The Supermatrix Technique: A Simple Framework for Hypothesis Testing with Missing Data

Kyle M. Lang

University of Kansas Department of Psychology

Todd D. Little

Texas Tech University Institute for Measurement, Methodology, Analysis, and Policy

Author Note

This research was funded in part by a grant (NSF0066969) from the National Science Foundation (T. Little & W. Wu, co-PIs), and in part by the Center for Research Methods and Data Analysis (when Todd D. Little was director; 2009-2013). Todd D. Little is now director of the Institute for Measurement, Methodology, Analysis, and Policy at Texas Tech University.

Correspondence for this manuscript can be addressed to Kyle M. Lang, Department of Psychology, University of Kansas, Lawrence, KS 66045. Email: kylelang@ku.edu

Abstract

We present a new paradigm that allows simplified testing of multiparameter hypotheses in the presence of incomplete data. The proposed technique is a straight-forward procedure that combines the benefits of two powerful data analytic tools: multiple imputation and nested-model χ^2 difference testing. A Monte Carlo simulation study was conducted to assess the performance of the proposed technique. Full information maximum likelihood (FIML) and single regression imputation were included as comparison conditions against which the performance of the suggested technique was judged. The imputation-based conditions demonstrated much higher convergence rates than the FIML conditions. $\Delta\chi^2$ statistics derived from the proposed technique were more accurate than such statistics derived from both the FIML conditions and the regression imputation conditions. Limitations of the current work and suggestions for future directions are also addressed.

Keywords: missing data, hypothesis testing, multiple imputation, full information maximum likelihood, monte carlo simulation

The Supermatrix Technique: A Simple Framework for Hypothesis Testing with Missing Data

Several recent papers offer convincing evidence for the utility of planned missing data designs (e.g., Garnier-Villarreal, Rhemtulla, & Little, 2013; Rhemtulla, Jia, Wu, & Little, 2013). We seek to further motivate the use of these powerful tools by proposing a simple framework for conducting multiparameter hypothesis tests in the presence of incomplete data. The issue of accurately testing hypotheses when faced with missing data is certainly not a trivial concern, and single parameter hypothesis tests are not always adequate. Consider a hypothetical two-group panel model that tests for gender differences in how attachment style mediates the relationship between aggression and positive affect among elementary school students (i.e., a test of moderated mediation). Such a question can only be fully elucidated via a model comparison approach. To streamline these model comparison tests, we introduce a new approach (the "supermatrix" technique) to build upon and expand the extant hypothesis testing framework of modern missing data analysis.

The veracity of MI-based and FIML-based parameter estimation has been well established (Enders & Bandalos, 2001; Schafer & Graham, 2002), but these tools are still limited when it comes to hypothesis testing. When using FIML estimation, both single parameter and multiparameter hypothesis tests are easily conducted. However, there are a number of situations in which MI may be preferred over FIML. MI can easily accommodate a large number of auxiliary variables, for example, but FIML is not as flexible. Because FIML can only use auxiliary variables that are included in the analysis model, it will tend to exhibit convergence problems when a large number of auxiliary variables are introduced (Enders, 2010). MI also has the advantage when calculating scale scores or parcels from incomplete data. In these circumstances, MI can simply impute the item-level data before any composite scores are

calculated. FIML, on the other hand, never "fills in" the missing data points. So, there is no way to sum or average these incomplete variables without losing the partial item-level information (Enders, 2010).

While they may have certain advantages over FIML, traditional implementations of MI come with their own set of limitations. MI-based single parameter hypothesis tests are easily conducted by using Rubin's Rules (Rubin, 1987) to construct corrected Wald statistics or confidence intervals. Rubin's Rules, however, are not applicable to aggregating the likelihood ratio statistics (χ^2) required for nested model $\Delta\chi^2$ tests, and they still rely on standard errors that can be influenced by the method of scale setting (Gonzalez & Griffin, 2001). This is a nontrivial limitation when testing hypotheses that compare two different, model-based, understandings of a psychological phenomenon. Rhemtulla, et al (2013), for example, show how planned missing data designs can be used to improve statistical inference in growth curve modeling. However, testing the shape of the population change trajectory in a growth curve modeling framework requires a model comparison approach. In the supermatrix technique, we seek to develop a tool which can address the limitations described above by offering a straight-forward method for conducting $\Delta\chi^2$ -based model comparison tests with multiply imputed data.

Previous work has developed χ^2 -distributed statistics that can be calculated in the context of multiply imputed data structures (e.g., Browne, 1984; Meng & Rubin, 1992). Unfortunately, these statistics entail complicated calculations that can make them difficult to implement, and they have yet to be widely incorporated into statistical analysis software. Others have sought simpler solutions by focusing on pooling the multiple χ^2 replicates, or pooling the multiple imputed data sets prior to the analysis phase of a study. Unfortunately, the simplest implementation of this strategy has been shown to provide suboptimal results. Naively averaging

the χ^2 replicates derived from the analysis phase of ordinary MI will lead to biased assessments of model fit (Asparouhov & Muthén, 2010; Lang, 2013).

Lee and Cai (2012) develop a principled pooling approach based on averaging the multiple covariance matrices produced by a Bayesian multiple imputation. After applying a post hoc correction adapted from Browne (1984), their approach produces an accurate χ^2-distributed model fit statistic, but they did not explicitly examined their approach as a hypothesis testing tool. So, it is unclear whether or not the post hoc correction is necessary to conduct accurate hypothesis tests. The rational of the supermatrix technique is very similar to the rational of the Lee and Cai (2012) method, but its development has focused on conducting accurate $\Delta\chi^2$-based hypothesis tests without the need to appeal to a two-stage estimator.

Wood, White, and Royston (2008) have suggested a "stacked data" approach that is particularly germane to the current work. Their approach aggregates the *m* imputed data sets by stacking them one atop the other into a $mN \times p$ data frame where *N* is the original sample size and *p* is the number of variables. When this stacked data frame was analyzed with a series of weighted linear regression models, the resulting Type I error rates and power were comparable to the Rubin's Rules pooled estimates, even with simple weighting schemes.

The current work develops a technique that is mathematically equivalent to the W_1 technique from Wood et al. (2008). The W_1 weights (i.e., $W = m^{-1}$) correct for the redundancy introduced by stacking the *m* imputed data sets. The supermatrix technique applies an equivalent correction that is tailored to covariance structure modeling. Our approach is designed to aggregate the multiply imputed data sets in a way that will allow researchers to use a principled tool to treat their missing data (i.e., multiple imputation), yet still maintain the simplicity of estimating a single analysis model.

The Proposed Technique

The first step in the supermatrix approach is creating some number of plausible imputations. Once some number ($m > 1$) of imputations have been created, the m imputed data sets are "stacked" one atop the other to create a single data frame whose number of rows is equal to the number of observations in the original data set times the number of imputations. For example, consider a data set that contains $N = 250$ observations of $p = 25$ variables. If you create $m = 20$ imputations, the final dimensions of this combined data frame will be $mN \times p = 20(250) \times 25 = 5000 \times 25$. A single covariance matrix is then calculated to summarize all mN observations of this aggregate data frame. This covariance matrix (i.e., the supermatrix) then contains the sufficient statistics to which the analysis models are fit. In the example above, the supermatrix technique would produce a 25×25 covariance matrix that that summarize the 20 imputed data sets. Figure 1 gives a graphical representation of the supermatrix process.

<Include Figure 1 about here>

Because the aggregate covariance matrix produced by the supermatrix technique is envisioned as a proxy for the complete data, the number of observations ascribed to the underlying data set is explicitly set to the original sample size N. This constraint adjusts the likelihood ratio statistics to correct for the spurious inflation of sample size entailed in treating the stacked m imputed data sets as a single data frame.

