1-1-2014

Planned Missing Data Designs & Small Sample Size: How Small is Too Small?

Fan Jia
University of Kansas

E. Whitney G. Moore
Wayne State University, whitneymoore@wayne.edu

Richard Kinai
University of Kansas

Kelly S. Crowe
University of Kansas

Alexander M. Schoemann
East Carolina University

See next page for additional authors

Recommended Citation

Jia, F., Moore, E. W. G., Kinai, R., Crowe, K. S., Schoemann, A. M., & Little, T. D. (2014). Planned missing data designs with small sample sizes: How small is too small? *International Journal of Behavioral Development, 38* (5), 435-452. doi: 10.1177/0165025414531095
Available at: http://digitalcommons.wayne.edu/coe_khs/54

Authors
Fan Jia, E. Whitney G. Moore, Richard Kinai, Kelly S. Crowe, Alexander M. Schoemann, and Todd D. Little

Planned Missing Data Designs with Small Sample Sizes: How Small is Too Small?

Fan Jia,[1] E. Whitney G. Moore,[2] Richard Kinai,[1] Kelly S. Crowe,[1]
Alexander M. Schoemann[3] and Todd D. Little[4]

[1] University of Kansas, [2] University of North Texas, [3] East Carolina University,
[4] Texas Tech University

Abstract

Utilizing planned missing data (PMD) designs (ex. 3-form surveys) enables researchers to ask participants fewer questions during the data collection process. An important question, however, is just how few participants are needed to effectively employ planned missing data designs in research studies. This paper explores this question by using simulated three-form planned missing data to assess analytic model convergence, parameter estimate bias, standard error bias, mean squared error (MSE), and relative efficiency (RE).Three models were examined: a one-time point, cross-sectional model with 3 constructs; a two-time point model with 3 constructs at each time point; and a three-time point, mediation model with 3 constructs over three time points. Both full-information maximum likelihood (FIML) and multiple imputation (MI) were used to handle the missing data. Models were found to meet convergence rate and acceptable bias criteria with FIML at smaller sample sizes than with MI.

Keywords: planned missing designs, simulation, full information maximum likelihood (FIML), multiple imputation (MI), 3-form survey

Introduction

High quality approaches to infer population information in the presence of missing data, such as full-information maximum likelihood (FIML) and multiple imputation (MI), have been around for decades (Dempster, Laird, & Rubin, 1977; Rubin, 1976); however, implementing either of these procedures was very time-consuming and computationally intense. With the rapid advance of computer technology, modern approaches are now readily available in most current statistical software packages (Enders, 2010; Enders & Gottschall, 2011; Graham, Cumsille, & Elek-Fisk, 2003). Specifically, either FIML or MI can be used to recapture the missing information and represent the population's characteristics very well (Dempster et al., 1977; Enders, 2010; Graham, Taylor, & Cumsille, 2001; Graham et al., 2006; Rubin, 1976). The benefits of using either FIML or MI for handling missingness is that they increase power compared to traditional methods of handling missing data (e.g. listwise deletion), while estimating unbiased parameter values.

Now that FIML and MI are widely available, an alternate view of missing data has become viable; namely, to plan for and easily address past limitations caused by missing data (Graham, 2009). This planned missing data (PMD) approach involves intentionally introducing missingness into the data collection in a way that results in missing data patterns that are "missing completely at random." Methodologists have started investigating the strengths and limitations of using a PMD approach, and enough empirical evidence now exists to justify integrating this approach as a routinely used research methodology (Graham, 2009). An important strength for developmental researchers is the ability to collect longitudinal data, while minimizing the burden placed upon their participants. The purpose of this paper is to provide clearer guidance for researchers using PMD survey designs (specifically the three-form design;

see below) with small sample sizes (N = 60 - 300).

Background

A PMD is based upon a particular characteristic pattern of missingness among the data: completely random missingness. There are three categories or types of missingness in data: missing completely at random (MCAR), missing at random (MAR), and unintentional, but systematic missing data (i.e., missing not at random, MNAR; Graham, 2009; Rubin, 1976; Schafer & Graham, 2002). When data is missing completely at random (MCAR) there is no systematic cause for the missingness. When the reason for the data's missingness is captured by other variables in the data, then it is called a missing at random (MAR) process (Graham, 2009; Rubin, 1976; Schafer & Graham, 2002). Lastly, MNAR data is missing for a reason that was not captured in anyway by the other variables measured. Since the cause of missingness is not measured in MNAR situations, we lack relevant information to inform the FIML or MI process when missing data is MNAR. Both MCAR and MAR data can be recaptured by FIML or MI approaches by using the relationships (covariances) that exist within the observed data to inform the missing information. To know how much the covariances are able to inform the missing information, the proportion of information lost due to missing data can measured. This value is the fraction of missing information (FMI). The lower the FMI, the less information lost, and the higher the quality of estimation (Enders, 2010; Savalei & Rhemtulla, 2012).

The ability to recapture data that is missing due to either a MCAR or MAR process is the foundation for PMD theory and practice. The data that is missing by design is, by definition, MCAR, and, therefore, is easily recaptured. When either MCAR or MAR data is present, then the existing data can potentially inform the modern missing data technique's (see FIML and MI specifics below) process to recapture the missing information (Dempster et al., 1977; Enders, 2010; Graham et al., 2001; Graham et al., 2006; Rubin, 1976).

For PMD designs to be effective and have the ability to result in improved data quality, FIML or MI need to be used to recapture the data that was missing by design. There are, however, slight differences that may affect the success of their application. FIML estimation is done during model estimation (Graham, 2009). Therefore, only the variables in an analysis model, plus a select few auxiliary variables – variables not part of the analysis model that are included only to inform the estimation of any MAR mechanism – may inform model parameter estimates (Graham, 2009; Graham, et al., 2003). FIML estimation is now often the default in SEM software (e.g., Mplus and `lavaan`). MI, infers the descriptive statistics of the dataset's variables using all possible information from all answered items. MI was originally intended for use with datasets having a large number of variables, particularly when those variables may not all be used in a single model (Rubin, 1987). The imputed values are based upon the relationships among the present data. Each dataset that is imputed will have slightly different values in the missing data positions. The greater the variability in values between imputations, the greater the uncertainty due to the missingness present. By using multiple imputed datasets (20 is the suggested minimum), the parameter estimates and standard error values are less biased than if only a single imputed dataset were used (Graham, et al., 2007; Rubin, 1996; Van Buuren, Boshuizen, & Knook, 1999). Although researchers agree that with enough imputations, MI results are equivalent to FIML results, there is not yet consensus on the number of imputations that should be used (Bodner, 2008; Savalei & Rhemtulla, 2012). Once multiple datasets have been imputed, then the analysis model is run upon each dataset and the results combined by Rubin's Rules (see Rubin, 1987), which is now an automatic step in many software packages. One benefit of MI is that imputed items can be averaged (e.g., such as subscale scores) prior to model estimation. When using FIML, items averaged to form subscale values result from averaging across missing values, which may lead to more biased estimates (Schafer & Graham, 2002).

Purpose

The purpose of this paper is to provide preliminary guidance on sample size minimums for researchers using PMD 3-form survey designs. After fitting models to the data, the convergence of the PMD models by FIML and MI were examined. These results provide guidance to researchers regarding the minimum sample size FIML and MI may need in order to use these designs. The PMD models' relative bias of parameter estimates and standard errors, mean squared error (MSE), and relative efficiency (RE) were also assessed. The less the relative bias, the more the PMD models' values are properly representing the true parameter values as specified in the simulation (Schafer & Graham, 2002). These results suggest the minimum sample size FIML and MI may need in order to obtain accurate and precise parameter estimates.

