Pooling Methods for Likelihood Ratio Tests in Multiply Imputed Data Sets

Simon Grund [1,2], Oliver Lüdtke [1,2], and Alexander Robitzsch [1,2]

[1] Leibniz Institute for Science and Mathematics Education

[2] Centre for International Student Assessment

Author Note

Simon Grund 0000-0002-1290-8986

Oliver Lüdtke 0000-0001-9744-3059

Alexander Robitzsch 0000-0002-8226-3132

All additional files concerning this article, including scripts, data, and the supplemental materials are available at: https://osf.io/u9s4k

Correspondence concerning this article should be addressed to Simon Grund, Leibniz Institute for Science and Mathematics Education, 24118 Kiel, Germany; (+49)-431-880-5653; grund@leibniz-ipn.de.

Abstract

Likelihood ratio tests (LRTs) are a popular tool for comparing statistical models. However, missing data are also common in empirical research, and multiple imputation (MI) is often used to deal with them. In multiply imputed data, there are multiple options for conducting LRTs, and new methods are still being proposed. In this article, we compare all available methods in multiple simulations covering applications in linear regression, generalized linear models, and structural equation modeling (SEM). In addition, we implemented these methods in an R package, and we illustrate its application in an example analysis concerned with the investigation of measurement invariance.

Keywords: missing data, multiple imputation, model comparison, likelihood ratio test

Pooling Methods for Likelihood Ratio Tests in Multiply Imputed Data Sets

Model comparison is one of the most common tasks in empirical research. By comparing models, researchers evaluate how well each model from a set of candidate models fits the observed data and whether restrictions placed on the model parameters change the fit of the model in a meaningful way. To this end, researchers often use multiparameter tests such as the Wald test, the likelihood ratio test (LRT), or the score test (for an overview, see Buse, 1982; Cox, 2006). In practice, model comparisons are often complicated by missing data, for example, due to unintended nonresponse or planned missingness (Graham et al., 2006; Rhemtulla et al., 2014; Rhemtulla & Hancock, 2016). Multiple imputation (MI) is one of the most common methods for dealing with missing data (Enders, 2010; Schafer & Yucel, 2002) as it is designed to replace missing values by creating multiple copies of the data, each with the missing values "filled in" with plausible replacements (Rubin, 1987). These data sets are then analyzed separately with conventional statistical methods, and the results are pooled to obtain a final set of parameter estimates and inferences.

The use of MI can be advantageous when there are multiple models of interest because it separates the treatment of missing data from the analysis. This ensures that the same assumptions about the missing data underlie the estimation and comparison of all models. On the other hand, the model comparison itself can become challenging when using MI because it requires the results to be pooled across the imputed data sets. Pooling methods for various types of analyses have been proposed in the missing data literature (for an overview, see Reiter and Raghunathan, 2007). The main focus of this article is on model comparisons with LRTs, but similar methods exist for Wald and score tests (Li, Raghunathan, et al., 1991; Mansolf et al., 2020). Several methods have been proposed for pooling LRTs (e.g., see Schafer, 1997), and new methods are still being developed (Chan & Meng, 2019). However, their properties are still poorly understood, and little is known about which methods are the most reliable for pooling LRTs in typical applications in psychology and related fields.

The present article is aimed at reducing this gap by evaluating the methods currently available for pooling LRTs in MI and by facilitating their application with an R package in which

all the methods are implemented. The article is structured as follows. First, we briefly review the statistical foundation of LRTs in complete data before we outline the methods that have been proposed so far for pooling LRTs in MI. In this context, we also discuss the theoretical and practical advantages and disadvantages of each method. Then, we present the results of three simulation studies in which we evaluated the performance of each method in applications typical for regression analysis, analyses with generalized linear models (GLMs), and structural equation modeling (SEM). Finally, we illustrate the application of theses methods with an example analysis using SEMs and the R package `mitml` (Grund et al., 2021).

Complete-data LRTs

In complete data sets, an LRT is performed by comparing two models with respect to the log-likelihood of the data, $\ell(\theta) = \log f(Y|\theta)$, given a set of model parameters θ. Specifically, the LRT is performed by calculating

$$d = -2\left[\ell(\hat{\theta}_0) - \ell(\hat{\theta})\right] \ , \tag{1}$$

where $\hat{\theta}$ denotes the estimated parameters under the full model, and $\hat{\theta}_0$ denotes the estimated parameters under a restricted (or *null*) model. The LRT statistic d (also known as the *deviance*) is typically compared with a χ^2 distribution with k degrees of freedom, where k is the number of restricted elements in θ_0 (for a general treatment, see Held & Bové, 2014).

Pooling methods for LRTs

In MI, each of the m imputed data sets is analyzed separately. This results in m sets of parameters for the full model, $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(m)}$, and the null model, $\hat{\theta}_0^{(1)}, \ldots, \hat{\theta}_0^{(m)}$. The parameter estimates themselves can easily be pooled simply by using Rubin's (1987) rules. By contrast, it is a lot straightforward to pool the LRT statistics $d^{(1)}, \ldots, d^{(m)}$, which are based on the log-likelihood of the imputed data under the full model, $\ell_1(\hat{\theta}^{(1)}), \ldots, \ell_m(\hat{\theta}^{(m)})$, and the null model, $\ell_1(\hat{\theta}_0^{(1)}), \ldots, \ell_m(\hat{\theta}_0^{(m)})$. In the following, we briefly describe the different methods that have been proposed for this task thus far.

Pooled χ^2 statistics (D_2)

Li, Meng, et al. (1991) proposed that LRTs be pooled by pooling the corresponding χ^2 statistics (or equivalently the p values) that are obtained from the LRTs in each of the imputed data sets. This method has been labeled D_2 in the missing data literature (Schafer, 1997). The D_2 statistic is calculated as

$$D_2 = \frac{\bar{d} - \frac{k(m-1)}{m+1} r_2}{k(1 + r_2)} \tag{2}$$

where $\bar{d}$ is the *average* LRT statistic across the imputed data sets, and r_2 is an estimate of the *average relative increase in variance* (ARIV) due to nonresponse:

$$r_2 = \left(1 + \frac{1}{m}\right) \left[\frac{1}{m-1} \sum_{l=1}^{m} \left(\sqrt{d^{(l)}} - \overline{\sqrt{d}}\right)^2\right] . \tag{3}$$

Broadly speaking, the ARIV represents the average increase in the sampling variance in the parameters being tested that is due to missing data (see also Enders, 2010). In other words, it provides an (inverse) measure of the fraction of missing information (FMI) with respect to the parameters that differ between the models that are being compared. D_2 requires that the FMI of the parameters that are being tested is approximately equal, so that their increases in variance can be meaningfully summarized by the ARIV. The D_2 statistic is typically compared with an F distribution that has k numerator and ν_2 denominator degrees of freedom:

$$\nu_2 = k^{-3/m}(m-1)(1 + r_2^{-1})^2 . \tag{4}$$

Figure 1 provides a visual illustration of D_2 as well as the other pooling methods that we are considering here. The main principle of D_2 is to evaluate the LRT statistics separately in each data set and to pool the results directly into an average value that is adjusted in accordance with the ARIV to reflect the added uncertainty that is due to missing data. As a result, calculating D_2 requires only the χ^2 statistics (or equivalently the p values) from the m imputed data sets.

Pooled LRTs (D_3)

To provide an improved method for pooling LRTs in multiply imputed data, Meng and

Rubin (1992) recommended the D_3 statistic. This statistic is calculated as

$$D_3 = \frac{\tilde{d}}{k(1+r_3)} \,, \tag{5}$$

where $\tilde{d}$ is the average LRT statistic evaluated at the *pooled* parameter estimates (e.g., with the pooled values obtained through Rubin's rules), and r_3 is another estimate of the ARIV:

$$r_3 = \frac{m+1}{k(m-1)}(\bar{d} - \tilde{d}) \,. \tag{6}$$

Like D_2, the D_3 statistic is based on the assumption that the FMI of the parameters being tested is approximately equal. D_3 can be compared with an F distribution with k numerator and ν_3 denominator degrees of freedom:

$$\nu_3 = \begin{cases} 4 + (t-4)\left[1 + (1-2t^{-1})r_3^{-1}\right]^2 & \text{if } t = k(m-1) > 4 \,, \\ t(1+k^{-1})(1+r_3^{-1})^2/2 & \text{otherwise} \end{cases} \,. \tag{7}$$

Figure 1 again illustrates the main principle of D_3. In contrast to D_2, the D_3 statistic requires the LRT statistic to be evaluated *twice* for each of the imputed data sets: once at the parameter estimates in that particular data set (i.e., $\theta^{(l)}$ and $\theta_0^{(l)}$ for the lth data set); and once with the parameters fixed to their pooled values (i.e., $\bar{\theta}$ and $\bar{\theta}_0$).