Van Buuren (2012) suggests that stacking approaches like the supermatrix technique will produce unbiased point estimates of model parameters, but may lead to negatively biased standard errors for those estimates. Likewise, the supermatrix technique is expected to produce unbiased point estimates of model parameters, but we do not have confidence that it will produce accurate standard errors (see Wood et al., 2008, Appendix A, for justification of this claim).

Fortunately, significance tests that are based on nested model $\Delta\chi^2$ statistics are not directly affected by the tested parameters' standard errors. Such tests conducted with the supermatrix technique will not be affected by bias in these standard errors. So, we expect the supermatrix technique to produce accurate tests of hypotheses, provide they are based on the $\Delta\chi^2$.

A Monte Carlo simulation study was conducted to assess the performance of the supermatrix technique at varying levels of sample size and percent missing. Of greatest interest are the convergence rates of the analysis models and the accuracy of nested model $\Delta\chi^2$ statistics. Specifically, it is hypothesized that:

1) The supermatrix technique will produce higher convergence rates than FIML estimation.
2) Supermatrix-based $\Delta\chi^2$ statistics will be negligibly different from complete data-based $\Delta\chi^2$ statistics.
3) Supermatrix-based $\Delta\chi^2$ statistics will outperform $\Delta\chi^2$ statistics derived from data treated with single regression imputation.

Methods

Data Simulation

The data for the Monte Carlo study were simulated using R 2.15 (R Development Core Team, 2011). The simulated data were generated according to the multi-trait multi-method (MTMM) factor analytic model shown in Plate 1 of Figure 2. An MTMM structure was chosen for the data generation model to more accurately represent the nuanced process by which psychological phenomena arise in reality. The data generating model employed two "common" factors and two "covariate" factors to simulated two "scales" of ten items each. Each of the common factors uniquely predicted the ten items of one of the scales, while the covariate factors were allowed predict all twenty simulated items.

All common factor loadings were specified to be equal in the population with a value of λ =.6. The covariate factors were each allowed to have their own, small association with the items of the two scales. That is, the items of the first scale loaded onto the first covariate factor at λ =.2, while the second scale's variables loaded onto the first covariate factor at λ =.1. Likewise, the first scale's variables loaded onto the second covariate factor at λ =.05, while the second scale's items did so at λ =.1. Common factors had variances fixed to $\psi_{1,1} = \psi_{2,2} = 1.0$ and covaried with one another at $\psi_{2,1} = .5$. The two covariate factors were specified to be independent of one another and the common factors (i.e., $\psi_{4,3} = \psi_{4,2} = \psi_{4,1} = \psi_{3,2} = \psi_{3,1} = 0$). They had variances of $\psi_{3,3} = 1.0$ and $\psi_{4,4} = 3.0$. Variances of all unique factors were specified to be a constant θ = .64. This model structure was chosen to simulate items arising from two strongly correlated hypothetical constructs (e.g., empathy and prosocial attitudes) that are also subject the lesser influence of two uncorrelated covariates (e.g., mindfulness and extraversion).

After populating the parameter matrices with the values described above, the R function rmvtnorm (Genz et al., 2012) was used to simulate multivariate normal factor scores and error terms. The final simulated data sets were derived by including these factor scores and error terms in the factor analytic data model represented by Equation 1.

$$\mathbf{Y} = \eta\Lambda^T + \Theta \quad (1)$$

Where $\mathbf{Y}$ is an $N \times 20$ matrix of simulated data, η is an $N \times 4$ matrix of factor scores, Λ is a 20×4 matrix of factor loadings and Θ is an $N \times 20$ matrix of residual error components. Finally, the two columns of covariate factor scores where merged with $\mathbf{Y}$. Thus, the final simulated data set was an $N \times 22$ matrix consisting of the twenty scale variables to which the analysis models would be fit (i.e., empathy and prosocial attitudes items, from the example above) and the two

auxiliary variables that would be used as predictors of the missing data process (i.e., mindfulness and extraversion, from the example above).

<Include Figure 2 about here>

Parsimony Error

To further improve the ecological validity of the current work, the data were generated according to a model that was more complex than the subsequently specified analysis models. This "hidden" simplification is analogous to the parsimony error introduced by researchers attempting to portray highly complex psychological phenomena with simplified mathematical models. To implement this misspecification, several off-diagonal elements of the residual covariance matrix described above were set to small, non-zero values (θ = .03). This alteration introduced a small residual covariance between every fifth unique factor. When these residual covariances were not included in the analysis models, the result was a mild degree of model misspecification present in all conditions.

Missing Data Imposition

Missing data were imposed according to a missing at random (MAR) process as defined by Rubin (1976). A function was written in R 2.15 that ran iteratively through all twenty of the simulated scale variables. Equation 2 quantifies the decision rule by which the missing values were imposed on each variable.

$$P(R_i = 1 \mid X) = \begin{cases} 1, \text{ if } \Phi^{-1}(\mathrm{X}_i) \leq PM + \delta \\ 0, \text{ if } \Phi^{-1}(\mathrm{X}_i) > PM + \delta \end{cases}, \; i = 1, 2, ..., N \qquad (2)$$

Where R_i is the i^{th} entry in an $N \times 1$ pattern matrix that equals 1 when the i^{th} observation of the focal scale variable is missing and 0 when it is observed, X is an exogenous missing data predictor (i.e., auxiliary variable), $\Phi^{-1}(\cdot)$ represents the inverse normal cumulative distribution

function, *PM* is the proportion of missing data, δ is a small disturbance factor that ensures no row in the treated data frame is entirely empty, and *N* is the total sample size. For half of the items, *X* was specified to be the $N \times 1$ matrix of factor scores from the first covariate factor in the data generating model. For the other half of the items, *X* was specified to be the analogous matrix of factor scores from the second covariate factor.

In terms of our running example, this function recreates a scenario in which the chance that some of a subject's empathy or prosocial attitudes items are missing is determined by their levels of mindfulness and extraversion. Because this function imposes the missing data according to a simple probit regression model, the Rubin (1976) definition of a MAR process is replicated as closely as possible. That is, the missingness can be considered a pure random sample of the complete data, after conditioning on the predictor of missingness.

Comparison Conditions

Three comparison conditions were included against which the performance of the supermatrix technique was judged. As an optimal, control condition, the analysis models were fit to the complete data. Because the supermatrix technique is being developed as a practical hypothesis testing tool, we are not principally concerned with ensuring that the estimates it produces are penalized for the missing data. Rather, the current work presupposes that an optimal missing data analysis should produce results that are equivalent to those derived from fully observed data, all validity issue being equal. Thus, we set the complete data-based $\Delta\chi^2$ values as the ideal to be reproduced by the supermatrix.

For the second comparison condition, FIML estimation was used to fit the analysis models directly to the incomplete data. FIML was chosen because it has been shown to demonstrate optimal statistical properties for a wide range of missing data problems (Enders,

2010; Enders & Bandalos, 2001). To ensure that the FIML conditions were optimally implemented, the predictors used to impose the missing data (i.e., mindfulness and extraversion, in our running example) were included in the analysis models via the saturated correlates approach (Graham, 2003) to fully satisfy the MAR assumption.

Finally, the missing data were treated with a single regression imputation. This technique was chosen because it could be argued that aggregating the multiple imputations with the supermatrix approach may be "washing out" the between imputation variance that is the foundation of the unique benefits of MI. If this is true, the supermatrix technique may reduce to a needlessly complex single regression imputation.