Methods

The simulation designs involved three latent variable models: a one-time point, cross-sectional confirmatory factor analysis (CFA) model, a two-time-point CFA model, and a three-time-point mediation model. Simulations for sample sizes ranging from 60 to300 were conducted to examine the minimum sample size necessary to successfully complete an SEM analysis of 3-form PMD design data collected cross-sectionally (Figure 1) and longitudinally (Figure 2 and 3). For the guidance from these simulations to be as applicable as possible, an additional 5% MCAR missingness was included on top of the PMD design, so that the impact of a commonly present amount of missingness in real-world datasets would also be present and accounted for. The models' parameters – loadings, within time latent covariances, autoregressive coefficients and cross-lag coefficients – were specified to be within the ranges commonly seen in psychological and developmental research. The specifics for each model are detailed below. To streamline the methods section, the longitudinal studies' expanded methodology will build upon the cross-sectional model's methodology provided initially. All study designs include performance comparisons between FIML and MI.

3-Form PMD Survey Design

The 3-form PMD survey design is being adopted by researchers in the social sciences, and so it will be the PMD design focus of this article. Other, more complex, multi-form designs are also, of course, possible. The basic principle for the 3-form design is that scale items are divided into four blocks: X, A, B, and C (Graham, 2009; Graham et al., 2006; Moore, 2011; Rhemtulla et al., 2012). The X-block is also called the common block and is the block included in all survey versions (e.g., Version 1, 2, and 3; Graham, 2009; Graham et al., 2006; Moore, 2011; Rhemtulla et al., 2012). The X-block is comprised of demographic information and often includes the strongest and most representative indicator of each construct. Each construct's remaining indicators are then distributed across the A-, B-, and C-blocks as evenly as possible based upon their psychometric properties (see Moore, 2011 for a more detailed 3-form design example; and Little, Jorgensen, Lang, & Moore, 2013 for an example of analyses with data collected using a 3-form design). Finally, three survey versions are designed by combining two of the A-, B-, and C-blocks of items together with the X block. For example, Version 1 would have the X-block items; along with the items from blocks A and B (See Table 1). In this paper's example, the X set contained one item from each latent construct, and the other items were evenly spread across A, B and C sets. This distribution approach resulted in 33% of the items being in the X-block, and 22% in each of the A-, B-, and C-blocks. Therefore, each survey version presents only 78%

of all the study items to each participant (See Table 1).

Data Generation

Model 1. The simplest latent structure model used for data generation was a cross-sectional CFA model (Model 1) with three latent variables, η_1, η_2, and η_3. Each latent variable was indicated by three items (see Figure 1). Latent variances were fixed at 1 for model identification. A range of population values representing the variety of conditions seen in applied studies were simulated. Specifically, the factor loadings ranged from 0.7 to 0.85, and were simulated at 0.05 increments (i.e., 0.7, 0.75, 0.80, and 0.85). The factor loadings in a given model had the same value across each construct (i.e., tau-equivalent). Additionally, the within-time correlations between latent variables were simulated to range between 0.2 and 0.5 in increments of 0.1 (i.e., 0.2, 0.3, 0.4, and 0.5). Simulating each of these possible combinations resulted in 16 conditions (4 loadings x 4 covariances) to be assessed.

Model 2. The two-time-point CFA model (Model 2) had the same latent variables and indicators as Model 1, but measured at two time points (see Figure 2). The Time 1 latent variances were fixed at 1, Time 2 latent variances and all latent covariances were freely estimated. The population values of the factor loadings and within time covariances were simulated with the same ranges as for the cross-sectional model (Model 1) mentioned above. The conditions used to derive the population cross-lag covariance matrices for Model 2 were based upon autoregressive coefficient values simulated to range from .4 to .9 in .1 increments, and cross-lag coefficient values simulated for a range from .1 to .4 in increments of .1. These conditions were used to derive the population covariance matrices for the manifest variables using (simplified from equation 4.7 of Bollen 1989):

$$\Sigma = (1 - B)^{-1} \Psi (1 - B)^{-1'},$$

where Σ represents the manifest covariance matrix, B is the coefficient matrix for latent endogenous variables, and Ψ is the latent covariance matrix. Simulating each possible combination resulted in 340 conditions (4 loadings x 4 covariances x 6 autoregressive coefficients x 4 cross-lagged coefficients, minus 44 combinations that produced negative latent variances).

Model 3. The most complex, three-time-point mediation model (Model 3) had the same latent variables and indicators as Model 1 and Model 2, but measured at three time points (see Figure 3). The Time 1 latent variances were fixed at 1, while the other latent variances were freely estimated. The population values of factor loadings and within-time covariances were simulated with the same ranges as for Model 1. Model 3's autoregressive coefficients and cross-lag coefficients were the same values as we used to deriving the Model 2 parameters. Therefore Model 3 had the same number of conditions as Model 2.

Because modeling with sample sizes was the primary focus of these simulations, we assessed each model's conditions with sample sizes from 60 to 300 at increments of 20 (e.g., 60, 80, 100, 120, 140,…, 300). After examining these results, the threshold range for each model to reach convergence and accurate estimates was then more precisely investigated. This was done by conducting additional simulations in which the sample sizes assessed were in increments of five..

Missing values were imposed based on a three-form planned missing data design. In the longitudinal models (Model 2 and Model 3), we systematically switched the form each participant received. Specifically, participants who responded to Version 1 at time one, responded to Version 2 at time two, and Version 3 at time three. In the last step, 5% of unplanned MCAR missingness was imposed on top of the planned missingness. Therefore, the percentage of missing data in each survey version was 28%.

Two hundred samples (replications) were drawn from each condition, and the three models described above were estimated for each of the samples using FIML and MI. In simulation studies there is always a trade-off between the number of replications and conditions examined. A large number of replication (e.g., 1000) is usually recommended to improve precision (Bentler, 1995). However, Skrondal (2000, page 157) argues that this “practice unduly favors precision at [the] expense of external validity.” The empirical applications of statistical methods deserve more attention than “the exaggerated precision of convention” standard. Using a comparatively small, but not too small, number of replications (i.e., 200) allows us to examine more simulation conditions, and thereby achieve a higher level of validity without scarifying too much precision.

All samples were generated and analyzed in R using the `simsem` package (Version 0.5-0; Pornprasertmanit, Miller, & Schoemann, 2012). The `simsem` package was designed to automate Monte Carlo Simulations that used either a CFA or SEM analytical framework. Using functions in `simsem`, data were first generated and then modified (i.e., simulated three-form planned missing data design and unplanned MCAR). The models analyzed with the FIML approach were run through `lavaan` (Rosseel, 2012). The models analyzed by the MI approach were run through `Amelia` (Version 1.6.1; Honaker, King, & Blackwell, 2011), and the resultant 20 imputed datasets were run through `lavaan` and the results combined by Rubin’s (1987) Rules. The decision to use 20 imputed datasets was made based on the findings from previous studies and the computational effort needed. Rubin (1987) suggested 2 to 10 imputations were sufficient in most realistic situations. Then, von Hippel, (2005) in response to Hershberger and Fisher (2003), supported Rubin’s guidance, and argued that the marginal gains of using more than 10 imputations is outweighed by the computational costs. In 2007, Graham, Olchowski and

Gilreath – who also noted that computational cost is worth considering – suggested using at least 20 imputed datasets. Thus, given the huge number of simulation conditions already required a heavy computational effort, in this study, we considered 20 imputed datasets a reasonable number.

Analysis

The effectiveness of FIML and MI to properly estimate the true parameter values for each model across sample size conditions was assessed over the following characteristics: model convergence, relative biases of parameter estimates and standard errors, MSE, and RE.