Pooled LRTs using "stacked" data (D_4)

Recently, Chan and Meng (2019) proposed yet another method for pooling LRTs that relies on the construction of a "stacked" data set, that is, a combined data set with the imputed data stacked on top of each other (for some earlier ad hoc methods based on stacked data, see also Lang and Little, 2014; Wood et al., 2008). In keeping with the previous literature, we refer to this method as D_4. Let $\ell_S(\theta_S) = m^{-1} \log f(Y_S|\theta_S)$ denote the log-likelihood of the stacked data. Then D_4 is calculated as

$$D_4 = \frac{\hat{d}_S}{k(1+r_4)} \,, \tag{8}$$

where $\hat{d}_S = -2\left[\ell_S(\hat{\theta}_{0,S}) - \ell_S(\hat{\theta}_S)\right]$ is the test statistic of the LRT obtained from the stacked data set, and r_4 is yet another estimate of the ARIV,

$$r_4 = \frac{m+1}{k(m-1)}(\bar{d} - \hat{d}_S) \ . \tag{9}$$

Because r_4 can take on negative values (similar to r_3), Chan and Meng (2019) recommended that the estimator of the ARIV in Equation 9 be replaced by $r_4^+ = \max(0, r_4)$ to ensure that the resulting expression is positive.[1] The D_4 statistic can be compared with an F distribution that has k numerator and ν_4 denominator degrees of freedom:

$$\nu_4 = k(m-1)\left[1 + (r_4^+)^{-1}\right]^2 \ . \tag{10}$$

Figure 1 provides an illustration of D_4. In contrast to D_3, the D_4 statistic relies on the stacked data set instead of pooled parameter estimates for pooling the LRT. As a result, calculating D_4 requires only the evaluation of the LRT statistic in the stacked data in addition to the m individual LRT statistics obtained from the imputed data sets.

Comparisons of Pooling Methods for LRTs

The pooling methods presented above have different advantages and disadvantages. From a statistical point of view, D_2 has been criticized because its sole reliance on χ^2 values constitutes an inefficient use of the information available about the models that are being compared (Li, Meng, et al., 1991; see also Enders, 2010; Schafer, 1999; van Buuren, 2018). Consequently, Li, Meng, et al. (1991) have found that D_2 can be either too liberal or too conservative depending on the number of parameters that are being tested and the FMI. Grund et al. (2016), Liu and Sriutaisuk (2020), and van Ginkel and Kroonenberg (2020) replicated these findings but also showed that D_2 can be fairly robust if the number of imputations is sufficiently large and the FMI is not too extreme. From a practical perspective, D_2 can be an attractive option because it is easy to compute, and the required χ^2 values are routinely provided by statistical software.

---

[1] In addition to the D_4 method as presented here, Chan and Meng (2019) also propose a "robust" version of D_4, which is based on a different (and more robust) estimate of the ARIV but requires the much stronger assumption that the relative increase in variance is equal for *all* parameters in the full model (i.e., not only the parameters that are being tested). For simplicity's sake, we focus on the non-robust version of D_4. However, interested readers can find additional details about and simulation results on the performance of the robust and non-robust versions of D_4 in Supplement A.

The D_3 statistic has often been recommended as a more reliable method for pooling LRTs in MI (Meng and Rubin, 1992; see also Enders, 2010; Schafer, 1999). Previous studies have found that D_3 usually works well but also that it tends to be slightly conservative and can suffer from low power, especially in conditions with large FMI (e.g., Grund et al., 2016; Liu & Enders, 2017; van Ginkel & Kroonenberg, 2020). However, D_3 has also been criticized because it is not invariant to the parameterization of the models that are being compared (e.g., Schafer, 1999). Specifically, because computing D_3 requires a re-evaluation of the likelihood function at the pooled values of the estimated parameters, its exact value depends on what parameterization is used to pool the parameter estimates across the imputed data sets (e.g., the residual variance in a regression model can be pooled by averaging the estimated variances or standard deviations). In addition, using D_3 in practice can be difficult because it requires full access to the likelihood function, and software that implements D_3 is still somewhat rare and often specific to certain types of models.[2]

Finally, D_4 has recently been proposed as another alternative with multiple improvements over D_3 (see Chan & Meng, 2019). Specifically, D_4 is invariant to the parameterization of the models that are being compared. In addition, Chan and Meng (2019) argued that the performance of D_4 should be at least as good as (or better than) D_3. From a practical point of view, D_4 is a promising alternative because it requires only the construction of a stacked data set that is then analyzed with standard methods without the need to re-evaluate the likelihood at user-defined values. As a result, D_4 can be implemented more easily, both in statistical software and by researchers who are proficient with statistical modeling environments such as R (R Core Team, 2020). However, there are currently no software packages that implement D_4 and no studies that have evaluated it, especially in comparison with other methods and in applications that are common in psychology and other fields.

In the following, we present the results of three simulation studies in which we compared the performance of these pooling methods in three different scenarios that represent common

---

[2] Software that implements D_3 for at least some types of models includes the R packages `mice` (van Buuren & Groothuis-Oudshoorn, 2011), `mitml` (Grund et al., 2021), and `semTools` (Jorgensen et al., 2020) as well as the commercial packages M*plus* (Asparouhov & Muthén, 2008), Stata (Medeiros, 2008), and SAS (Mistler, 2013).

applications of LRTs in psychology and the social sciences with varying complexity: linear regression (Study 1), logistic regression (Study 2), and structural equation modeling (SEM, Study 3).

Study 1: Linear Regression

Data generation

In Study 1, we were interested in applications of the LRT for comparing models in the context of regression analyses. The data were generated from a linear regression model with a standardized outcome variable y and k standardized, normally distributed predictor variables x_j ($j = 1, \ldots, k$). For person i ($i = 1, \ldots, n$),

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} + e_i \ , \quad e_i \sim N(0, \sigma^2) \ , \tag{11}$$

where β_j denotes the regression coefficient for the jth predictor variable, and σ^2 denotes the residual variance. The predictors were correlated according to some correlation ρ and had identical regression coefficients ($\beta_j = \beta$). The value of the regression coefficients was chosen in such a way that a certain amount of variance (R^2) was explained by the predictor variables (for a previous use of this design, see also Liu & Enders, 2017).

We then introduced missing values in the predictor variables x_j according to a missing completely at random (MCAR) mechanism (Rubin, 1987), where the probability of missing data is denoted by p. The missing values were generated in all predictors simultaneously such that there were only two patterns of missing data: one with all predictors observed and one with all predictors missing. This allowed us to investigate the effects of data that were unsystematically missing with a simple pattern, where the FMI in the parameters of interest was fully determined by the percentage of missing data.[3]

---

[3] The methods investigated in this paper can also be used when the data are missing at random (MAR). Under MAR, the missing data occur in a systematic manner, where the propensity of missing data is related to the observed data but unrelated to unobserved data once the observed data are taken into account (see Rubin, 1987). We did not investigate any systematic missing data mechanisms (e.g., MAR) because we did not expect that the differences between the procedures would differ meaningfully from those that we could observe under MCAR (see also Liu & Enders, 2017).

Simulated conditions

We varied the sample size ($n = 50, 100, 200, 500$), the number of predictor variables ($k = 2, 4, 6$), the proportion of variance explained by the predictor variables ($R^2 = 0, .02, .13, .26$), and the percentage of missing data ($p = 10\%, 30\%, 50\%$). Each condition was replicated between 1,000 and 3,000 times depending on the sample size.[4]

Treatment of missing data and analyses

The missing data were imputed with the R package `mice` (van Buuren & Groothuis-Oudshoorn, 2011). In line with current recommendations in the missing data literature (e.g., Bodner, 2008; Graham et al., 2007; von Hippel, 2020), we generated 100 imputed data sets, each after 20 iterations of the imputation algorithm. The model of interest was the regression model in Equation 11, and we used the LRT to compare this model with a restricted model that included only the intercept (i.e., no predictor variables). In this context, applying the LRT is equivalent to simultaneously testing all regression coefficients against zero. To implement the LRT for the imputed data, we used the methods outlined above:

D_2**:** pooled χ^2 values based on separate LRTs (Li, Meng, et al., 1991)
D_3**:** pooled LRT (Meng & Rubin, 1992)
D_4**:** pooled LRT based on "stacked" data (Chan & Meng, 2019)

To provide a means of comparison, we also applied the LRT to the complete data (CD) and after listwise deletion (LD), whereby all cases with missing data were removed from the analysis. All methods were evaluated in terms of their Type 1 error rates (at the $\alpha = 5\%$ level) in conditions with $R^2 = 0$ and their statistical power in conditions with nonzero effects ($R^2 > 0$). The computer code needed to run this and the other two simulations is provided on the OSF project page (https://osf.io/u9s4k).

[4] Conditions with small samples ($n = 50$) were replicated 3,000 times, those with moderate samples ($n = 100$) 2,000 times, those with large sample ($n = 200$) 1,500 times, and those with very large samples ($n = 500$) 1,000 times. As a result, the Monte Carlo SEs for the Type 1 error rates ranged from 0.4% to 0.7%, assuming that the actual Type 1 error rates were close to the nominal value of 5%.