Simulation Parameters

There were two simulation parameters varied in this study: sample size (*N*) and percent missing (*PM*). Because considering only a small discrete set of simulation parameters can lead to a considerable loss of interesting detail, the current study varied the simulation parameters in small steps (i.e., $N = \{100, 120, \ldots, 980, 1000\}$, $PM = \{2, 4, \ldots, 48, 50\}$). Thus, for every replication there were 1150 crossed levels of *N* and *PM*. Such fine-grained specification of percentages missing may seem unnecessary since missingness imposed intentionally via planned missing data designs will occur as relatively large fixed percentages (e.g., *PM*=25% for a classic three-form design). However, the inevitable addition of unplanned missingness to this planned missingness means that, even in a planned missing data design, the observed rates of nonresponse can vary arbitrarily.

Within each of these 1150 cells, eight analysis models were estimated using the R package lavaan (Rosseel, 2012). "Full" and "restricted" confirmatory factor analysis (CFA) models were fit to the complete data with ordinary maximum likelihood (ML) estimation, to the

incomplete data using FIML estimation, and to the imputed data produced by the supermatrix technique or regression imputation. The full model was a two factor CFA in which each factor predicted one of the two simulated scales, and the factors were allowed to freely covary.

The restricted model was identical to the full model except that the latent covariance was fixed to $\psi_{2,1} = 0$. This constraint offered the means to assess Hypotheses 2 and 3 by facilitating significance tests of the latent covariance via nested model $\Delta\chi^2$ tests. The analysis models associated with the FIML conditions were identical to those described above except that they also incorporated the predictors of the missing data process via the saturated correlates technique (see Plate 2 of Figure 2).

In terms of our running example, the full model was a CFA with correlated empathy and prosocial attitudes factors, while the restricted model forced empathy and prosocial attitudes to be independent. Thus, the $\Delta\chi^2$ tests discussed below would be testing the significance of the correlation between empathy and prosocial attitudes.

Test Statistics

Two test statistics were employed in this study: percentage relative bias (PRB) and root mean square error (RMSE). PRB was chosen because its interpretation makes it well suited to the *a priori* explication of thresholds by which to judge performance. PRB is simply the average bias in the estimated statistic rescaled as a percentage of the true statistic's magnitude. The formula for PRB is quite simple:

$$PRB = 100 \cdot \left(K^{-1} \sum_{i=1}^{K} \frac{\hat{T}_i - T}{T} \right) \quad (3)$$

Where T is the true value of the statistic, $\hat{T}_i$ is the estimated statistic for the i^{th} replication, and K is the number of replications. It should be noted that the fit statistics derived from the complete

data conditions were treated as the ideal, so the complete data-based values were considered the "true" values. The "bias" discussed throughout this paper, therefore, is not bias in the usual sense because the observed statistics are not compared back to any true population values.

The RMSE was chosen because it combines information on both bias and variability into a well-rounded measure of overall accuracy (Burton, Altman, Royston, & Holder, 2006). Because the formulation of these two statistics is somewhat, but not entirely, redundant (i.e., both PRB and RMSE contain a term quantifying bias but PRB ignores variability) the two in combination offer a picture of performance that is more nuanced than the sum of its parts. The formula for RMSE is also quite simple:

$$RMSE = \sqrt{K^{-1}\sum_{i=1}^{K}\left(\hat{T}_i - T\right)^2} = \sqrt{\left(\bar{\hat{T}} - T\right)^2 + \left(SE_{\hat{T}}\right)^2} \quad (4)$$

Where T is the true value of the statistic, $\hat{T}_i$ is the estimated statistic from the i^{th} replication, $\bar{\hat{T}}$ is the mean of the estimated statistic, $SE_{\hat{T}}$ is the empirical standard deviation of the estimated statistic, and K is number of replications.

In this study, PRB > 5 was considered to be an excessive degree of bias. In other words, if the estimated value of a missing data-based $\Delta\chi^2$ statistic deviated from the complete data-based value by more than 5% of complete data value's magnitude, that missing data technique was considered to perform unacceptably in that condition.

Procedure

For every replication a single data set was simulated according to the process described above. This data set was then used to fit the complete data control models. Next, missing data were imposed, and the FIML conditions were fit to this incomplete data. In the supermatrix conditions, these same incomplete data were imputed 100 times using the R package Amelia II

(Honaker, King, & Blackwell, 2011). These 100 imputed data sets were then submitted to the supermatrix treatment and analyzed via the models described above. Finally, the missing data were imputed once via regression imputation using the *mice.impute.norm.predict* function from the R package MICE (van Buuren & Groothuis-Oudshoorn, 2011). These data were then analyzed with the same models fit to the supermatrix covariance matrices.

Once all eight models were estimated, their χ^2 statistics were used to calculate the $\Delta\chi^2$ between the respective full and restricted models. Finally, the PRB and RMSE of these $\Delta\chi^2$ values were calculated for the supermatrix, regression imputation, and FIML conditions (i.e., using the complete data $\Delta\chi^2$ statistics as the true values). This process was repeated 500 times for every one of the 1150 crossed levels of percent missing and sample size. This resulted in a total of 3(missing data treatments) × 2(model constraints) × 46(*N*) × 25(*PM*) = 6900 total crossed conditions within each of the 500 replications.

Results

Convergence Rates

Hypothesis 1 was fully supported. All of the imputation-based models demonstrated perfect convergence. However, some FIML models had very low convergence rates, particularly in conditions with small *N* and high *PM*. Figure 3 shows the convergence rates for FIML-based models plotted by *N* and *PM*, while Table 1 shows the convergence rates for the FIML conditions with less than 80% convergence.

An obvious pattern emerges when considering these results. Convergence rates for FIML-based conditions were perfect for larger *N* and lower *PM*, but there is a distinct point where convergence rates begin to rapidly decrease as *N* decreases and *PM* increases. The

precipitous drop-off in Figure 3 and corresponding entries in Table 1 show that sample sizes lower than 200 tend to produce very low convergence rates, particularly when *PM* exceeds 30%.

<Include Figure 3 about here>

<Include Table 1 about here>

Significance Testing with the $\Delta\chi^2$

The most important finding of this study is the support of Hypothesis 2. Tables 2 and 3 show the $\Delta\chi^2$ values from a representative subset of sample sizes and percents missing. Table 2 only reports the results from replications in which the FIML models converged, while Table 3 reports all of the available results. It is apparent that the supermatrix-based $\Delta\chi^2$ values were quite accurate across all conditions tested (although they did tend to demonstrate a small degree of positive bias when $PM > 40$ and $N < 300$). However, the FIML-based $\Delta\chi^2$ tests performed poorly across a large proportion of the tested conditions. In stark contrast to the supermatrix-based $\Delta\chi^2$ values, the FIML-based $\Delta\chi^2$ tests demonstrate a consistent and appreciable negative bias. In fact, for all the tested levels of *N*, the FIML-based $\Delta\chi^2$ values reached an unacceptable degree of negative bias once *PM* exceeds 10%. Plate 1 of Figure 4 shows that, on average, the supermatrix-based $\Delta\chi^2$ values track the complete data versions almost perfectly. Plate 2 of Figure 4, on the other hand, shows the considerable discrepancy between the average FIML-based $\Delta\chi^2$ values and the complete data versions.