Model Convergence. To assess a model, it must converge. Therefore, our first step was to examine the effect of sample size and model complexity on the quantity and quality of the model's convergence when applying FIML or MI. The convergence rate helped us answer our first question of how small a sample size one could have when collecting data with a 3-form PMD design. Our next convergence question was on convergence rate differences between FIML or MI. For the FIML conditions we consider a replication as converged if the number of iterations of the replication is smaller than 250 (the default in `lavaan`). Then the rate of convergence for FIML results can be simply obtained by counting the number of converged replications in each condition and divided it by the total number of replications (i.e., 200). However, assessing the convergence rate of MI results is not as straightforward, since each replication is comprised of 20 imputed datasets. As there is no guidance about how many imputed datasets need to converge when using MI, we chose the strictest cutoff (i.e., 20 of 20 datasets converge) to make the convergence rate more comparable to FIML. In other words, we considered a replication convergent only when the model converged on all 20 of its imputed datasets. Then, for a given sample size in each model, we computed the average convergence rate across all parameter value conditions.

Relative bias of parameter estimates and standard errors (SE). Once a model converges, the next important question is how accurate are the models' values – parameter estimates. To determine the minimum sample size at which parameter estimates were unbiased, we focused on the model's factor loadings, and latent covariances/coefficients. This assessment was based upon the relative parameter bias, which refers to the percentage of the raw bias—the difference between the average value across replications and the true parameter value—relative to the population value. In other words, the relative parameter bias is the ratio determined by dividing the raw bias (i.e., difference) by the population value. Therefore, the smaller the absolute relative parameter bias, the more properly the true parameter value is estimated by the model. In addition to assessing the parameter estimates' bias, we also assessed the standard errors' bias. Standard errors (SE) represent the variability of the parameter estimate of interest (e.g., factor loading, latent covariance, or latent pathway). A parameter's SE is used in determining if the parameter estimate is significantly different from zero. Assessing the relative SE bias was similar to the relative parameter bias above. First, we evaluated how much each sample's SE deviated from its population value (i.e., raw bias). This raw value is divided by the population value to get the relative SE bias ratio. Although we are not able to set the population's SE value in our simulation study, the standard deviation of the empirical sampling distribution of parameter estimates serves this purpose. After computing the absolute the relative parameter and SE biases, the values for the same parameter type (e.g., factor loadings, within-time-point covariances) were averaged.

Mean squared error (MSE). Another measure of an estimate's overall accuracy, its MSE, was also used to evaluate the performance of FIML and MI at the different sample sizes. The

MSE is equal to the squared bias of the estimate plus the variance of the estimate. Of the two components, the first measures the estimator's bias (e.g., accuracy), while the second measures the variability of the estimator (precision). Generally, a good estimator has small combined variance and bias. That is, when performing well, FIML and MI would produce accurate and precise parameter estimates. Thus, the smaller the parameter's MSE, the better the model fits the data.

Relative efficiency (RE). Relative efficiency (RE) measures the amount of information loss due to missing data. It is negatively related to the fraction missing of information (FMI). RE is computed as a ratio of the sampling variances (i.e., squared standard errors) of the complete data estimates to the missing data estimates (Rhemtulla, Jia, Wu, inpress). Ranging from 0 to 1, RE is computed for each parameter and could be interpreted as the loss of statistical power caused by missing data (Savalei & Rhemtulla, 2012). For example, data was collected with missingness from 100 participants, and a parameter is found to have an RE of .8, then a complete dataset could produce the same parameter's information with only 80 (0.8*100) participants.

Criteria. Several criteria are examined to determine how small a sample size is acceptable. The first criterion (C1) is that convergence rate is greater than .9 (Muthén & Muthén ,2002). The second criterion (C2) is that the parameter estimate's relative bias does not exceed .05 and standard error bias does not exceed .1 (Hoogland and Boomsma's, 1998). C1 and C2 are commonly used criteria and they two together help avoid nonconvergence, bias or inaccuracy in a statistic procedure. In addition, Hoogland and Boomsma's (1998) proposed a more stringent criterion (C3), suggesting that the mean standard error bias across parameters should be smaller than .05. Larger sample size might be needed when C3 is employed.

We ran analyses of variance (ANOVAs) to investigate the effects of the design factors on the magnitude of the relative bias of the parameter estimates and standard errors. Four univariate ANOVAs were performed on FIML point estimate bias, FIML standard error bias, MI point estimate bias, and MI standard error bias. The simulation design factors (e.g. factor loading values, sample size) were all treated as the between-subject factors. Due to the large number of total replications (with the sample size increments of 20, there were 208,000 conditions for Model 1 and 4,992,000 for Model 2 and Model 3), statistical significance tests were not reported; instead, we examined the partial η^2 for each effect and only reported and interpreted those that exceeded .01, which is considered a "small effect" according to Cohen (1973).

Results

Model 1

Figure 4a depicts convergence rates for Model 1. When N = 60, FIML tended to converge better than MI, though both had very low convergence rates (.71 and .55). As expected, when sample size increased, so did the convergence rates. A 90% convergence rate was reached by FIML when N=90, and by MI when N=110. Both FIML and MI achieved 100% convergence when N=180.

Next, the relative bias of the parameters and standard errors for this model's factor loadings (λ; See Figure 5a) and latent covariances (ϕ; See Figure 5b) were examined. Even when sample size was as small as 60, the biases in the point estimates of loadings were still trivial for both FIML and MI. The top-right panel shows that the SE bias for loadings estimated by FIML was less than MI. As N increased, the difference in SE bias estimated by FIML and MI decreased. With larger sample sizes (N > 150) FIML and MI performed equally well, with bias values at or below .05. The latent covariances' relative bias was much higher than the corresponding factor loadings' relative bias at all sample sizes (N = 60 - 300). Specifically, when applying FIML, once N ≥ 70, the latent covariances in Model 1 were accurately estimated. While MI estimates of Model 1's latent covariances were accurate when N ≥ 115. Different than the loadings, the latent covariance SE bias showed no notable difference between FIML and MI at all sample sizes. Therefore, this single-time point CFA model's parameters were acceptably estimated with small sample sizes by both FIML (N ≥ 70) and MI (N ≥ 115).

None of the design factors had any notable effect on the point estimate bias, with either FIML or MI. The noticeable effects ($\eta^2 \geq .01$) were present only on SE bias. When estimating factor loadings with MI, the SE bias decreased with increases in the N ($\eta^2 = .014$), factor loadings ($\eta^2 = .016$) and latent covariance ($\eta^2 = .042$). In the estimation of latent covariances, N had effect on SE bias with both FIML and MI ($\eta^2 = .08$, and .05, respectively).

MSE and RE in Model 1 (see Appendix 1) did not show notable patterns. MSEs were almost identical (MSE < .05) when applying either FIML or MI. REs remained stable across samples size. On average, FIML tended to be less efficient than MI when estimating factor loadings, but more efficient for covariances.

In addition to Figure 5, Table 2 provides a decision guide for choosing a minimum sample size with Model 1. For FIML, N = 90 was the minimum number to get the model converged. This sample size was also large enough to achieve accurate estimates of the major parameters in the model. MI results showed a different pattern. A sample size of 110 was the minimum number for a three construct model to converge. However, to obtain accurate estimation of latent covariances, which are often a parameter of interest in a simple CFA model, the sample size should be no smaller than 115. To meet the most stringent criterion for SE -- mean standard error bias across parameters should be smaller than .05 -- sample sizes of 120 and 155 appear to be needed for FIML and MI, respectively, when fitting cross-sectional models similar to Model 1.

Model 2

Figure 4b depicts convergence rates for Model 2. When N = 60, convergence for both FIML (27%) and MI (56%) was much lower than for Model 1. However, FIML's convergence rate achieved 93% by N=80, and 99% with N=110. When MI was applied, Model 2 converged better than Model 1 at every sample size examined. MI's convergence rate achieved 90% with N=90, and 99% with N = 135. Both MI and FIML attained acceptable convergence rates with smaller samples sizes for Model 2 compared to Model 1.