Results

Type 1 error rates. The Type 1 error rates of all the procedures are summarized in Table 1. In conditions with only a few missing values (10%), all procedures had Type 1 error rates close to the nominal value of 5% in most cases. The only exceptions to this rule were with respect to CD, LD, and D_2, which showed a noticeable increase in Type 1 errors in conditions with small samples ($n = 50$) and many predictors ($k = 6$). Because these problems were also present in CD, this primarily reflected the behavior of the LRT in small samples. By contrast, D_3 and D_4 tended to show slightly lower rates of Type 1 errors in comparison with CD and maintained the nominal level throughout.

In conditions with large amounts of missing data (50%), the differences between the procedures were more pronounced. Specifically, with LD and D_2, the Type 1 error rates were often larger, reflecting more liberal statistical inferences, whereas those obtained with D_3 and D_4 were smaller, reflecting that these procedures were more conservative. These differences were especially large in conditions with small samples ($n = 50$) and many predictors ($k = 6$), where the procedures led to Type 1 error rates that were well above (13.6% for LD and 19.6% for D_2) or below (0% for D_3 and 0.2% for D_4) the nominal value. In conditions with larger samples, the Type 1 error rates obtained with LD, D_3, and D_4 again came close to the nominal value, even in these otherwise extreme conditions; however, those obtained with D_2 sometimes remained slightly too high.

Power. For the statistical power to detect nonzero effects, the results closely matched those for the Type 1 error rates and are summarized in Figure 2 for conditions with a medium effect size ($R^2 = .13$). Specifically, the power tended to be higher for LD and D_2, which reflected a more liberal behavior, and slightly lower for D_3 and D_4, which reflected a more conservative behavior. These differences were especially large in conditions with small samples ($n = 50$), many predictors ($k = 6$), and more missing data (50%), and tended to be smaller in less extreme conditions. Naturally, all procedures exhibited lower statistical power than CD. The overall pattern of results was similar for small ($R^2 = .02$) and large effects ($R^2 = .26$). However, the differences

between the procedures were less pronounced in these conditions because the statistical power was generally much lower or higher regardless of which method was used to conduct the LRT (for additional details, see Supplement C in the online supplemental materials).

Effects of the Number of Imputations

To investigate the potential impact of the number of imputations on model comparisons in MI, we repeated selected conditions from Study 1 with a different number of imputations. For this purpose, we chose a fixed sample size of $n = 100$ with a fixed percentage of missing data (30%) and varied the number of predictors ($k = 2, 4, 6$), the amount of variance explained ($R^2 = 0, .13$), and the number of imputations ($m = 3$, 5, 10, 20, 50, 100, 200, 500, 1,000). These conditions were replicated 2,000 times, and the results are summarized in Figure 3. The Type 1 error rates ($R^2 = 0$) were not strongly affected by the number of imputations but showed a slight tendency for D_2 to be more conservative with fewer imputations and for D_3 and D_4 to be more liberal then they were when the number of imputations was large. The power ($R^2 = .13$) tended to increase with more imputations, especially for D_2 and to a lesser extent also for D_3 and D_4. These difference were stronger when more parameters were tested. Overall, these results illustrate the importance of the number of imputations and suggest that at least 20 imputations should be used to obtain stable results.

Study 2: Logistic Regression

Data generation

In Study 2, we were interested in using the LRT to compare models in the context of generalized linear models (GLMs). To this end, we generated data from a logistic regression model with a binary outcome variable y and k standardized predictor variables x_j ($j = 1, \ldots, k$). For person i ($i = 1, \ldots, n$),

$$\log\left(\frac{P(y_i = 1|X = x_i)}{1 - P(y_i = 1|X = x_i)}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \ , \tag{12}$$

where all predictors were correlated in an identical manner as before. The coefficients were chosen

in such a way that the predictors explained a certain amount of variance (R^2) in the latent response propensity underlying the logistic regression model (i.e., by simulating the binary responses on the basis of a latent response propensity y_i^* with residuals that followed a logistic distribution with variance $\pi^2/3$; for details, see Agresti, 2007). Missing data were induced as before.

Simulated conditions, imputation, and analysis

We varied the sample size ($n = 50, 100, 200, 500$), the number of predictor variables ($k = 2, 4, 6$), the proportion of variance explained by the predictor variables ($R^2 = 0, .02, .13, .26$), and the percentage of missing data ($p = 10\%, 30\%, 50\%$). Each condition was replicated between 1,000 and 3,000 times as before.

The missing data were imputed in the same manner as before. The model of interest was the logistic regression model in Equation 12, and we again used an LRT to compare this model with a reduced model that included only an intercept. We used the same methods as before to pool the LRTs in the imputed data, and all methods were again evaluated with respect to Type 1 error rates and power.

Results

The results for the Type 1 error rates and the power are summarized in Table 2 and Figure 4, respectively. Similar to the previous study, all procedures provided acceptable Type 1 error rates except for some of the conditions with small samples ($n = 50$), many predictors ($k = 6$), and a large percentage of missing values (50%). In these conditions, LD and D_2 often showed elevated Type 1 error rates, whereas D_3 and D_4 often showed Type 1 error rates below the nominal value. As before, the Type 1 error rates obtained with LD, D_3, and D_4 approached the nominal value in conditions with larger samples, whereas those obtained with D_2 sometimes remained slightly too high. The results for the statistical power again followed the same pattern, whereby the power tended to be slightly higher for LD and D_2 and slightly lower for D_3 and D_4. These differences were most pronounced in conditions with small samples ($n = 50$), many predictors ($k = 6$), and a large percentage missing data (50%); in less extreme conditions, these

differences were still visible but generally smaller.

Study 3: SEM

Data generation

In Study 3, we were interested in using the LRT for model comparisons in the context of SEM. The data were generated from a two-factor confirmatory factor analysis (CFA) model with k items per factor and correlated residuals (or uniquenesses) between the corresponding items of each factor, as shown in Figure 5. This model is a common choice in applications of CFA in longitudinal research or in studies with multiple ratings of the same construct (e.g., self- vs. other ratings), where the assumption of independent unique factors is often not justified (e.g., Jöreskog, 1979; Little, 2013; Marsh, 1989). For identification, the factors were standardized with a mean of zero and a variance of one.

We inserted missing values into all the items that loaded on the second factor in accordance with an MCAR mechanism, where the probability of missing data is denoted by p. Similar to before, we induced the missing data in all the second-factor items simultaneously so that there were only two patterns of missing data: one in which all items were observed; and one in which the items of the first factor were observed and those of the second factor were missing.

Simulated conditions, imputation, and analysis

We varied the sample size (n = 100, 200, 500, 1,000), the number of items per factor (k = 4, 6, 8), the correlation between the residuals (ρ = 0, .10, .20), and the percentage of missing data (p = 10%, 30%, 50%). We fixed the correlation between the two factors to ϕ = .50. The factor loadings ($\mathbf{\Lambda}$) were chosen such that (a) the loadings for the corresponding items on each factor were equal and (b) their values were evenly distributed across the range of 0.70 to 0.95 as shown in Table 3. This corresponds to item reliabilities of roughly between .50 and .90. Each condition was replicated between 1,000 and 3,000 times as before.

Missing data were imputed in the same manner as before. To this end, we generated 100

imputations but increased the number of iterations per imputation to 40. The model of interest was the two-dimensional CFA model in Figure 5, and we used an LRT to compare this model with a restricted model in which the correlations between the residuals were fixed to zero. In this context, the LRT is used as a test for the presence or absence of correlated unique factors between corresponding items. Both models were estimated with standard maximum likelihood methods using the R package `lavaan` (Rosseel, 2012). To pool the LRT in the imputed data, we used the same methods as before. However, in addition to these methods, we also conducted the LRT after using full-information maximum likelihood (FIML), which is routinely used in SEMs to handle missing data (Arbuckle, 1996; Enders, 2010). This method was included primarily as an additional means of comparison.

Results

Type 1 error rates. The results for the Type 1 error rates are summarized in Table 4. Overall, all procedures showed acceptable Type 1 error rates except for some of the conditions in which the sample size was small (n = 100) or the amount of missing data was large (50%). In these conditions, the Type 1 error rates were sometimes too high for LD and FIML and too low for D_3 and D_4. In contrast to the previous studies, D_2 also showed Type 1 error rates below the nominal value in some of these conditions but did so much less often than D_3 and D_4. Similar to the previous studies, the Type 1 error rates obtained with LD, FIML, D_3, and D_4 came close to the nominal value as the sample size increased, even in these otherwise extreme scenarios. This was also true for D_2; however, the value approached by D_2 appeared to be slightly above the nominal value of 5%, albeit not very far.

Power. The results for the statistical power to detect a nonzero correlation between the residuals are summarized in Figure 6 for conditions with weakly correlated residuals (ρ = .10). In line with the results for the Type 1 error rates, the statistical power tended to be higher for LD and FIML than for D_3 and D_4. The power obtained with D_2 was usually lower than for LD and FIML but higher than for D_3 and D_4. However, in large samples (n = 500), D_2 usually showed the

highest statistical power. The differences between the procedures was most pronounced in conditions with the smallest sample sizes ($n = 100$), the largest number of items ($k = 8$), and the largest amount of missing data (50%). However, these differences generally tended to be smaller than those observed in the previous studies. Because conditions with a larger number of items also allowed for a more reliable measurement of the latent variables, the power was usually higher in conditions with many ($k = 8$) versus only a few ($k = 4$) items, regardless of which method was used to conduct the LRT.