<Include Figure 4 about here>

The RMSE values associated with the FIML-based $\Delta\chi^2$ statistics also indicate poor performance. Two problematic patterns are apparent when referring to Table 2. First, in the same way that the FIML-based PRB values are larger in absolute magnitude than their supermatrix-

based counterparts, the FIML-based RMSE values are also universally larger than the analogous supermatrix-based values. Second, the PRB of both methods demonstrates a monotonic inflation as *PM* increases, independent of *N*, but the RMSE values associated with the FIML-based $\Delta\chi^2$ increase with both rising *PM* and increasing *N*. Since the RMSE accounts for not only bias, but also efficiency, this discrepancy suggests that the variation of the FIML-based $\Delta\chi^2$ values increases as *N* increases. When looking at the supermatrix-based $\Delta\chi^2$ values, we see that changing *N* does not have such an influence.

<Include Table 2 about here>

Hypothesis 3 was also supported. Although, Table 3 shows that regression imputation demonstrated negligible bias across all tested conditions, the RMSE values tell a different story. Across most conditions, the regression imputation-based RMSE values were nearly twice as large as the supermatrix-based values. This inversion in the patterns of PRB and RMSE implies much higher variability across replications (and thus lower anticipated efficiency) for regression imputation than the supermatrix technique.

<Include Table 3 about here>

Discussion

This study demonstrates the utility of the supermatrix technique as a parsimonious framework for hypothesis testing with missing data. The supermatrix approach showed high rates of convergence and a consistent ability to produce $\Delta\chi^2$ statistics that reflect the complete data standard, both in terms of bias and efficiency. By way of comparison, the performance of the supermatrix technique was contrasted with two other common missing data tools: FIML estimation and single regression imputation. The former was chosen because it is the current

gold standard in missing data analysis, and the latter was chosen because it may fill a similar role to the supermatrix technique.

When compared to FIML estimation, the performance of the supermatrix technique was quite good. The supermatrix technique exhibited much higher rates of convergence. In fact, all of the supermatrix conditions achieved 100% convergence, but the FIML conditions had considerable issues with convergence. Once sample sizes dropped below 200 and rates of missing increased much beyond 30%, the FIML models began to exhibit very low convergence rates. This finding is not terribly surprising given that FIML estimates are based on only the observed responses. This means that mixing such high percents missing with small sample sizes essentially decreases the sample size below what is required to satisfy the large sample assumption of ML. Because imputation techniques (including the supermatrix technique) operate by "filling in" the holes in the incomplete data set, they circumvent this issue.

This is a nontrivial finding. Higher convergence rates in a Monte Carlo simulation directly translate to higher probabilities of model convergence in substantive studies. Thus, the poor convergence rates observed for the FIML conditions have meaningful implications. Namely, researchers can anticipate a higher probability of achieving convergence if they treat their missing data with the supermatrix technique rather than employing FIML estimation, especially when faced with small sample sizes and high rates of nonresponse.

The performance of the supermatrix technique as a tool for conducting $\Delta\chi^2$ tests was also very promising. In all conditions tested (except for a few conditions with very high rates of missing and small sample sizes), the supermatrix-based $\Delta\chi^2$ values showed negligible bias. This finding suggests that the supermatrix technique can produce unbiased assessments of parameter significance under a paradigm that allows hypotheses to be represented as competing models.

Additionally, assessing parameter significance with $\Delta\chi^2$ tests, rather than Wald statistics, allows for statistical inference without the need to appeal to standard errors which are easily biased by missing data and sensitive to model parameterization.

While unbiased supermatrix-based $\Delta\chi^2$ statistics were expected, an unanticipated result was the poor performance of the FIML-based $\Delta\chi^2$ statistics. In all but those conditions with the lowest percents missing (i.e., *PM* < 10%), the FIML-based $\Delta\chi^2$ values were unacceptably negatively biased. Although the effect size associated with this test (i.e., $\psi_{2,1}$ = *r* = .5) was too large to assess rejection rates, the substantial degree of negative bias suggests that researchers seeking to capture a small effect with FIML-based $\Delta\chi^2$ tests would likely face considerably inflated Type II error rates.

To compound this poor performance, the FIML-based $\Delta\chi^2$ values were additionally sensitive to sample size. The PRB of both supermatrix-based and FIML-based $\Delta\chi^2$ statistics demonstrated a monotonic increase as a function of the increasing rates of missingness, but neither was sensitive to changes in sample size. While the RMSE of the supermatrix-based $\Delta\chi^2$ statistics followed a similar pattern, the RMSE of the FIML-based values also increased along with increasing sample size. Because the RMSE captures efficiency as well as bias, and the bias of the FIML-based $\Delta\chi^2$ values was independent of sample size, this inflation of the RMSE implies increasing variation in the estimated χ^2 values as sample size increases. The absence of such a relationship for the supermatrix-based values suggests that the supermatrix-based $\Delta\chi^2$ statistics will remain much more stable with changing sample size. Thus, in addition to having

less overall bias, the supermatrix-based $\Delta\chi^2$ statistics can be expected to demonstrate greater stability than their FIML-based counterparts.

The supermatrix technique was also compared to single regression imputation. Regression imputation was unbiased, but the RMSE values associated with it were nearly twice as large as those derived from the supermatrix technique. This finding shifts the performance balance in favor of the supermatrix technique, because neither the supermatrix technique nor regression imputation was unacceptably biased, but regression imputation demonstrated considerably lower efficiency.

Limitations and Future Directions

The scope of this study was limited and represented only a subset of situations where researchers may consider using the supermatrix technique. All simulated data were multivariate normally distributed, and only simple, cross-sectional, single group models were considered. We acknowledge that the reality experienced by applied researchers is likely to be more complex. We have no reason, however, to expect the results would not hold over a wide array of modeling situations.

Future extensions of this work should address some specific issues. Model complexity should be varied to represent more of the situations encountered by applied researchers. This should include variation in the pattern of constraints used to produce the $\Delta\chi^2$ statistics, since all of the $\Delta\chi^2$ values in this study were derived by constraining single parameters. We have no reason to suspect that these results do not generalize to tests of multiple parameters, but future work should confirm this. Also, future work should be optimized to scrutinize rejection rates, because the data generation model chosen for this study disallowed testing power and Type I error rates. Finally, the current study was not designed to assess the accuracy of the model

parameters themselves. The work of previous authors (e.g., Rubin, 1987; Satorra & Bentler, 1994; van Buuren, 2012; Wood et al., 2008) and the principles of point estimation that motivate Rubin's Rules suggest that the supermatrix technique should produce unbiased point estimates of model parameters, but this should be confirmed.

Conclusion

In summary, the supermatrix technique shows great promise. When compared to FIML estimation, the supermatrix technique demonstrated far higher convergence rates, lower bias, and higher efficiency. The supermatrix technique also outperformed regression imputation, at least in terms of efficiency. These two findings in concert suggest that, of the three techniques examined in this study, the supermatrix technique should produce the most accurate and consistent nested model comparison tests, in practice. In the end, the results of this study suggest that the supermatrix technique is a useful tool for researchers seeking to test model-based hypotheses in the presence of incomplete data.

References

Asparouhov, T., & Muthén, B. (2010, July). Chi-square statistics with multiple imputation [Computer software manual]. (Technical Appendix. Muthén & Muthén: Los Angeles)

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*. doi: 10.1111/j.2044-8317.1984.tb00789.x

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine, 25*(24), 4279–4292. doi: 10.1002/sim.2673

Enders, C. K. (2010). *Applied missing data analysis.* New York, NY: The Guilford Press.