Next, the parameter bias and SE bias for this two-time point CFA model's parameters were examined: factor loadings (λ; See Figure 6a), within-time-point latent covariances ($\phi_{WT;}$ Figure 6b), and between-time-point latent covariances ($\phi_{BT;}$ Figure 6c). Neither FIML, nor MI, produced notable parameter bias in estimating factor loadings; although FIML performed slightly better than MI. Even when sample size was as small as 60, the biases in point estimates of loadings were still below the .05 cutoff value for both FIML and MI. The top- right panel (Figure 6a) shows that the SE biases for loadings when FIML or MI was applied were overlapped and smaller than .1 at all sample sizes. The parameter bias of within-time-point latent covariances and between-time-point latent covariances told the same story: unbiased estimates were generated when N = 65 for FIML and N = 115 for MI. The SE bias of covariances for FIML and MI were very close and all smaller than .1, and both gradually decreased as N increased. The mean SE bias of covariances reached .05 or less for FIML at N = 100, and for MI at N = 200.

None of the design factors had any notable effect on the point estimate bias, with either FIML or MI. For the FIML conditions, two effects were detected for the SE bias in factor loadings: N (η2 = .102), and factor loadings (η2 = .044). As N and the population value of factor loadings increased the bias in factor loadings SE became smaller. N also was found to have a negative effect on the SE bias of within-time-point latent covariances and between-time-point latent covariances (η2 = .015 and 0.017, respectively). For the MI conditions, the SE bias in factor loadings was influenced by two design factors, N (η2 = .012) and within-time-point covariance (η2 = .013). The increase in N and the decrease in the population value of within-time-point covariance resulted in the decrease of SE bias in factor loadings. In addition, N influenced the SE bias of within-time-point covariance (η2 = .017).

Similar to Model 1, MSEs for FIML and MI in Model 2 (see Appendix 2) were almost identical and all smaller than .05. The REs for the MI conditions were found to be lower than those for the FIML conditions (see Appendix 2).

In addition to Figure 6, Table 4 provides a decision guide for choosing a minimum sample size with Model 2. For FIML, N = 80 was the minimum number to get the model converged with unbiased parameter estimates. MI results showed a slightly different pattern. A sample size of 90 was the minimum number for our two-time point model to converge. Similar to the patterns seen for Model 1’s cross-sectional CFA, a greater sample size (N ≥ 115) was needed to obtain accurate latent covariance estimates, which usually are the parameter of interest in a simple CFA model. However, for more precise estimation (mean SE < .05), 100 and 200 are suggested as the smallest sample sizes for FIML and MI, respectively, when fitting two-time point models, such as Model 2.

Model 3

Figure 4c depicts the convergence rates for the three-time point, mediation model (Model 3). When N = 80, convergence for both FIML (0%) and MI (30%) was poor for Model 3. FIML’s convergence rate achieved 90% by N=130. MI’s convergence rate achieved 90% with N=160.

As expected, this more complex model needed a larger sample size for either FIML or MI to successfully attain a 90% convergence rate.

Next, the parameter bias and SE bias for this model's parameters were examined: factor loadings (λ; See Figure 7a), autoregressive coefficients (β_{AR}; Figure 7b), and mediation pathways (β_M, i.e., $\beta_{5,1}$ and $\beta_{9,5}$; see Figures 3 and 7c). Since FIML did not converge until N = 90, and MI also performed poorly at N = 60 and 80, we only present the bias and MSE when $N \geq 90$. As long as Model 3 converged, the factor loadings' parameter biases were smaller than .05 (see Figure 7a). Comparatively, loadings' SE bias was not surprisingly sensitive to convergence, especially for FIML. Bias was greater than .10 with FIML when N < 100. As sample size increased, SE bias for both FIML and MI diminished. However, neither consistently had SE bias values of .05 or less.

The latent pathways' relative parameter bias and SE bias decreased as sample size increased (Figure 7b). However, the revealed patterns for FIML and MI were both different from each other, and from the prior models. First, when FIML was applied to N = 90, there was a .1 bias; however, when N = 95, the FIML autoregressive pathways' bias averaged .05, and continued to improve as sample size increased. The SE bias for the FIML estimated autoregressive paths showed a similar pattern to loadings, but with values slightly larger at all sample sizes: attaining .10 by N = 105, and decreasing as N increased. Second, when MI was used to handle missing data, the autoregressive pathway estimates' bias averaged .07 when N = 90, and then gradually decreased until it reached .05 at N = 175; while the SE bias gradually decreased and fell below .10 at N = 105. Since this is a mediation model, we were keenly interested in the mediation paths: $\beta_{5,1}$ and $\beta_{9,5}$. As depicted in Figure 7c, FIML had extremely poor performance in estimating this mediation until $N \geq 125$. The trend for the SE bias when FIML estimated this mediation was similar to the autoregressive effects, only with slightly higher bias values. When MI was applied, the mediation pathway estimates' bias tended to bounce around the .07 value, until N ≥ 170, at which point acceptable (.05) estimate bias was reached. The SE bias for the MI estimated mediation pathways was .11 at N = 90, and dropped as N increased.

The design factors had no effect on the point estimate bias for either the FIML or MI conditions. Notable effects of N on the estimates' bias of autoregressive and mediation paths (η^2 = .04 and .013, respectively) were found for the FIML conditions. For the MI conditions, the SE bias in factor loadings was influenced by the population value of factor loadings (η2 = .027). As the value of factor loadings increased from .7 to .85, the marginal mean of the factor loading's SE bias changed from negative to positive.

The MSE and RE values for Model 3 can be found in Appendix 3. Similar to Model 2, the factor loadings' MSE values showed no difference from applying FIML or MI, and were minimal (MSE < .03). For the autoregressive and the mediation effects, MSE values generated by FIML and MI were also very close. The factor loadings' RE values for FIML conditions were smaller than those for MI conditions. In contrast, when estimating the autoregressive and the mediation effects FIML had higher RE than MI.

Table 5 serves as a decision guide when choosing sample size for a three-time-point mediation model estimated using data collected utilizing a PMD 3-form survey design. More than with the prior models, the sample size needed to have acceptable convergence and unbiased estimation of the parameters differed. Therefore, taking all of these aspects into account the minimum sample size needed with FIML would be 130 to attain an acceptable convergence rate and unbiased latent parameter estimates; while, with MI, one might need to have a sample size no smaller than 175.

Discussion

We examined the ability to use 3-form PMD design with small sample sizes. The three simulated models were a small, cross-sectional model with three latent constructs; a moderate, two-time point model with three latent constructs at each time point; and a large, three-time point mediation model with three latent constructs at each time point. Three-form PMD missingness, and 5% MCAR missingness were imposed on the simulated data and then analyzed in a CFA or SEM framework. These analyses were conducted with either FIML estimation or 20 multiply imputed datasets.

There were three major model characteristics used to assess the ability of FIML and MI to handle missing data when a 3-form PMD approach to data collection was utilized. Each of these characteristics displayed expected trends. The first trend was that convergence rates increased as sample size increased. This was true for both FIML and MI across all three models, and all conditions. In addition, FIML consistently reached the 90% convergence rate with a small sample size than MI. The second trend identified was that the relative parameter bias diminished as sample size increased. Again, FIML estimation performed better than MI with respect to the sample size needed to produce estimates that met the acceptable relative bias cutoff. The third trend seen across all three models and conditions was that SE bias trended gradually down as the sample size increased. This trend was so consistent across models, conditions, and estimators, such that the SE bias for FIML and MI were often indistinguishable.