Example Analysis

To illustrate the use of the pooling methods for LRTs, we used data from the EIKA study, a multi-cohort study conducted in Bremen, Germany, that investigated the psychological and socioeconomic antecedents of academic achievement in disadvantaged youths (EIKA, 2006). The data included self- and parent ratings of the Big Five personality traits for 943 ninth-grade students as well as information about their sex, cognitive ability, and reading and math proficiency. For simplicity, we focused only on extraversion. In this example, we aimed to investigate the extent to which measurement invariance held for the self- and parent ratings of extraversion (Meredith, 1993; Millsap & Olivera-Aguilar, 2012) and provide a step-by-step example for conducting LRTs in multiply imputed data. For simplicity, we focused on the steps involved in conducting the LRTs. The full computer code and the example data can be found in Supplement B and on the OSF project page (https://osf.io/u9s4k). Throughout the example, we used the R packages `mice` to impute the missing data (van Buuren & Groothuis-Oudshoorn, 2011), `lavaan` to fit the analysis models (Rosseel, 2012), and `mitml` to implement and pool the LRTs (Grund et al., 2021).

In the EIKA study, each personality trait was measured with a set of eight adjective pairs (e.g., "quiet" vs. "talkative") rated on a 5-point Likert scale, some of which were negatively worded and reverse-coded for analysis. The self- and parent ratings had between 19.1% and 33.5% missing data. The students' sex, cognitive ability, and reading and math proficiency were missing in 0.3% to 1.5% of the cases and were used as auxiliary variables (e.g., Collins et al., 2001). To assess measurement invariance, we compared the following specifications of a two-

dimensional confirmatory factor analysis (CFA) model: configural (form only, H_f), metric (loadings, $H_{\mathbf{\Lambda}}$), scalar (loadings and intercepts, $H_{\mathbf{\Lambda\tau}}$). All models included correlated residuals for the item pairs as well as correlated residuals between the negatively-worded items to account for possible method effects (see also Marsh, 1996). To set the scale of the latent variables, we set the factor means and variances to zero and one, respectively, in the configural model and relaxed them as needed in the metric and scalar invariance models (see Millsap & Olivera-Aguilar, 2012). All models were estimated with standard maximum likelihood procedures. Given the discrete nature of the item responses, this should be regarded as an approximation (see also DiStefano & Morgan, 2014; Lei & Shiverdecker, 2020).

Below, we provide the computer code needed to conduct the LRTs for the assessment of measurement invariance with the multiply imputed data. Given the list of imputed data sets, the first step is to fit the configural, metric, and scalar invariance models to the imputed data using the `with()` and the `cfa()` commands.

```
# fit configural, metric, and scalar invariance models
mi.configural <- with(eika.imp,
  cfa(model = configural, estimator = "ML", meanstructure = TRUE, std.lv = TRUE),
  include.data = TRUE)

mi.metric <- with(eika.imp,
  cfa(model = metric, estimator = "ML", meanstructure = TRUE, std.lv = TRUE),
  include.data = TRUE)

mi.scalar <- with(eika.imp,
  cfa(model = scalar, estimator = "ML", meanstructure = TRUE, std.lv = TRUE),
  include.data = TRUE)
```

Then, the `anova()` command can be used to compare the models with LRTs. This command supports the pooling of the LRTs by using D_2, D_3, or D_4. For example, to use D_4 to pool the LRTs, the command is as follows:

```
# compare models
anova(mi.configural, mi.metric, mi.scalar, method = "D4")

# Call:
```

```
#
# anova.mitml.result(object = mi.configural, mi.metric, mi.scalar,
#     method = "D4")
#
# Model comparison calculated from 100 imputed data sets.
# Combination method: D4
#
# Model 1: lavaan::lavaan(model = configural, ...)
# Model 2: lavaan::lavaan(model = metric, ...)
# Model 3: lavaan::lavaan(model = scalar, ...)
#
#             F.value      df1      df2    P(>F)      RIV
#    1 vs 2     1.046        7 8337.863    0.396    0.405
#    2 vs 3     3.868        7 6752.251    0.000    0.471
#
# Models were automatically ordered via 'logLik' (by decreasing complexity).
# Data for stacking were automatically extracted from the fitted models.
```

In addition to using D_4, we also conducted model comparisons with D_2 and D_3 as well as with FIML, for which we used the same auxiliary variables as in MI (see also Enders, 2008; Graham, 2003). The results are summarized in Table 5 and indicated that the data supported metric but not scalar invariance for the self- and parent ratings. On the basis of these findings, we also investigated whether the data supported partial invariance. To this end, we fitted a partial scalar invariance model that released the equality constraint that had been placed on the intercepts of Item 3 ($H_{\boldsymbol{\Lambda\tau}(3)}$). Using D_4:

```
# fit partial (scalar) invariance model
mi.partial.scalar <- with(eika.imp,
  cfa(model = partial.scalar, estimator = "ML", meanstructure = TRUE, std.lv = TRUE),
  include.data = TRUE)

# compare models for metric and partial scalar invariance
anova(mi.metric, mi.partial.scalar, method = "D4")

# Call:
#
# anova.mitml.result(object = mi.metric, mi.partial.scalar, method = "D4")
#
# Model comparison calculated from 100 imputed data sets.
# Combination method: D4
#
# Model 1: lavaan::lavaan(model = metric, ...)
```

```
# Model 2: lavaan::lavaan(model = partial.scalar, ...)
#
#             F.value      df1      df2    P(>F)      RIV
#    1 vs 2     1.105        6 6101.472    0.357    0.454
#
# Models were automatically ordered via 'logLik' (by decreasing complexity).
# Data for stacking were automatically extracted from the fitted models.
```

This indicated that scalar invariance held at least partially for all items except Item 3 (see Table 5). Consistent with our simulation results, we found that FIML tended to provide larger $\Delta\chi^2$ values than MI. However, the results were fairly consistent across the different methods, and all methods led to the same conclusions.

Discussion

The present article was concerned with pooling methods for LRTs in multiply imputed data sets. In writing this article, we had multiple goals: to provide a review of all methods currently available for pooling LRTs, to evaluate and compare these methods in the context of typical applications in psychology and other fields, and to facilitate their application with a user-friendly implementation in the R package `mitml` (Grund et al., 2021). To this end, we elaborated on the advantages and disadvantages of these methods, evaluated their performance in three simulation studies, and illustrated their use in a step-by-step example that was concerned with testing for measurement invariance in SEMs.

Overall, our findings indicate that all methods can provide reliable ways to pool LRTs in multiply imputed data but also that they can behave differently in more challenging situations (e.g., small samples, model comparisons that involve a large number of parameters, large amounts of missing data). Specifically, we found that D_2 can be too liberal, especially if the percentage of missing data is high, whereas D_3 and D_4 can sometimes be too conservative. For D_2 and D_3, these findings are consistent with the existing literature on the performance of these methods (e.g., Grund et al., 2016; Liu & Enders, 2017; van Ginkel & Kroonenberg, 2020). For the novel D_4 procedure, our findings suggest that the procedure can be considered to be at least as reliable as—if not more reliable than—the D_3 procedure. This is extremely encouraging because computing

D_4 requires only the construction of a "stacked" data set, which strongly simplifies the application of LRTs with multiply imputed data and should allow for much simpler implementations in statistical software that can cover a wide variety of statistical models. The `mitml` package (Grund et al., 2021) implements D_4—in addition to D_2 and D_3—through the generic `logLik()` interface, which can be used with many models estimated in R. This includes (generalized) linear models, multilevel models, and SEMs, among others.

Missing data are the norm rather than the exception in empirical research. They can be caused by the unintended dropout or nonresponse of study participants as well as planned missing data designs, which have become more and more popular in psychology and related fields (Graham et al., 2006; Rhemtulla et al., 2014; Rhemtulla & Hancock, 2016). In this article, we focused on MI in its traditional role as a tool for handling missing data. However, the methods considered here also have relevance in other applications. For example, imputed data also occur in large-scale studies, where MI is used to impute latent variable scores (Mislevy, 1991), in causal inference with the potential outcomes framework (Westreich et al., 2015), and in the analysis of synthetic data, where MI is used to overcome disclosure limitations and allow for reproducible analyses with sensitive data (Reiter, 2002; see also Raghunathan, in press).

