this has not been attempted. We urge software developers to implement both the “saturated correlates” model and the proposed two-stage methodology so that “inclusive” strategies can more easily be studied.

Finally, all of the results in this paper assume normally distributed data. For a new method, this is a good start. However, since data are often nonnormal as well as incomplete, an extension of this method would be necessary. Yuan and Bentler (2000) gave the developments for the two-stage approach that allow for nonnormal MCAR data. Extending these to MAR data



and to auxiliary variables is straight-forward once the necessary consistency results become available (Yuan, 2006). In the case of nonnormal data, we predict that the residual based statistic that did best in this study will no longer work well, as it becomes similar to the ADF test statistic. The scaled statistic may do well, and other statistics, such as F statistics (Yuan & Bentler, 1998) are possible. With both normal and nonnormal data, the mean-and-variance adjusted chi-square (Satorra & Bentler, 1994) can also be considered, which may improve on the inflated performance of the mean-adjusted or “scaled” chi-square considered in this study.

Finally, the question of whether auxiliary variables are always useful still does not have a clear answer and requires more study. Based on one simulation study, Collins et. al. (2001) concluded that “the inclusion of these variables is at worst neutral, and at best extremely beneficial” (p. 348). However this conclusion was based primarily on bias and root mean square error measures, that is, on parameter estimates and standard errors, and not test statistics. Our small simulation shows that the test statistics did somewhat worse when auxiliary variables were included. We suspect that this is mainly because a bigger set of variables is considered. We thus conclude that whatever rule of thumb researchers use to determine whether their sample is large enough to conduct SEM analyses, these rules should take into account the auxiliary variables and not just the main variables being modeled.

## References

- Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, *49*, 155–173.
- 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.
- Arbuckle, J. (2007). *Amos 7.0 User's Guide*. Spring House, PA: Amos Development Corporation.
- Bentler, P. M. (2007). *EQS6 Structural Equations Program Manual*. Encino, CA: Multivariate Software, Inc.
- Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.
- Brown, C. H. (1983). Asymptotic comparison of missing data procedures for estimating factor loadings. *Psychometrika*, *48*, 269-291.
- Browne, M.W. (1984). Asymptotically distribution free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62-83.
- Chen, F., Bollen, K. A., Paxton, P., Curran, P., & Kirby, J. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods and Research*, *25*, 223–251.
- 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.

- Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*, 16-29.
- Dolan, C.V., & Molenaar, P. C. M. (1991). A comparison of four methods of calculating standard errors of maximum-likelihood estimators in the analysis of covariance structure. *British Journal of Mathematical and Statistical Psychology, 44*, 359-368.
- Ferguson, T. S. (1996). *A course in large sample theory*. London: Chapman & Hall/CRC.
- Enders, C. K. (2004). The impact of missing data on sample reliability estimates: implications for reliability reporting practices. *Educational and Psychological Measurement, 64*, 419-436.
- Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood for missing data in structural equation models. *Structural Equation Modeling, 8*, 430-457.
- Enders, C. K., & Peugh, J.L. (2004). Using an EM covariance matrix to estimate structural equation models with missing data: choosing an adjusted sample size to improve the accuracy of inferences. *Structural Equation Modeling, 11*, 1-19.
- Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structured equation models. *Structural Equation Modeling, 10*, 80-100.
- Graham, J. W., & Hofer, S. M. (1995). *EMCOV reference manual [Computer software]*. Los Angeles: University of Southern California, Institute for Prevention Research.
- Graham, J. W., Hofer, S. M., Donaldson, S. I., MacKinnon, D. P., & Schafer, J. L. (1997). Analysis with missing data in prevention research. In K. Bryant, M. Windle, & S. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 325-366). Washington, DC: American Psychological Association.

- Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate normal data with small sample size. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 1-29). Thousand Oaks, CA: Sage.
- Hu, L., Bentler, P.M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*, 351-362.
- Jamshidian, M., & Bentler, P. M. (1999). ML estimation of mean and covariance structures with missing data using complete data routines. *Journal of Educational and Behavioral Statistics*, *24*, 21–41.
- Jöreskog, K. G., & Sorbom, D. (1997). *LISREL 8: User's reference guide*. Lincolnwood, IL: Scientific Software International.
- Kenward, M. G., & Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science*, *13*, 236-247.
- Kim, K.H., & Bentler, P.M. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika*, *67*, 609-624.
- Little, R. J. A., & Rubin D. B. (2002). *Statistical analysis with missing data*. New York: Wiley.
- Magnus, J.R., & Neudecker, H. (1999). *Matrix differential calculus with applications in statistics and econometrics* (revised edition). Chichester UK: Wiley.
- Muthen, L., & Muthen, B. (2006). *Mplus user's guide*. Los Angeles.
- Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling*. Mahwah, New Jersey: Lawrence Erlbaum.
- Rovine, M. J. (1994). Latent variables models and missing data analysis. In A. von Eye, & C. C. Clogg (Eds.), *Latent Variables Analysis: Applications for Developmental Research* (pp. 181-225). Thousand Oaks, CA: Sage.

- Satorra, A., & Bentler, P.M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C.C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.
- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall.
- Schafer, J. L. (1999). *NORM: Multiple Imputation of incomplete Multivariate Data Under a Normal Model* [Computer software].
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147-177.
- Schafer, J. L., & Olsen, M.K. (1998). Multiple imputation for multivariate missing data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*, 545-571.
- van Driel, O. P. (1978). On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika, 43*, 225-243.
- Yuan, K.-H. (2006). *Normal theory ML for missing data with violation of distribution assumptions*. Unpublished manuscript. University of Notre Dame.
- Yuan, K. -H. Bentler, P.M. (1998). Normal theory based test statistics in structural equation modeling. *British Journal of Mathematical and Statistical Psychology, 51*, 289-309.
- Yuan, K.-H., & Bentler, P.M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 165-200*.
- Yuan, K.-H., & Hayashi, K. (2006). Standard errors in covariance structure models: asymptotics versus bootstrap. *British Journal of Mathematical and Statistical Psychology, 59*, 397-417.

Table 1. Results of Enders and Peugh (2004) Study 1 (no auxiliary variables) for the selected conditions, proportion of missing data 15%.

	<b>Direct ML*</b>	<b>Ad-hoc*</b>	<b>SE L1</b>	<b>SE L4</b>	<b>SE L7</b>	<b>SE L10</b>	<b>SEL13</b>	<b>SE16</b>	<b>Fcorr</b>
<b>No aux</b>									
N=200	15.0%	97.0%	92.4%**	92.2%	93.3%	93.5%	92.6%	94.4%	93.4%
N=400	8.0%	94.0%							<i>not reported</i>
N=600	7.0%	92.0%	92%***	94.1%	93.6%	93.9%	93.6%	95.2%	95.5%

Note: The first two columns give the rejection rates of the direct ML and the ad-hoc two-stage test statistic. The remaining columns give coverage of the 95% confidence interval for selected loadings and for factor correlation, based on ad-hoc standard errors (equation (15) in the present article). E&P did not report coverage for the intermediate sample size of 400.

\*First and last columns of Table 2 of E&P; \*\*Table 3, row 14, of E&P; \*\*\* Table 4, row 14, of E&P.

Table 2. Replication of Enders and Peugh (2004) for the selected conditions with new standard errors and test statistics, proportion of missing data 15%.

<b>Direct</b>											
	<b>ML</b>	<b>Ad-hoc</b>	<b>Scld</b>	<b>Res</b>	<b>SE L1</b>	<b>SE L4</b>	<b>SE L7</b>	<b>SE L10</b>	<b>SEL13</b>	<b>SE16</b>	<b>Fcorr</b>
<b>No aux</b>											
N=200	15.6%	97.1%	25.7%	10.6%	95.1%	93.4%	93.3%	93.6%	95.4%	93.1%	93.3%
N=400	8.7%	94.3%	11.9%	8.1%	95.3%	94.4%	93.2%	94.9%	94.3%	94.0%	94.5%
N=600	8.7%	92.6%	11.7%	7.7%	96.0%	95.3%	93.9%	96.7%	94.8%	94.6%	94.8%
<b>Aux</b>											
N=200	19.5%	97.9%	35.8%	17.6%	96.0%	94.8%	95.0%	96.2%	95.4%	93.4%	93.7%
N=400	10.2%	94.6%	14.7%	8.9%	96.2%	95.7%	95.8%	96.5%	95.7%	94.5%	93.2%
N=600	8.1%	93.2%	10.9%	7.6%	96.0%	95.3%	95.1%	94.5%	96.0%	95.0%	93.4%