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: A Multidisciplinary Journal, 8*(3), 430–457. doi: 10.1207/S15328007SEM08035

Garnier-Villarreal, M., Rhemtulla, M., & Little, T. D. (2013). Two-method planned missing designs for longitudinal research. *International Journal of Behavioral Development*. Manuscript submitted for review

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2012). mvtnorm: Multivariate normal and t distributions [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=mvtnorm (R package version 0.9-9992)

Gonzalez, R., & Griffin, D. (2001). Testing parameters in structural equation modeling: every "one" matters. *Psychological Methods, 6*(3), 258–269. doi: 10.1037//1082-989X.6.3.258

Graham, J. W. (2003). Adding missing-data-relevant variables to fiml-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 10*(1), 80–100. doi: 10.1207/S15328007SEM10014

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software, 45*(7), 1–47. Retrieved from http://www.jstatsoft.org/v45/i07/

Lang, K. M. (2013). Streamlining missing data analysis by aggregating multiple imputations at the data level: A Monte Carlo simulation to assess the tenability of the supermatrix approach. (Master's thesis, University of Kansas, 2013). URI: http://hdl.handle.net/1808/11718

Lee, T., & Cai, L. (2012). Alternative multiple imputation inference for mean and covariance structure modeling. *Journal of Educational and Behavioral Statistics, 37*(6), 675–702. doi: 10.3102/1076998612458320

Meng, X. L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika, 79*(1), 103–111. doi: 10.1093/biomet/79.1.103

R Development Core Team. (2011). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org (ISBN 3-900051-07-0)

Rhemtulla, M., Jia, F., Wu, W., & Little, T. D. (2013). Planned missing designs to optimize the efficiency of latent growth parameter estimates. *International Journal of Behavioral Development.* Manuscript submitted for review

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592. doi: 10.1093/biomet/63.3.581

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

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

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of state of the art. *Psychological Methods, 7*(2), 147–177. doi: 10.1037//1082-989X.7.2.147

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67.

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: Chapman & Hall/CRC.

Wood, A. M., White, I. R., & Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine, 27*(17), 3227–3246. doi: 10.1002/sim.31

Figures & Tables

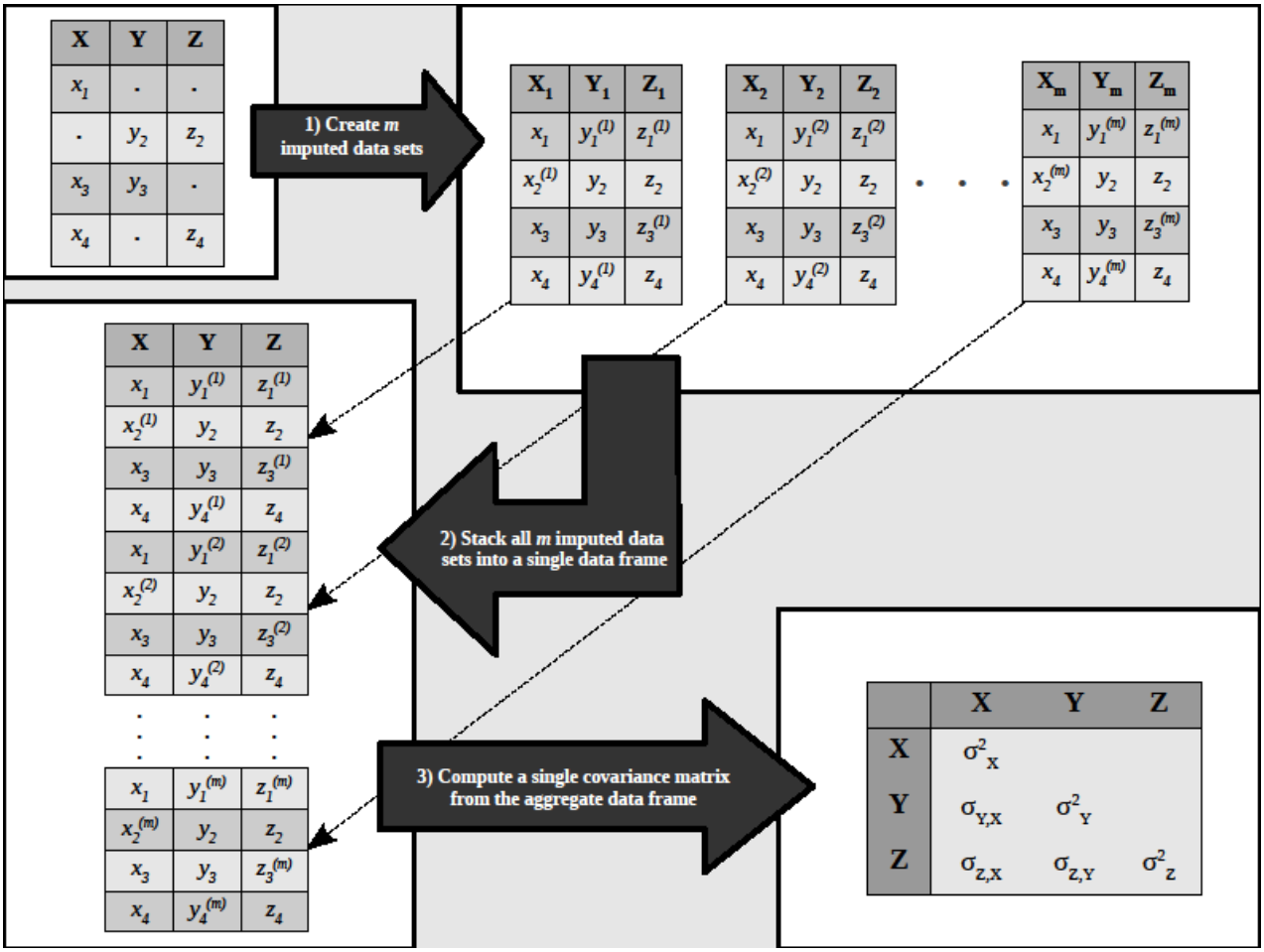

Figure 1: A graphic representation of the supermatrix technique as applied to a data set with 4 observations of 3 variables