As in all studies, this study has multiple limitations and suggests different avenues for additional research. In this article, we focused on model comparisons with LRTs. However, statistical models are also often compared in terms of goodness-of-fit indices, especially in SEMs (e.g., West et al., 2012). Previous research has shown that D_3 can be used to construct information criteria (Claeskens & Consentino, 2008) and fit indices in multiply imputed data (Asparouhov & Muthén, 2008; Enders & Mansolf, 2018) with properties similar to FIML. However, it has also been shown that computing fit indices for SEM is not straightforward even with FIML because the incomplete-data fit function implies different population values for popular fit indices such as the CFI and RMSEA (Lai, 2020; Zhang & Savalei, 2020). Future research should consider this problem further, including novel methods such as D_4 and possibly its robust version suggested by Chan and Meng (2019; for additional details, see Supplement A). Similarly, model comparisons sometimes cannot be conducted with standard LRTs. For example, when SEMs are estimated with

categorical or nonnormal data, it has been recommended that researchers use robust or limited-information procedures (DiStefano & Morgan, 2014; Lei & Shiverdecker, 2020) that require adjustments to the χ^2 difference tests in model comparisons (e.g., Satorra and Bentler, 1994; Yuan and Bentler, 2000; see also Shi et al., 2020). Liu and Sriutaisuk (2020) showed that D_2 can be used in this setting with acceptable statistical properties. Nonetheless, more research is needed on methods for pooling χ^2 difference tests in multiply imputed data as well as on how existing methods can be adjusted to accommodate categorical and nonnormal data. In this context, we hope that the present article provides a first step toward a more complete understanding of pooling methods for LRTs and similar procedures in multiply imputed data.

References

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Wiley.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques*. Erlbaum.

Asparouhov, T., & Muthén, B. O. (2008). *Chi-square statistics with multiple imputation* (Technical Appendix). http://www.statmodel.com/

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal*, *15*, 651–675. https://doi.org/10.1080/10705510802339072

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *The American Statistician*, *36*, 153–157.

Chan, K. W., & Meng, X.-L. (2019). *Multiple improvements of multiple imputation likelihood ratio tests*. arXiv: 1711.08822 [math,stat]. https://arxiv.org/abs/1711.08822

Claeskens, G., & Consentino, F. (2008). Variable selection with incomplete covariate data. *Biometrics*, *64*, 1062–1069. https://doi.org/10.1111/j.1541-0420.2008.01003.x

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330–351. https://doi.org/10.1037/1082-989X.6.4.330

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge University Press.

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 425–438. https://doi.org/10.1080/10705511.2014.915373

EIKA. (2006). *Projekt Entwicklung und Implementierung eines neuen Konzeptes zur Eingliederung Jugendlicher in die Berufs- und Arbeitswelt in Schulen mit erhöhtem Förderbedarf* (Unpublished data set). Humboldt-Universität zu Berlin, Friedrich-Alexander-Universität Erlangen Nürnberg. Berlin, Erlangen-Nürnberg.

Enders, C. K. (2008). A note on the use of missing auxiliary variables in full information maximum likelihood-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*, 434–448. https://doi.org/10.1080/10705510802154307

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Enders, C. K., & Mansolf, M. (2018). Assessing the fit of structural equation models with multiply imputed data. *Psychological Methods*, *23*, 76–93. https://doi.org/10.1037/met0000102

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*, 80–100. https://doi.org/10.1207/S15328007SEM1001_4

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*, 206–213. https://doi.org/10.1007/s11121-007-0070-9

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*, 323–343. https://doi.org/10.1037/1082-989X.11.4.323

Grund, S., Lüdtke, O., & Robitzsch, A. (2016). Pooling ANOVA results from multiply imputed datasets: A simulation study. *Methodology*, *12*, 75–88. https://doi.org/10.1027/1614-2241/a000111

Grund, S., Robitzsch, A., & Lüdtke, O. (2021). *Mitml: Tools for multiple imputation in multilevel modeling (Version 0.4-0)*. http://CRAN.R-project.org/package=mitml

Held, L., & Bové, D. S. (2014). *Applied Statistical Inference: Likelihood and Bayes*. Springer. https://doi.org/10.1007/978-3-642-37887-4

Jöreskog, K. G. (1979). Statistical models and methods for analysis of longitudinal data. In K. G. Jöreskog, D. Sörbom, & J. Magidson (Eds.), *Advances in factor analysis and structural equation models* (pp. 129–170). Abt books.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., Rosseel, Y., Miller, P., Quick, C., Garnier-Villarreal, M., Selig, J., Boulton, A., Preacher, K., Coffman, D., Rhemtulla, M.,

Robitzsch, A., Enders, C., Arslan, R., Clinton, B., Panko, P., Merkle, E., Chesnut, S., . . . Ben-Shachar, M. S. (2020, May 27). *semTools: Useful tools for structural equation modeling* (Version 0.5-3). https://CRAN.R-project.org/package=semTools

Lai, K. (2020). Correct Estimation Methods for RMSEA Under Missing Data. *Structural Equation Modeling: A Multidisciplinary Journal*, *0*(0), 1–12. https://doi.org/10.1080/10705511.2020.1755864

Lang, K. M., & Little, T. D. (2014). The supermatrix technique: A simple framework for hypothesis testing with missing data. *International Journal of Behavioral Development*, *38*(5), 461–470. https://doi.org/10.1177/0165025413514326

Lei, P.-W., & Shiverdecker, L. K. (2020). Performance of Estimators for Confirmatory Factor Analysis of Ordinal Variables with Missing Data. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 584–601. https://doi.org/10.1080/10705511.2019.1680292

Li, K. H., Meng, X.-L., Raghunathan, T. E., & Rubin, D. B. (1991). Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica*, *1*, 65–92. http://www.stat.sinica.edu.tw/statistica/

Li, K. H., Raghunathan, T. E., & Rubin, D. B. (1991). Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, *86*, 1065–1073. https://doi.org/10.1080/01621459.1991.10475152

Little, T. D. (2013). *Longitudinal structural equation modeling*. The Guilford Press.

Liu, Y., & Enders, C. K. (2017). Evaluation of multi-parameter test statistics for multiple imputation. *Multivariate Behavioral Research*, *52*, 371–390. https://doi.org/10.1080/00273171.2017.1298432

Liu, Y., & Sriutaisuk, S. (2020). Evaluation of model fit in structural equation models with ordinal missing data: An examination of the D_2 method. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 561–583. https://doi.org/10.1080/10705511.2019.1662307

Mansolf, M., Jorgensen, T. D., & Enders, C. K. (2020). A multiple imputation score test for model

modification in structural equation models. *Psychological Methods*, *25*(4), 393–411. https://doi.org/10.1037/met0000243

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement*, *13*, 335–361. https://doi.org/10.1177/014662168901300402

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*, 810–819. https://doi.org/10.1037/0022-3514.70.4.810

Medeiros, R. (2008, November 16). *Likelihood ratio tests for multiply imputed datasets: Introducing milrtest* (Presentation). Fall North American Stata Users Group Meeting, San Francisco, CA, United States. https://www.stata.com/meeting/proceedings/

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

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543. https://doi.org/10.1007/BF02294825

Millsap, R. E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. *Handbook of structural equation modeling*. Guilford Press.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177–196. https://doi.org/10.1007/BF02294457

Mistler, S. A. (2013). A SAS macro for computing pooled likelihood ratio tests with multiply imputed data. *Proceedings of the SAS Global Forum*. http://support.sas.com/

R Core Team. (2020). *R: A language and environment for statistical computing (Version 4.0)*. Vienna, Austria. https://www.R-project.org/

Raghunathan, T. E. (in press). Synthetic data. *Annual Review of Statistics and Its Application*. https://doi.org/10.1146/annurev-statistics-040720-031848

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, *18*, 531–543.

Reiter, J. P., & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, *102*, 1462–1471. https://doi.org/10.1198/016214507000000932

Rhemtulla, M., Jia, F., Wu, W., & Little, T. D. (2014). Planned missing designs to optimize the efficiency of latent growth parameter estimates. *International Journal of Behavioral Development*, *38*, 423–434. https://doi.org/10.1177/0165025413514324

Rhemtulla, M., & Hancock, G. R. (2016). Planned missing data designs in educational psychology research. *Educational Psychologist*, *51*, 305–316. https://doi.org/10.1080/00461520.2016.1208094

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Sage.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC Press.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*, 3–15. https://doi.org/10.1177/096228029900800102

Schafer, J. L., & Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, *11*, 437–457. https://doi.org/10.1198/106186002760180608

Shi, D., Maydeu-Olivares, A., & Rosseel, Y. (2020). Assessing Fit in Ordinal Factor Analysis Models: SRMR vs. RMSEA. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 1–15. https://doi.org/10.1080/10705511.2019.1611434

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). CRC Press.

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

https://doi.org/10.18637/jss.v045.i03

van Ginkel, J. R., & Kroonenberg, P. M. (2020). Much ado about nothing: Multiple imputation to balance unbalanced designs for two-way analysis of variance. *Methodology*, *16*(4), 335–353. https://doi.org/10.5964/meth.4327

von Hippel, P. T. (2020). How many imputations do you need? A two-stage calculation using a quadratic rule. *Sociological Methods & Research*, *49*, 699–718. https://doi.org/10.1177/0049124117747303

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling*. Guilford Press.