The differences in performance by FIML and MI are understandable. In theory, FIML and MI produce equivalent results when the input data and model are the same, and the number of imputations are infinite (Graham, Olchowski, & Gilreath, 2007; Collins, Schafer, & Kam, 2001; Savalei & Rhemtulla, 2012). In our study, the first condition – that the input data and model used with FIML and MI are the same – was met. However, the second condition – infinite imputations for MI – was not met, as only 20 imputed datasets were used for each model replication. Given this limitation on imputed datasets, FIML outperforming MI is theoretically expected (Savalei & Rhemtulla, 2012), and exactly what our study's findings illustrated. With very few exceptions, FIML produced more accurate parameter estimates than MI, and needed a smaller sample size to meet the 90% convergence rate. In fact, as long as a sample size was large enough for the FIML estimated model to converge, the parameter estimates were accurate. MI, on the other hand, needed a greater sample size to meet the 90% convergence rate, and estimate parameters with an acceptably small level of bias. As discussed, in this paper the 20 imputations were chosen for both theoretical and practical reasons. "Equivalent results" between FIML and MI only exist when the number of imputations goes to infinity. Empirical studies showed that to obtain identical estimates as FIML, one may need as many as 50,000 imputations (Savalei & Rhemtulla, 2012). Applied researchers are recommended to use a relative large number of imputations to obtain better estimates with MI when computer power allows, although "identical estimates as FIML" are not usually necessary. To illustrate, we did a follow-up study over a small set of conditions for each model, with N = 200. We found that the increase in the number of imputations (e.g number of imputations = 20, 100, 500) slightly reduced parameter bias toward FIML values, while little effect was seen on the SE bias (see Appendix 4).

A final trend illustrated by this study's results is the relationship between sample size needs, model complexity, and data available to inform the estimation process. This relationship is a good example of the importance of FMI to model convergence (Longford, 2005). Among the three models, the two-time point CFA model (i.e. Model 2) attained the 90% convergence rate at the lowest sample sizes (N_{FIML}= 80 and N_{MI}= 90). FMI is an informative measure of the information lost due to missingness for each parameter. Therefore, unlike percent missing data (i.e., global measure of dataset missingness), FMI is computed for each parameter. FMI is equal to 1 minus the ratio of sampling variances of the complete data estimate to the missing data estimate. Therefore, FMI ranges from 0 to 1, and a higher value of FMI suggests a stronger impact of missing data (Savalei & Rhemtulla, 2012). Savalei & Rhemtulla (2012) proposed a 3-step procedure to compute FMI via FIML. Following these steps we found that with N = 200 the average FMI across all parameters for Model 1, Model 2 and Model 3 were 0.31, 0.26 and 0.36, respectively. These values suggest that the reason Model 2 converged more successfully at the smaller sample sizes is because there was more information available to inform the model's complete covariance matrix compared to Models 1 and 3. **Application**

An important application of this finding then is that when preparing to collect longitudinal data with a 3-form PMD design, it is important to know the FMI for the different data collection waves to ensure that the covariance matrix is covered well enough to estimate a model that will converge on a model with acceptably unbiased parameter estimates. For example, say a researcher conducted a power analysis, obtained the FMI for a two time-point CFA model, and determine the minimum sample size necessary for that model was 80 participants. When this researcher went to analyze the first-wave only data, s/he may not have enough information for the estimator to converge on a model. If the researcher had run the power analysis, and learned the FMI on a cross-sectional CFA, then s/he would have known a higher minimum sample size was actually necessary for cross-sectional model to converge. This is an important finding for applied researchers, as relationships are often published from both the single wave and multi-wave data that are collected during longitudinal studies.

There are ways to increase a model's convergence rate that depend on FIML or MI being used. The inclusion of auxiliary variables can improve the FMI for a model estimated with FIML. In this study all available variables were included in the estimated model. However, often in applied research there are variables not included in the model that may still be able to inform the estimation of the model's covariance matrix. The traditional benefit of the MI approach over FIML is that all the variables collected could be included in the imputation process. Therefore, any "auxiliary" information is included. MI estimated models can benefit, however, from a greater number of imputed datasets. As, mentioned earlier, when all else is the same, then the greater the number of imputed datasets used, the more equivalent to FIML the MI results will be.

In conclusion, when applied researchers utilize PMD designs, the minimum sample size necessary for analytic models to successfully converge with acceptably unbiased parameter estimates it is not simply that a larger sample is better. How much more of a sample size is better is dependent upon the complexity of the analytic model, the missing data technique used, criteria for estimation assessment, and the FMI. This study's preliminary results suggest that when researchers use a PMD design with factor loadings greater than .7 and covariances of .2 to .5, to obtain reasonable convergence (converenge rate > .9) and parameter and SE estimates (parameter bias < .05, and SE bias < .1), for one time-point, cross-sectional models a minimum of N = 90 may be sufficient when FIML is utilized; while N = 115 may be sufficient when MI is utilized. As the model complexity increases with multiple time-points, it will be important for researchers to take into account the estimator and the parameters' FMI to determine an appropriate increase in sample size. This study's initial findings suggest at a minimum N = 80 for FIML and N = 115 for MI approaches with two time-point CFAs; and N = 130 for FIML and N = 175 for MI approaches with three- construct mediation models. When estimating the sample size necessary for a study, it will be important for researchers to include a power analysis that reflects the amount and pattern of missingness present due to the PMD approach utilized.

References

Bentler, P. M. (1995). *EQS structural equations program manual*: Multivariate Software.

Bollen, K. A. (1989). Structural equations with latent variables: Wiley-Interscience.

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement, 33, 107–112.*

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*(4), 330.

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

Enders, C. K. (2001). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement, 61*(5), 713-740.

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

Enders, C. K., & Gottschall, A. C. (2011). Multiple Imputation Strategies for Multiple Group Structural Equation Models. [doi: 10.1080/10705511.2011.532695]. *Structural Equation Modeling: A Multidisciplinary Journal, 18*(1), 35-54. doi: 10.1080/10705511.2011.532695

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology, 60*, 349-376.

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for Handling Missing Data. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (Vol. 2, pp. 87-114). Hoboken, NJ: John Wiley & Sons Inc.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31*(2), 197-218.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How Many Imputations are Really Needed? Clarifications of Multiple Imputation Theory. Prevention Science, 2007(8), 206-213.

Graham, J. W., Taylor, B. J., & Cumsille, P. E. (2001). Chapter 11: Planned Missing-Data Designs in Analysis of Change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 335-353). Washington, D.C.: American Psychological Association.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned Missing Data Designs in Psychological Research. *Psychological Methods, 11*(4), 323-343.

Hershberger, S. L., & Fisher, D. G. (2003). A note on determining the number of imputations for missing data. *Structural Equation Modeling, 10*(4), 648-650.

Honaker, J., King, G., Blackwell, M. (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software, 45(7), 1-47*. URL http://www.jstatsoft.org/v45/i07/.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling An overview and a meta-analysis. *Sociological Methods & Research, 26*(3), 329-367.

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

Moore, E. W. G. (2011). Planned Missing Data Survey Design. crmda.ku.edu.

Muthén, L. K., & Muthén, B. O. (2002). How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(4), 599-620.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling, 8*(2), 287-312.

Pornprasertmanit, S., Miller, P., & Schoemann A. (2012). simsem: SIMulated Structural Equation Modeling. R package version 0.4-6. http://CRAN.R-project.org/package=simsem.

Rhemtulla, M., Jia, F., Wu, W., & Little, T. D. (in press). Planned missing data designs to optimize the efficiency of latent growth parameter estimates. *International Journal of Behavior Development.*

Rhemtulla, M., & Little, T. D. (2012). Planned Missing Data Designs for Research in Cognitive Development. *Journal of Cognition and Development, 13*(4), 425-438.