Plate 1: Data Generating Model

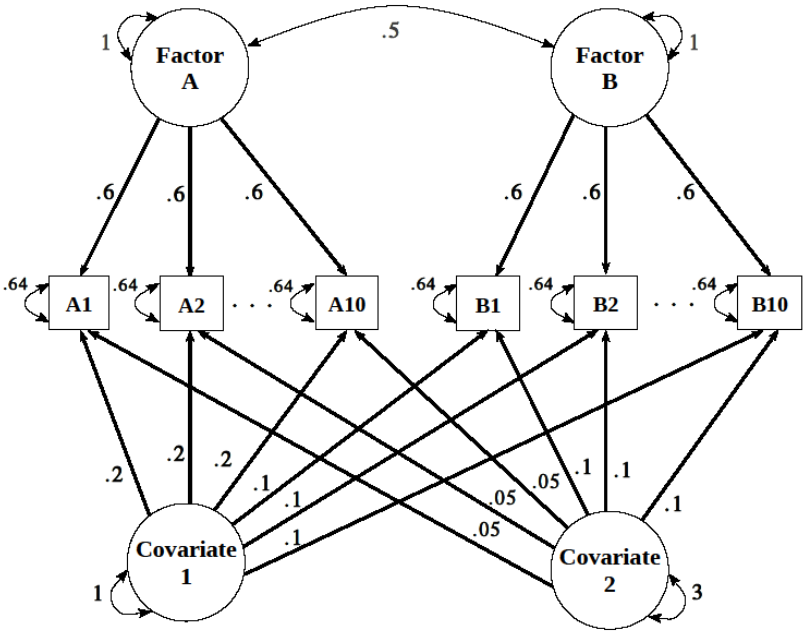

Plate 2: Analysis Model for the FIML Conditions

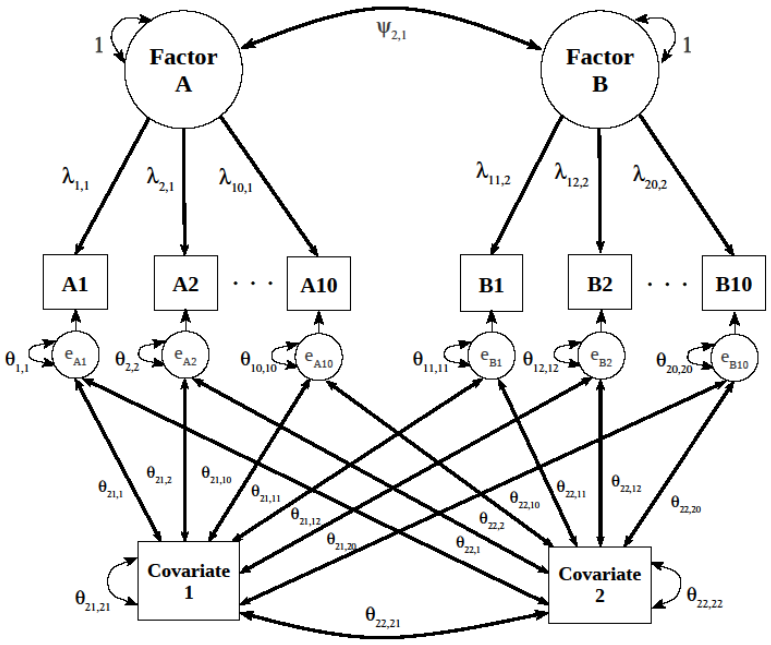

Figure 2: Path diagrams for selected models used in the simulation study

Note: Mean Structures are not shown

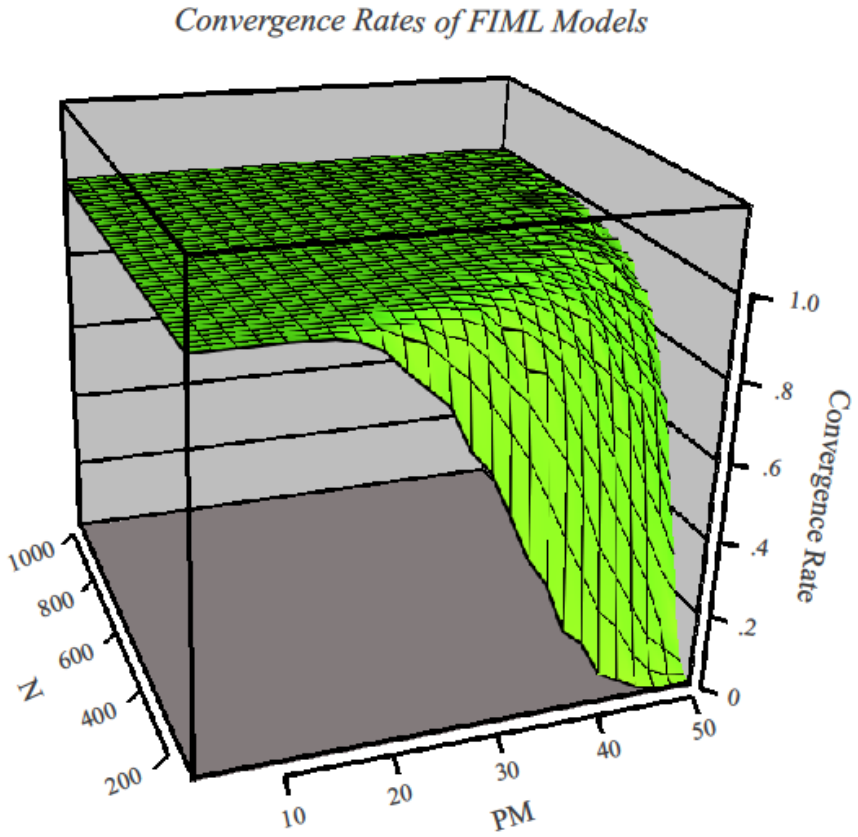

Figure 3: Convergence rates for FIML models plotted by sample size and percent missing

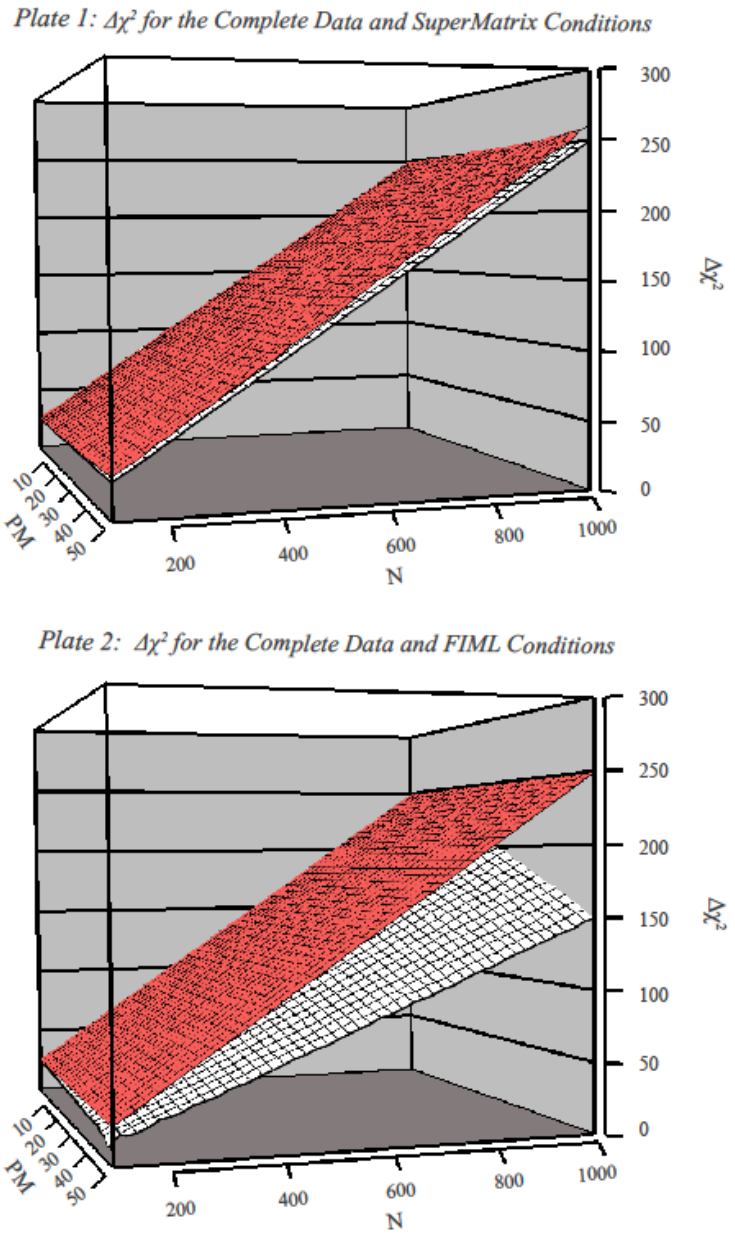

Figure 4: $\Delta\chi^2$ values for complete data, supermatrix & FIML conditions plotted by sample size and percent missing

Plate 1: Upper surface=Supermatrix conditions, Lower surface=Complete data conditions