Westreich, D., Edwards, J. K., Cole, S. R., Platt, R. W., Mumford, S. L., & Schisterman, E. F. (2015). Imputation approaches for potential outcomes in causal inference. *International Journal of Epidemiology*, *44*, 1731–1737. https://doi.org/10.1093/ije/dyv135

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. https://doi.org/10.1002/sim.3177

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*, 165–200. https://doi.org/10.1111/0081-1750.00078

Zhang, X., & Savalei, V. (2020). Examining the effect of missing data on RMSEA and CFI under normal theory full-information maximum likelihood. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 219–239. https://doi.org/10.1080/10705511.2019.1642111

Table 1
Study 1: Type 1 Error Rates in % ($R^2 = 0$)

					MI		
MD	k	n	CD	LD	D_2	D_3	D_4
10%	2	50	6.5	6.4	6.5	5.8	5.9
		100	5.6	6.1	6.1	5.9	5.9
		200	5.9	5.5	5.5	5.4	5.4
		500	7.2	7.0	7.2	6.8	6.8
	4	50	6.1	7.0	7.3	5.4	5.4
		100	6.2	6.3	6.2	5.5	5.5
		200	5.5	5.2	5.3	5.1	5.1
		500	4.5	4.4	4.4	4.5	4.5
	6	50	**8.0**	**8.3**	**8.7**	6.0	6.0
		100	6.7	7.4	**7.6**	6.6	6.7
		200	5.1	6.3	6.0	5.5	5.5
		500	5.6	5.1	5.3	5.1	5.1
30%	2	50	6.0	7.0	6.6	4.3	4.4
		100	5.1	5.3	5.5	4.0	4.2
		200	5.1	5.7	6.2	5.3	5.3
		500	3.8	4.6	4.8	4.2	4.2
	4	50	7.3	**7.9**	**8.4**	3.0	3.1
		100	5.3	5.7	6.6	4.3	4.4
		200	5.5	4.4	5.0	4.0	4.1
		500	5.2	4.8	5.8	4.8	4.9
	6	50	**8.1**	**9.0**	**11.4**	**1.9**	**2.3**
		100	6.6	7.0	**8.8**	4.4	4.7
		200	5.5	5.3	6.7	4.0	4.1
		500	5.4	4.6	5.9	4.3	4.3
50%	2	50	6.1	7.4	**8.3**	2.8	3.0
		100	5.7	6.3	7.0	4.2	4.2
		200	4.6	5.3	6.0	4.6	4.6
		500	5.3	5.4	7.0	5.2	5.2
	4	50	6.8	**10.1**	**13.5**	**1.0**	**1.4**
		100	6.2	6.9	**9.7**	2.8	3.0
		200	5.0	5.7	**8.3**	4.2	4.3
		500	5.0	5.8	6.3	4.8	4.9
	6	50	**8.7**	**13.6**	**19.6**	**0.0**	**0.2**
		100	5.8	**8.2**	**13.2**	**2.2**	2.8
		200	5.9	5.8	**9.4**	3.7	3.8
		500	5.9	5.8	**9.4**	4.9	5.0

Note. Bold numbers refer to Type 1 error rates outside the [2.5%; 7.5%] interval. k = number of predictors; n = sample size; CD = complete data; LD = listwise deletion; D_2, D_3, D_4 = pooling methods.

Table 2
Study 2: Type 1 Error Rates in % ($R^2 = 0$)

					MI		
MD	k	n	CD	LD	D_2	D_3	D_4
10%	2	50	6.6	6.7	6.7	6.1	6.1
		100	5.0	5.1	5.1	4.9	4.9
		200	4.4	4.1	4.3	4.2	4.2
		500	5.8	5.4	5.3	5.3	5.3
	4	50	6.8	**7.6**	**7.7**	6.4	6.4
		100	5.0	4.9	5.1	4.1	4.1
		200	6.3	6.7	6.9	6.5	6.5
		500	5.6	5.6	5.7	5.4	5.4
	6	50	**8.6**	**9.0**	**9.1**	6.4	6.6
		100	5.8	6.3	6.3	5.3	5.3
		200	4.9	5.0	5.5	4.8	4.8
		500	4.9	5.1	5.0	4.8	4.8
30%	2	50	6.0	5.9	6.2	3.9	4.1
		100	5.9	5.5	5.9	5.0	5.0
		200	4.9	5.1	5.4	4.9	4.9
		500	4.4	5.1	5.5	4.9	5.0
	4	50	6.7	**8.0**	**8.7**	3.3	3.6
		100	6.2	6.7	6.9	4.8	5.0
		200	4.9	6.3	6.9	5.4	5.4
		500	4.7	5.7	6.3	5.5	5.5
	6	50	7.4	**10.7**	**11.9**	**1.7**	**2.4**
		100	6.0	6.8	**7.6**	4.0	4.2
		200	5.6	5.4	6.4	4.5	4.6
		500	5.3	5.7	6.4	5.3	5.3
50%	2	50	6.0	**7.6**	**8.0**	2.5	3.0
		100	4.1	5.5	6.7	3.3	3.4
		200	5.5	5.9	6.2	4.8	4.9
		500	3.8	5.2	5.6	4.8	4.9
	4	50	6.3	**10.0**	**12.6**	**0.6**	**1.6**
		100	5.7	6.4	**8.2**	2.8	3.0
		200	5.3	5.1	**7.7**	3.5	3.6
		500	5.8	5.8	7.2	5.4	5.4
	6	50	**7.6**	**14.8**	**17.5**	**0.0**	**0.3**
		100	6.3	**7.9**	**12.6**	**2.1**	2.5
		200	5.3	6.2	**9.7**	3.6	3.7
		500	4.6	5.3	**9.0**	5.0	5.1

Note. Bold numbers refer to Type 1 error rates outside the [2.5%; 7.5%] interval. k = number of predictors; n = sample size; CD = complete data; LD = listwise deletion; D_2, D_3, D_4 = pooling methods.

Table 3
Specification of Factor Loadings $\mathbf{\Lambda} = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$ *in Study 3*

	$k = 4$		$k = 6$		$k = 8$	
	$\boldsymbol{\lambda}_1$	$\boldsymbol{\lambda}_2$	$\boldsymbol{\lambda}_1$	$\boldsymbol{\lambda}_2$	$\boldsymbol{\lambda}_1$	$\boldsymbol{\lambda}_2$
y_{11}	0.70	0	0.70	0	0.75	0
y_{12}	0.78	0	0.75	0	0.78	0
y_{13}	0.87	0	0.80	0	0.81	0
y_{14}	0.95	0	0.85	0	0.84	0
y_{15}			0.90	0	0.86	0
y_{16}			0.95	0	0.89	0
y_{17}					0.92	0
y_{18}					0.95	0
y_{21}	0	0.70	0	0.70	0	0.75
y_{22}	0	0.78	0	0.75	0	0.78
y_{23}	0	0.87	0	0.80	0	0.81
y_{24}	0	0.95	0	0.85	0	0.84
y_{25}			0	0.90	0	0.86
y_{26}			0	0.95	0	0.89
y_{27}					0	0.92
y_{28}					0	0.95

Note. k = number of items per factor.

Table 4
Study 3: Type 1 Error Rates in % (ρ = 0)

						MI		
MD	*k*	*n*	CD	LD	FIML	D_2	D_3	D_4
10%	4	100	6.6	6.6	6.7	6.5	6.0	5.9
		200	6.4	6.2	6.3	6.1	5.8	5.7
		500	4.3	4.3	4.3	4.3	4.2	4.2
		1000	4.3	5.6	5.1	5.3	5.6	5.6
	6	100	6.0	5.9	6.0	5.3	4.8	4.6
		200	6.2	6.3	6.4	5.8	5.7	5.7
		500	4.1	5.7	5.7	5.6	5.5	5.4
		1000	5.7	4.9	4.9	5.1	4.9	4.9
	8	100	6.2	6.7	6.6	5.2	4.7	4.6
		200	5.9	6.2	6.1	5.3	5.1	5.1
		500	5.3	4.7	4.6	4.3	4.4	4.4
		1000	4.8	5.3	5.2	4.8	5.1	5.1
30%	4	100	6.4	6.7	6.6	5.8	4.2	4.1
		200	6.5	5.9	5.8	5.5	5.0	4.9
		500	6.2	5.5	5.5	6.0	5.0	4.9
		1000	6.1	5.8	5.9	5.6	6.0	6.0
	6	100	6.2	7.1	7.0	5.1	2.9	2.6
		200	6.2	6.2	6.0	5.8	4.3	4.2
		500	6.0	6.2	6.1	6.3	4.9	4.8
		1000	5.1	5.0	5.0	6.2	4.9	4.9
	8	100	7.0	7.0	7.2	3.5	**1.8**	**1.7**
		200	6.5	6.5	6.5	5.4	3.8	3.6
		500	5.1	5.7	5.6	5.7	5.0	5.0
		1000	5.1	5.6	5.7	5.6	4.8	4.8
50%	4	100	6.1	**7.7**	7.1	6.6	**2.2**	**2.0**
		200	6.2	7.5	7.0	7.5	4.5	4.3
		500	5.5	5.3	5.2	6.6	4.4	4.4
		1000	3.8	5.5	5.6	6.6	5.3	5.3
	6	100	5.5	7.3	7.0	4.6	**0.8**	**0.7**
		200	6.0	6.0	6.2	5.8	2.8	2.6
		500	5.0	5.3	5.2	6.1	4.0	4.0
		1000	5.0	5.6	5.8	6.8	5.4	5.4
	8	100	6.8	**8.7**	**8.2**	2.7	**0.2**	**0.2**
		200	5.5	7.0	6.7	5.5	**1.7**	**1.6**
		500	5.3	5.5	5.6	6.7	3.6	3.5
		1000	4.7	5.1	5.1	6.9	3.9	3.9

Note. Bold numbers refer to Type 1 error rates outside the [2.5%; 7.5%] interval. *k* = number of predictors; *n* = sample size; CD = complete data; LD = listwise deletion; FIML = full-information maximum likelihood; D_2, D_3, D_4 = pooling methods.