Rhemtulla, M., Little, T. D., Moore, E. W. G., Gibson, K., & Wei, W. (2012). *Planned Missing Data Designs for Longitudinal Research*. Paper presented at the Society for Research on Adolescence, Vancouver, Canada.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48(2), 1-36*. URL http://www.jstatsoft.org/v48/i02/.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika, 63*(3), 581-592.

Rubin, D. B. (1987). Interval Estimation from Multiply-Imputed Data: A Case Study Using Agriculture Industry Codes. *Journal of Official Statistics, 3*, 375-387.

Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, *91*(434), 473-489.

Savalei, V., & Rhemtulla, M. (2012). On obtaining estimates of the fraction of missing information from FIML. *Structural Equation Modeling, 19*, 477-494.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research,* *8*(1), 3-15. doi: 10.1177/096228029900800102

Schafer, J. L., & Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Physchological Methods, 7*(2), 147-177. doi: 10.1037//1082-989X.7.2.147

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research, 35(2), 137-167.*

van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine, 18*(6), 681-694. doi: 10.1002/(sici)1097-0258(19990330)18:6<681::aid-sim71>3.0.co;2-r

von Hippel, P. T. (2005). TEACHER'S CORNER: How Many Imputations Are Needed? A Comment on Hershberger and Fisher (2003). *Structural Equation Modeling, 12*(2), 334-335.

Table 1
Three-Form PMD Survey Design

Construct & Item Number	PDM Block Assignment	Survey Version Assignment		
Construct 1, Item 1	X	1	2	3
Construct 1, Item 2	A	1		3
Construct 1, Item 3	B	1	2	
Construct 2, Item 1	X	1	2	3
Construct 2, Item 2	B	1	2	
Construct 2, Item 3	C		2	3
Construct 3, Item 1	X	1	2	3
Construct 3, Item 2	C		2	3
Construct 3, Item 3	A	1		3
Number of Items	9	7	7	7
% of Total Possible Items	100%	78%	78%	78%

Table 2

Decision Table of Sample Size Requirements for Model 1

Parameter Type	FIML			MI		
	C1 Only	C1 + C2	C3	C1 Only	C1 + C2	C3
λ	90	90	120	110	110	155
ϕ		90	120		115	155

Note. λ = Factor loadings; ϕ = Factor covariances. C1 = Convergence > 0.9; C2 = Parameter bias < 0.05 and SE bias < 0.1; C3 = mean SE bias < 0.05.

Table 3

Decision Table of Sample Size Requirements for Model 2

Parameter Type	FIML			MI		
	C1 Only	C1 + C2	C3	C1 Only	C1 + C2	C3
λ	80	80	100	90	90	200
ϕ_{WT}		80	100		115	200
ϕ_{BT}		80	100		115	200

Note. λ = Factor loadings; ϕ_{WT} = Within-time-point latent covariances; ϕ_{BT} = Between-time-point latent covariances. C1 = Convergence > 0.9; C2 = Parameter bias < 0.05 and SE bias < 0.1; C3 = mean SE bias < 0.05.

Table 4

Decision Table of Sample Size Requirements for Model 3

Parameter Type	FIML			MI		
	C1 Only	C1 + C2	C3	C1 Only	C1 + C2	C3
λ	130	130	160	160	160	> 300
β_{AR}		130	160		175	> 300
β_M		130	160		170	> 300

Note. λ = Factor loadings; β_{AR} = Autoregressive effects; β_M = Mediating effects. C1 = Convergence > 0.9; C2 = Parameter bias < 0.05, and SE bias < 0.1; C3 = mean SE bias < 0.05.

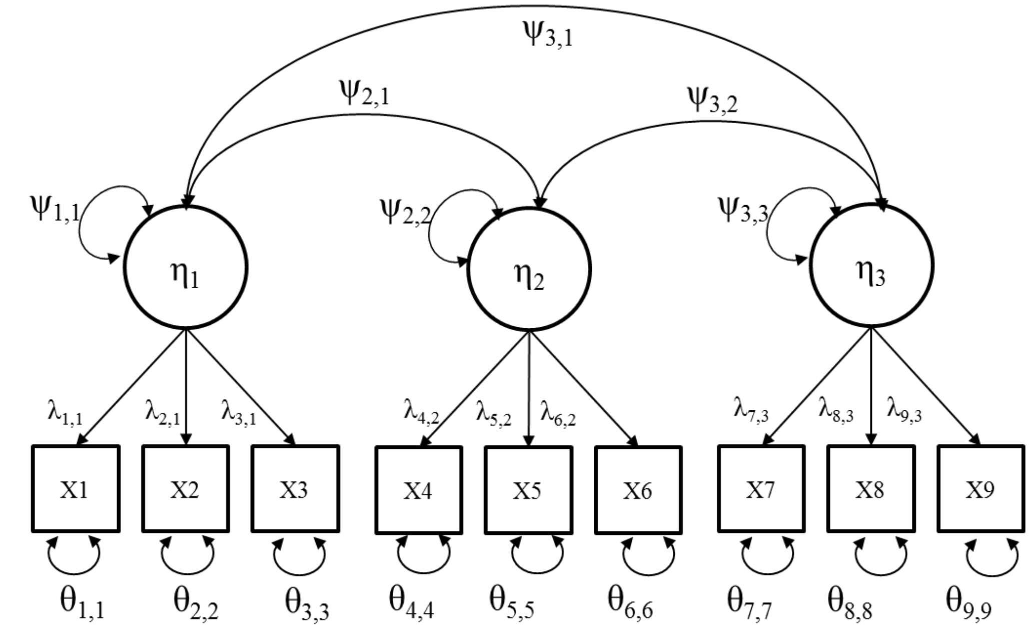

Figure 1. Model 1: Cross-sectional CFA model.

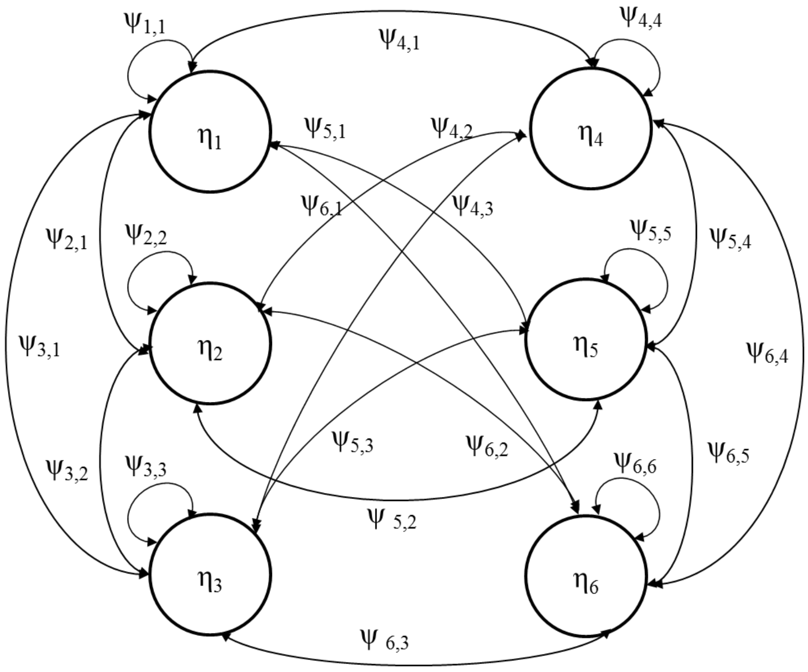

Figure 2. Model 2: Two-time-point saturated CFA model.