Plate 2: Upper surface=Complete data conditions, Lower surface=FIML conditions

N	PM	Convergence Rate	N	PM	Convergence Rate	N	PM	Convergence Rate
100	***26***	0.708	***140***	***32***	0.776	***200***	***40***	0.786
100	***28***	0.65	***140***	***34***	0.698	***200***	***42***	0.66
100	***30***	0.538	***140***	***36***	0.626	***200***	***44***	0.614
100	***32***	0.422	***140***	***38***	0.526	***200***	***46***	0.492
100	***34***	0.35	***140***	***40***	0.394	***200***	***48***	0.398
100	***36***	0.22	***140***	***42***	0.308	***200***	***50***	0.282
100	***38***	0.178	***140***	***44***	0.22	***220***	***42***	0.768
100	***40***	0.086	***140***	***46***	0.126	***220***	***44***	0.714
100	***42***	0.06	***140***	***48***	0.082	***220***	***46***	0.602
100	***44***	0.03	***140***	***50***	0.036	***220***	***48***	0.502
100	***46***	0.016	***160***	***36***	0.75	***220***	***50***	0.376
100	***48***	0.008	***160***	***38***	0.648	***240***	***44***	0.738
100	***50***	0	***160***	***40***	0.564	***240***	***46***	0.69
120	***28***	0.8	***160***	***42***	0.456	***240***	***48***	0.586
120	***30***	0.726	***160***	***44***	0.358	***240***	***50***	0.464
120	***32***	0.65	***160***	***46***	0.262	***260***	***46***	0.756
120	***34***	0.55	***160***	***48***	0.13	***260***	***48***	0.686
120	***36***	0.444	***160***	***50***	0.108	***260***	***50***	0.568
120	***38***	0.344	***180***	***38***	0.764	***280***	***46***	0.784
120	***40***	0.236	***180***	***40***	0.666	***280***	***48***	0.73
120	***42***	0.186	***180***	***42***	0.584	***280***	***50***	0.64
120	***44***	0.104	***180***	***44***	0.472	***300***	***48***	0.78
120	***46***	0.046	***180***	***46***	0.38	***300***	***50***	0.654
120	***48***	0.024	***180***	***48***	0.272	***320***	***50***	0.742
120	***50***	0.016	***180***	***50***	0.19	***340***	***50***	0.792

Table 1: Convergence rates for FIML conditions with convergence lower than 80%

Note: N=sample size, PM=percent missing

Percentage Relative Bias										*Root Mean Square Error*									
N	**PM**	**FIML**	**SM**	**RI**	**N**	**PM**	**FIML**	**SM**	**RI**	**N**	**PM**	**FIML**	**SM**	**RI**	**N**	**PM**	**FIML**	**SM**	**RI**
100	***2***	-1.20	0.30	0.33	***600***	***2***	-1.34	0.02	-0.13	***100***	***2***	1.25	1.24	1.66	***600***	***2***	3.44	2.76	3.89
100	***10***	-5.94	1.90	0.27	***600***	***10***	-6.02	0.80	-0.18	***100***	***10***	3.21	3.01	3.93	***600***	***10***	10.65	6.11	8.78
100	***20***	-13.88	2.81	-1.44	***600***	***20***	-13.30	1.14	-0.74	***100***	***20***	5.39	4.73	5.93	***600***	***20***	22.07	9.84	14.69
100	***30***	-23.87	2.49	-8.73	***600***	***30***	-21.17	2.10	-1.08	***100***	***30***	7.65	5.56	7.93	***600***	***30***	33.94	13.56	20.79
100	***40***	-33.53	7.98	-0.09	***600***	***40***	-30.38	2.95	-0.88	***100***	***40***	9.95	6.96	10.32	***600***	***40***	47.34	17.35	27.72
100	***50***	NA	NA	NA	***600***	***50***	-40.18	5.14	-0.44	***100***	***50***	NA	NA	NA	***600***	***50***	62.45	22.03	39.17
200	***2***	-1.40	0.07	-0.17	***700***	***2***	-1.27	0.06	-0.09	***200***	***2***	1.84	1.66	2.41	***700***	***2***	3.63	2.93	4.13
200	***10***	-6.22	1.08	-0.10	***700***	***10***	-6.47	0.41	-0.55	***200***	***10***	4.95	4.02	5.44	***700***	***10***	12.91	6.67	9.47
200	***20***	-13.20	2.17	-0.38	***700***	***20***	-13.21	1.30	-0.15	***200***	***20***	8.88	6.35	8.19	***700***	***20***	25.03	10.97	15.52
200	***30***	-21.68	3.01	-1.62	***700***	***30***	-20.92	2.40	0.23	***200***	***30***	13.15	8.49	11.70	***700***	***30***	38.55	14.98	22.03
200	***40***	-30.46	6.00	-4.07	***700***	***40***	-30.21	2.90	0.25	***200***	***40***	16.95	9.97	15.31	***700***	***40***	54.95	18.11	30.21
200	***50***	-40.79	9.32	-8.89	***700***	***50***	-39.90	5.66	0.18	***200***	***50***	22.67	12.77	22.71	***700***	***50***	71.73	25.44	41.81
300	***2***	-1.29	0.08	0.02	***800***	***2***	-1.23	0.10	-0.09	***300***	***2***	2.15	1.95	2.85	***800***	***2***	3.93	3.06	4.62
300	***10***	-6.52	0.52	-0.29	***800***	***10***	-6.19	0.55	-0.19	***300***	***10***	6.76	4.73	6.82	***800***	***10***	14.22	7.23	10.24
300	***20***	-13.16	1.65	-0.66	***800***	***20***	-13.23	1.35	-0.35	***300***	***20***	12.09	7.56	10.55	***800***	***20***	28.42	11.89	17.23
300	***30***	-21.59	2.35	-1.03	***800***	***30***	-21.18	2.07	-0.48	***300***	***30***	18.49	9.45	13.54	***800***	***30***	44.80	16.22	24.04
300	***40***	-31.34	3.48	-2.64	***800***	***40***	-30.27	2.48	-1.09	***300***	***40***	25.44	11.67	19.29	***800***	***40***	62.49	19.61	33.83
300	***50***	-40.54	7.18	-5.56	***800***	***50***	-40.48	4.16	0.70	***300***	***50***	32.59	16.26	27.05	***800***	***50***	82.82	25.49	46.58
400	***2***	-1.19	0.18	0.08	***900***	***2***	-1.16	0.18	-0.02	***400***	***2***	2.57	2.30	3.33	***900***	***2***	4.24	3.30	4.76
400	***10***	-6.27	0.69	-0.31	***900***	***10***	-6.23	0.59	-0.42	***400***	***10***	8.12	5.21	7.71	***900***	***10***	16.16	8.24	11.33
400	***20***	-13.07	1.63	-0.49	***900***	***20***	-13.34	1.05	-0.44	***400***	***20***	15.21	8.11	12.12	***900***	***20***	31.93	11.82	18.24
400	***30***	-21.54	2.26	-1.49	***900***	***30***	-21.30	1.66	-0.30	***400***	***30***	23.49	10.37	16.30	***900***	***30***	50.01	15.73	24.12
400	***40***	-30.84	3.23	-1.91	***900***	***40***	-29.99	3.10	-0.19	***400***	***40***	32.87	14.67	22.82	***900***	***40***	69.54	21.26	34.36
400	***50***	-40.48	5.21	-0.88	***900***	***50***	-39.95	4.55	0.06	***400***	***50***	42.65	17.12	31.97	***900***	***50***	91.79	27.61	49.63
500	***2***	-1.19	0.18	-0.01	***1000***	***2***	-1.26	0.06	-0.09	***500***	***2***	2.87	2.54	3.72	***1000***	***2***	4.73	3.58	5.01
500	***10***	-6.30	0.63	-0.34	***1000***	***10***	-6.20	0.61	-0.24	***500***	***10***	9.73	5.88	8.54	***1000***	***10***	17.56	8.48	11.66
500	***20***	-13.99	0.66	-1.30	***1000***	***20***	-13.68	0.62	-0.61	***500***	***20***	19.37	8.94	13.66	***1000***	***20***	36.19	12.87	18.75
500	***30***	-21.32	2.08	-0.55	***1000***	***30***	-21.20	1.67	-0.54	***500***	***30***	28.82	12.50	19.45	***1000***	***30***	55.11	17.36	27.53
500	***40***	-30.84	2.84	-1.31	***1000***	***40***	-30.53	2.21	-0.43	***500***	***40***	40.49	15.08	25.00	***1000***	***40***	78.53	21.55	38.74
500	***50***	-40.30	5.36	0.70	***1000***	***50***	-40.01	4.60	0.51	***500***	***50***	52.10	19.26	34.39	***1000***	***50***	101.77	29.51	54.65