Table 5

Results for the Assessment of Measurement Invariance for Self vs. Parent Ratings of Students' Extraversion in the Example Analyses

			FIML		D_2		D_3		D_4	
Model	Comparison	Δdf	$\Delta\chi^2$	p	$\Delta\chi^2$	p	$\Delta\chi^2$	p	$\Delta\chi^2$	p
Metric	H_f vs. $H_{\boldsymbol{\Lambda}}$	7	7.8	.351	6.5	.480	7.3	.397	7.3	.396
Scalar	$H_{\boldsymbol{\Lambda}}$ vs. $H_{\boldsymbol{\Lambda\tau}}$	7	28.0	.000	27.3	.000	27.1	.000	27.1	.000
Partial scalar	$H_{\boldsymbol{\Lambda}}$ vs. $H_{\boldsymbol{\Lambda\tau}(3)}$	6	6.6	.355	7.0	.318	6.6	.357	6.6	.357

Note. Numbers in parentheses indicate that the respective item parameters were unconstrained in the partial invariance model. The $\Delta\chi^2$ value for D_2, D_3, and D_4 was calculated as $kF_{k,\nu}$. $\boldsymbol{\Lambda}$ = loadings; $\boldsymbol{\tau}$ = intercepts; FIML = full-information maximum likelihood; D_2, D_3, D_4 = pooling methods.

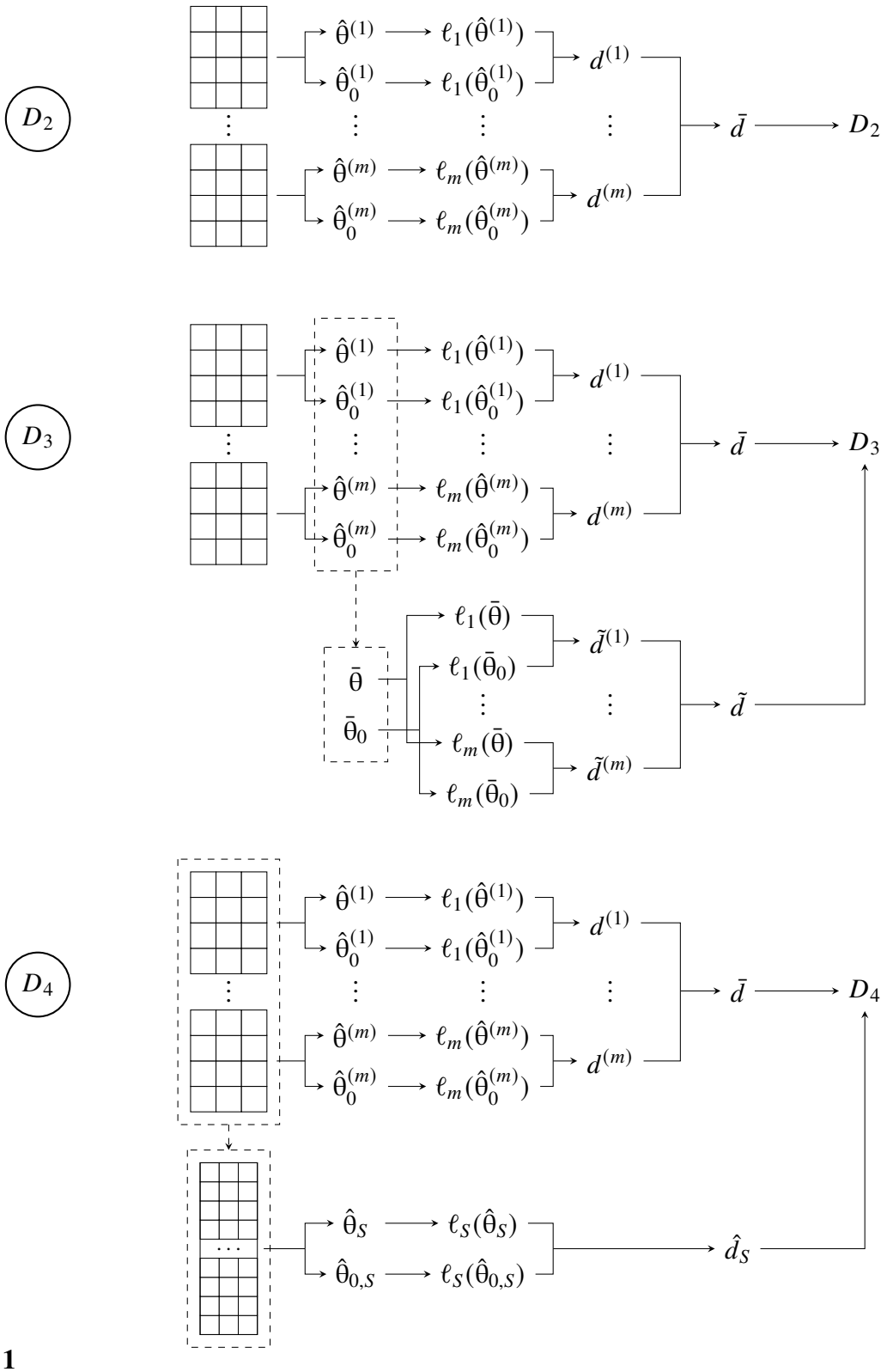

Figure 1

Schematic representation of pooling methods for LRTs under MI. The D_2 method is based only on the χ^2 statistics obtained from the multiply imputed data, whereas D_3 also requires a reevaluation of the log-likelihood at the pooled parameter estimates, and D_4 requires an evaluation of the log-likelihood in a "stacked" data set.