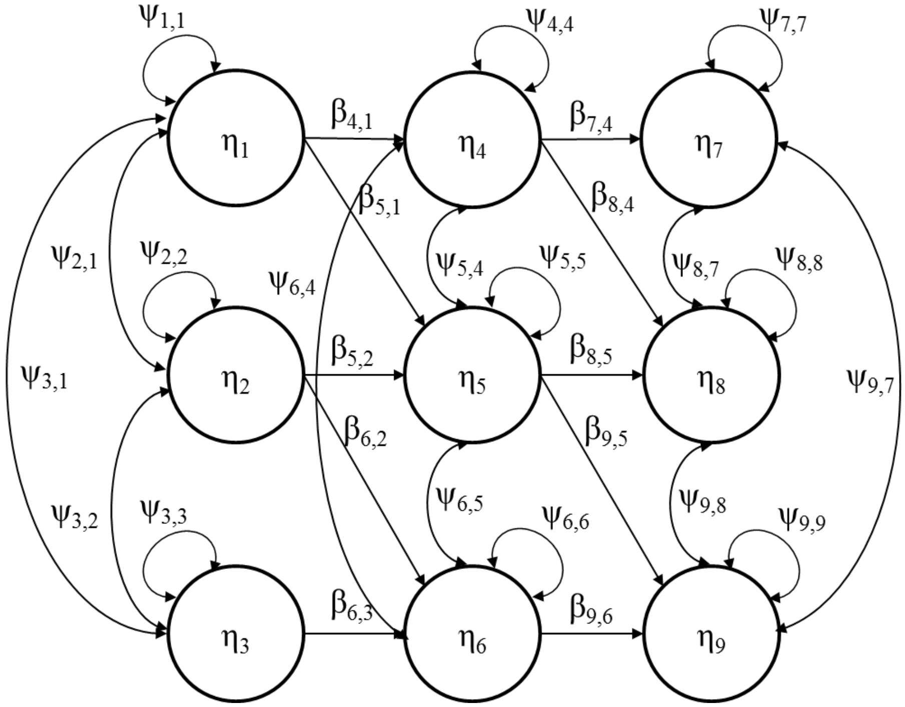

Figure 3. Model 3: Three-time-point mediation model

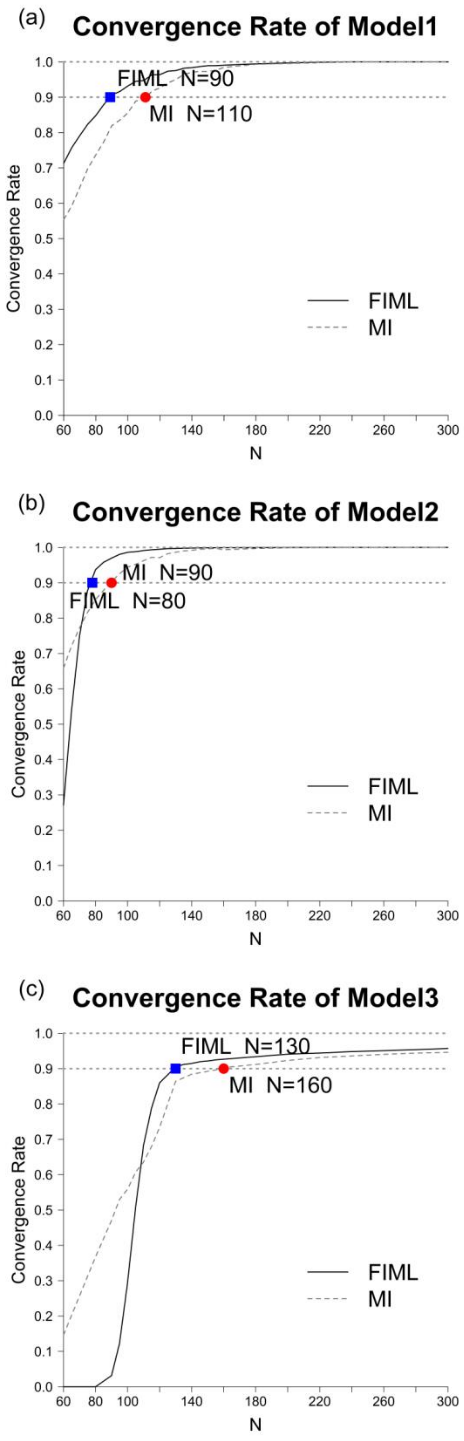

Figure 4. Convergence Rate

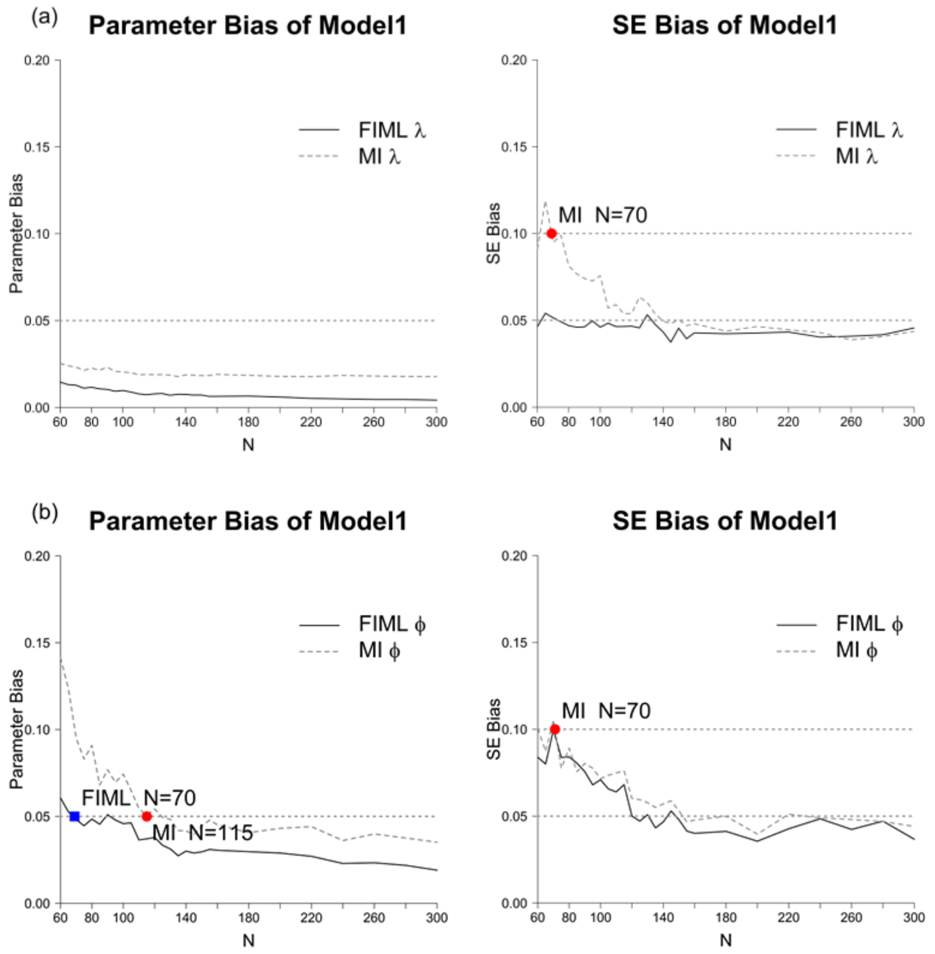

Figure 5. Parameter Bias and SE Bias for Model 1. λ and ϕ represent factor loadings and factor covariances, respectively.