Table 2: Percentage Relative Bias and Root Mean Square Error of the $\Delta\chi^2$ from replications in which FIML models converged

Note: N=sample size, PM=percent missing, SM=supermatrix, RI=regression imputation

Percentage Relative Bias										*Root Mean Square Error*									
N	**PM**	**FIML**	**SM**	**RI**	**N**	**PM**	**FIML**	**SM**	**RI**	**N**	**PM**	**FIML**	**SM**	**RI**	**N**	**PM**	**FIML**	**SM**	**RI**
100	***2***	-1.20	0.30	0.33	***600***	***2***	-1.34	0.02	-0.13	***100***	***2***	1.25	1.24	1.66	***600***	***2***	3.44	2.76	3.89
100	***10***	-5.94	1.90	0.27	***600***	***10***	-6.02	0.80	-0.18	***100***	***10***	3.21	3.01	3.93	***600***	***10***	10.65	6.11	8.78
100	***20***	-13.88	2.92	0.20	***600***	***20***	-13.30	1.14	-0.74	***100***	***20***	5.39	4.64	5.96	***600***	***20***	22.07	9.84	14.69
100	***30***	-23.87	3.89	-1.08	***600***	***30***	-21.17	2.10	-1.08	***100***	***30***	7.65	5.70	8.16	***600***	***30***	33.94	13.56	20.79
100	***40***	-33.53	6.53	-0.06	***600***	***40***	-30.38	2.95	-0.88	***100***	***40***	9.95	7.94	12.35	***600***	***40***	47.34	17.35	27.72
100	***50***	NA	14.52	12.67	***600***	***50***	-40.18	5.39	0.18	***100***	***50***	NA	10.69	17.49	***600***	***50***	62.45	22.40	38.93
200	***2***	-1.40	0.07	-0.17	***700***	***2***	-1.27	0.06	-0.09	***200***	***2***	1.84	1.66	2.41	***700***	***2***	3.63	2.93	4.13
200	***10***	-6.22	1.08	-0.10	***700***	***10***	-6.47	0.41	-0.55	***200***	***10***	4.95	4.02	5.44	***700***	***10***	12.91	6.67	9.47
200	***20***	-13.20	2.17	-0.38	***700***	***20***	-13.21	1.30	-0.15	***200***	***20***	8.88	6.35	8.19	***700***	***20***	25.03	10.97	15.52
200	***30***	-21.68	3.20	-0.98	***700***	***30***	-20.92	2.40	0.23	***200***	***30***	13.15	8.51	11.74	***700***	***30***	38.55	14.98	22.03
200	***40***	-30.46	6.13	-0.76	***700***	***40***	-30.21	2.90	0.25	***200***	***40***	16.95	10.20	15.58	***700***	***40***	54.95	18.11	30.21
200	***50***	-40.79	8.76	3.02	***700***	***50***	-39.90	5.63	0.43	***200***	***50***	22.67	13.62	23.40	***700***	***50***	71.73	25.42	41.95
300	***2***	-1.29	0.08	0.02	***800***	***2***	-1.23	0.10	-0.09	***300***	***2***	2.15	1.95	2.85	***800***	***2***	3.93	3.06	4.62
300	***10***	-6.52	0.52	-0.29	***800***	***10***	-6.19	0.55	-0.19	***300***	***10***	6.76	4.73	6.82	***800***	***10***	14.22	7.23	10.24
300	***20***	-13.16	1.65	-0.66	***800***	***20***	-13.23	1.35	-0.35	***300***	***20***	12.09	7.56	10.55	***800***	***20***	28.42	11.89	17.23
300	***30***	-21.59	2.36	-0.93	***800***	***30***	-21.18	2.07	-0.48	***300***	***30***	18.49	9.44	13.61	***800***	***30***	44.80	16.22	24.04
300	***40***	-31.34	3.66	-1.55	***800***	***40***	-30.27	2.48	-1.09	***300***	***40***	25.44	11.81	19.71	***800***	***40***	62.49	19.61	33.83
300	***50***	-40.54	7.76	1.05	***800***	***50***	-40.48	4.33	1.14	***300***	***50***	32.59	16.28	28.11	***800***	***50***	82.82	25.72	47.12
400	***2***	-1.19	0.18	0.08	***900***	***2***	-1.16	0.18	-0.02	***400***	***2***	2.57	2.30	3.33	***900***	***2***	4.24	3.30	4.76
400	***10***	-6.27	0.69	-0.31	***900***	***10***	-6.23	0.59	-0.42	***400***	***10***	8.12	5.21	7.71	***900***	***10***	16.16	8.24	11.33
400	***20***	-13.07	1.63	-0.49	***900***	***20***	-13.34	1.05	-0.44	***400***	***20***	15.21	8.11	12.12	***900***	***20***	31.93	11.82	18.24
400	***30***	-21.54	2.26	-1.49	***900***	***30***	-21.30	1.66	-0.30	***400***	***30***	23.49	10.37	16.30	***900***	***30***	50.01	15.73	24.12
400	***40***	-30.84	3.24	-1.81	***900***	***40***	-29.99	3.10	-0.19	***400***	***40***	32.87	14.63	22.83	***900***	***40***	69.54	21.26	34.36
400	***50***	-40.48	5.71	0.89	***900***	***50***	-39.95	4.60	0.21	***400***	***50***	42.65	17.09	31.60	***900***	***50***	91.79	27.67	49.89
500	***2***	-1.19	0.18	-0.01	***1000***	***2***	-1.26	0.06	-0.09	***500***	***2***	2.87	2.54	3.72	***1000***	***2***	4.73	3.58	5.01
500	***10***	-6.30	0.63	-0.34	***1000***	***10***	-6.20	0.61	-0.24	***500***	***10***	9.73	5.88	8.54	***1000***	***10***	17.56	8.48	11.66
500	***20***	-13.99	0.66	-1.30	***1000***	***20***	-13.68	0.62	-0.61	***500***	***20***	19.37	8.94	13.66	***1000***	***20***	36.19	12.87	18.75
500	***30***	-21.32	2.08	-0.55	***1000***	***30***	-21.20	1.67	-0.54	***500***	***30***	28.82	12.50	19.45	***1000***	***30***	55.11	17.36	27.53
500	***40***	-30.84	2.84	-1.26	***1000***	***40***	-30.53	2.21	-0.43	***500***	***40***	40.49	15.06	25.04	***1000***	***40***	78.53	21.55	38.74
500	***50***	-40.30	5.33	1.47	***1000***	***50***	-40.01	4.64	0.75	***500***	***50***	52.10	19.13	34.44	***1000***	***50***	101.77	29.76	55.24

Table 3: Percentage Relative Bias and Root Mean Square Error of the $\Delta\chi^2$ from selected conditions

Note: N=sample size, PM=percent missing, SM=supermatrix, RI=regression imputation