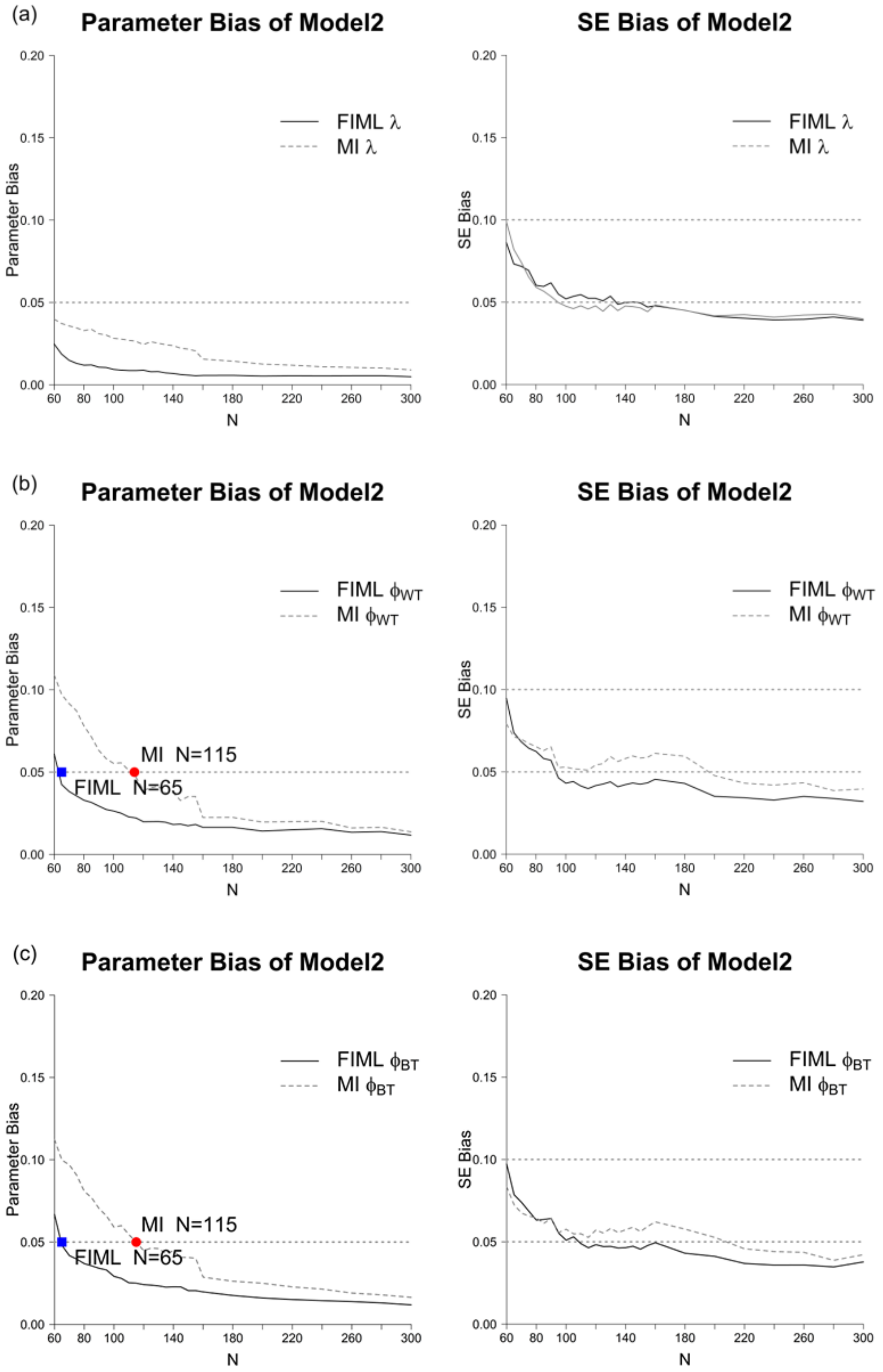

Figure 6. Parameter Bias and SE Bias for Model 2. λ, ϕ_{WT} and ϕ_{BT} represent factor loadings, within-time-point latent covariances, and between-time-point latent covariances, respectively.

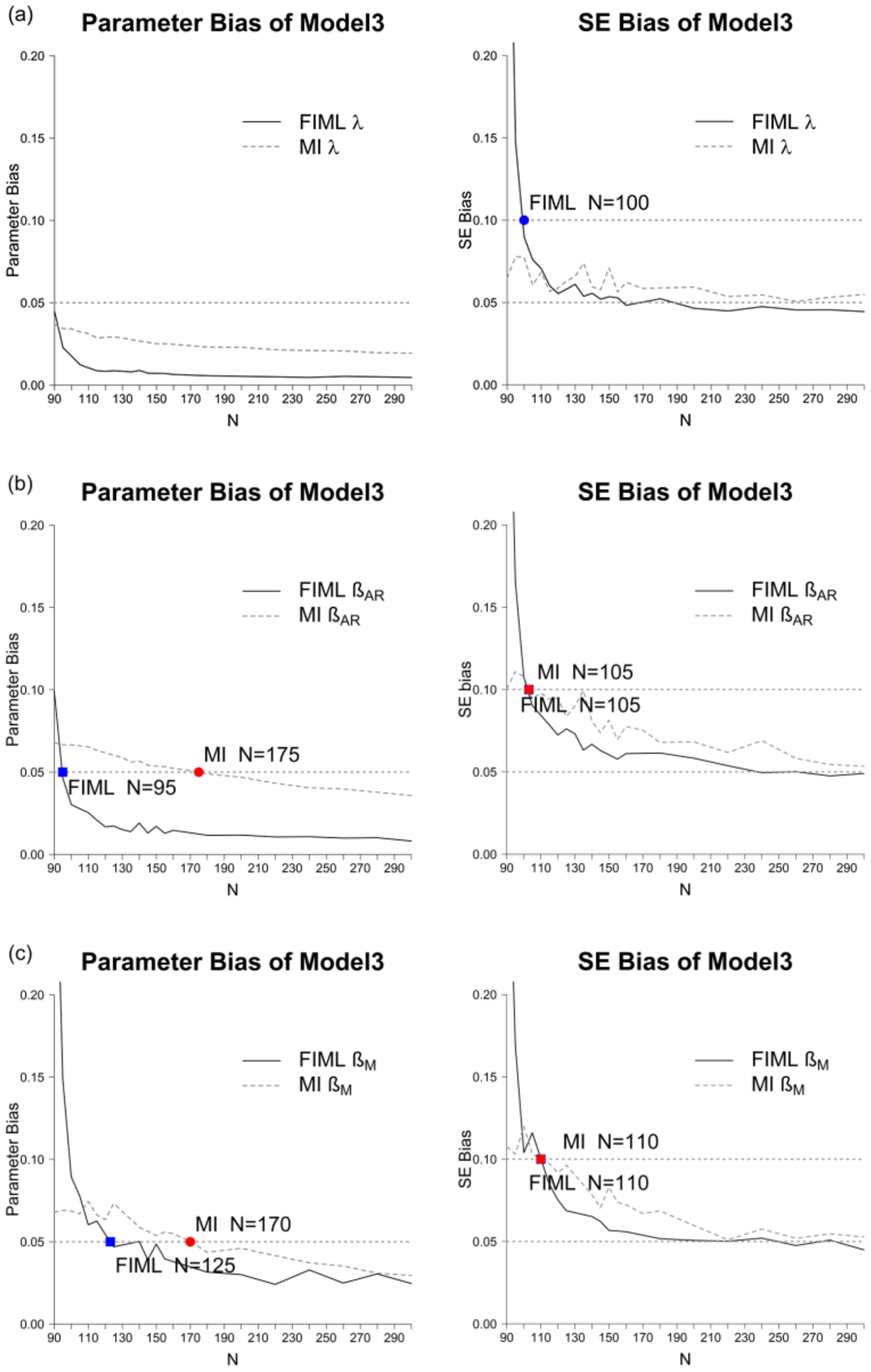

Figure 7. Parameter Bias and SE Bias for Model 3. λ, β_{AR} and β_M represent factor loadings, autoregressive effects and mediating effects, respectively.