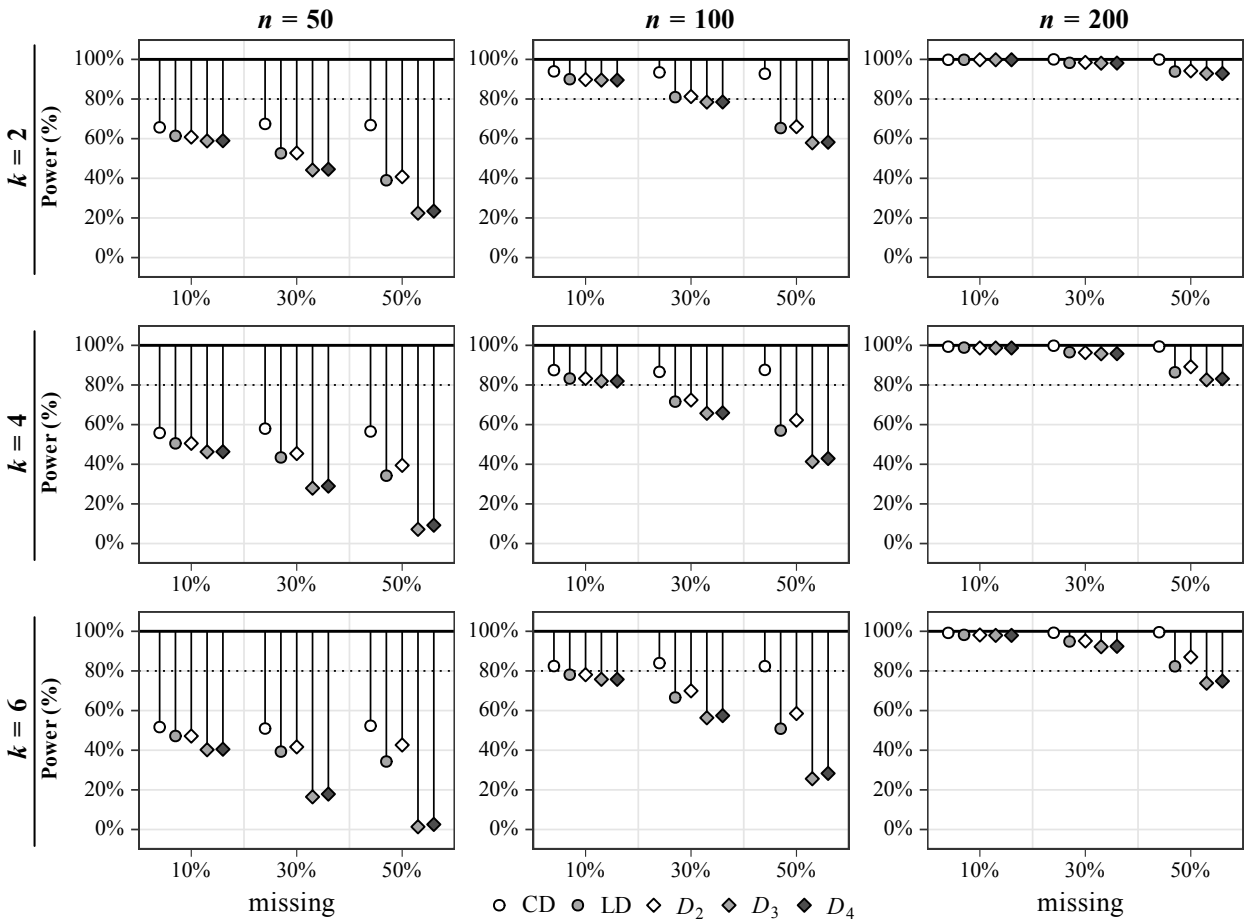

Figure 2

Power (in %) in conditions with medium effect size ($R^2 = .13$) in Study 1. n = sample size; k = number of predictors; CD = complete data; LD = listwise deletion; D_2, D_3, D_4 = pooling methods.

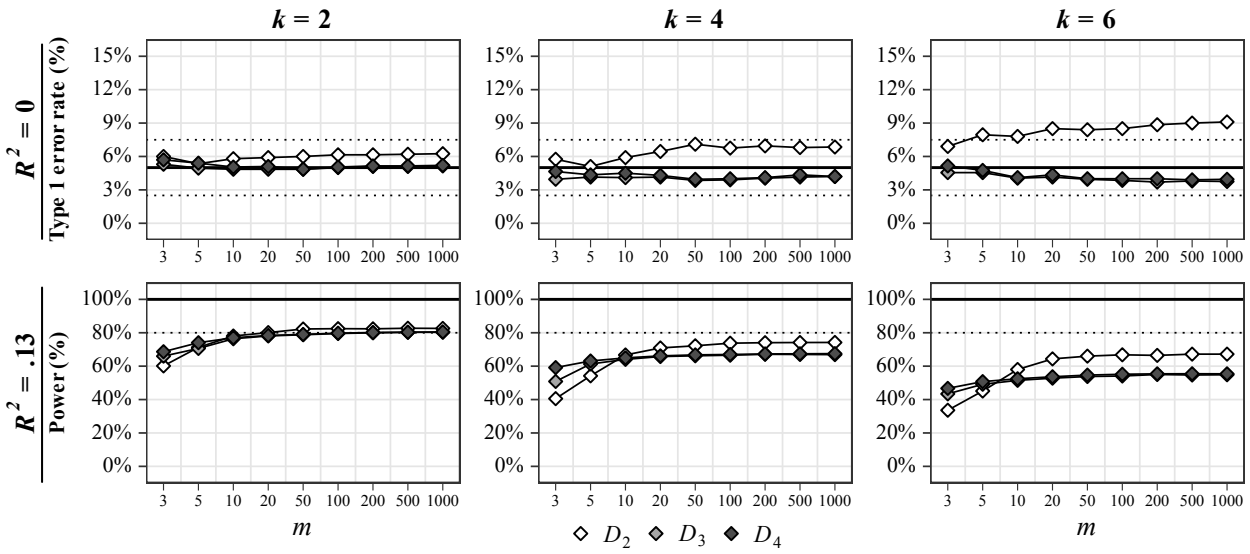

Figure 3

Type 1 error rates and power (in %) in conditions with moderate sample size (n = 100) depending on the number of imputations (m) in the context of Study 1. k = number of predictors; D_2, D_3, D_4 = pooling methods.

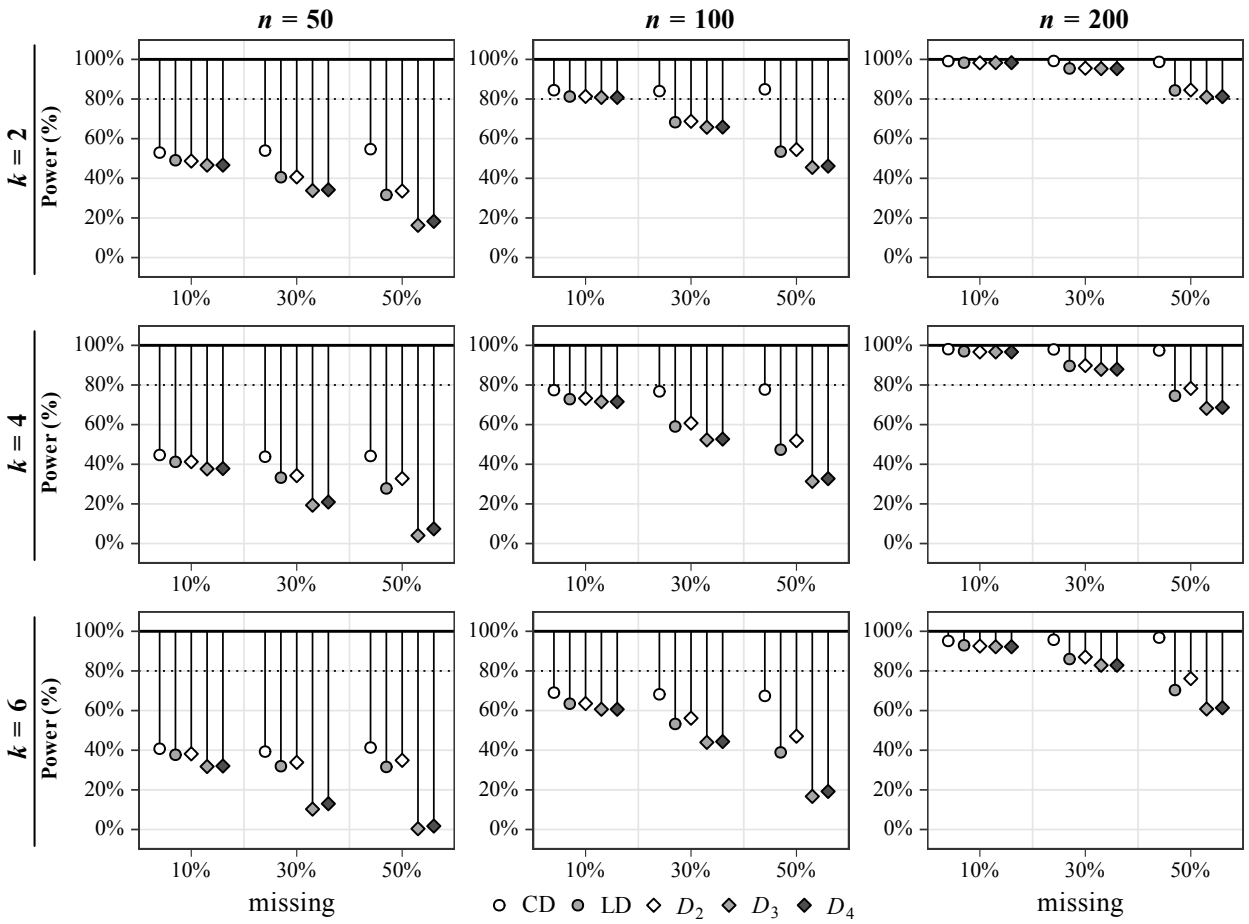

Figure 4

Power (in %) in conditions with medium effect size ($R^2 = .13$) in Study 2. n = sample size; k = number of predictors; CD = complete data; LD = listwise deletion; D_2, D_3, D_4 = pooling methods.

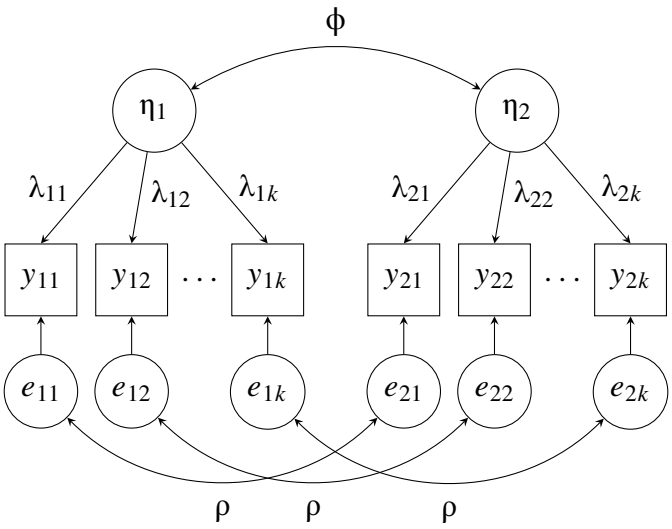

Figure 5
Data generating model in Study 3.

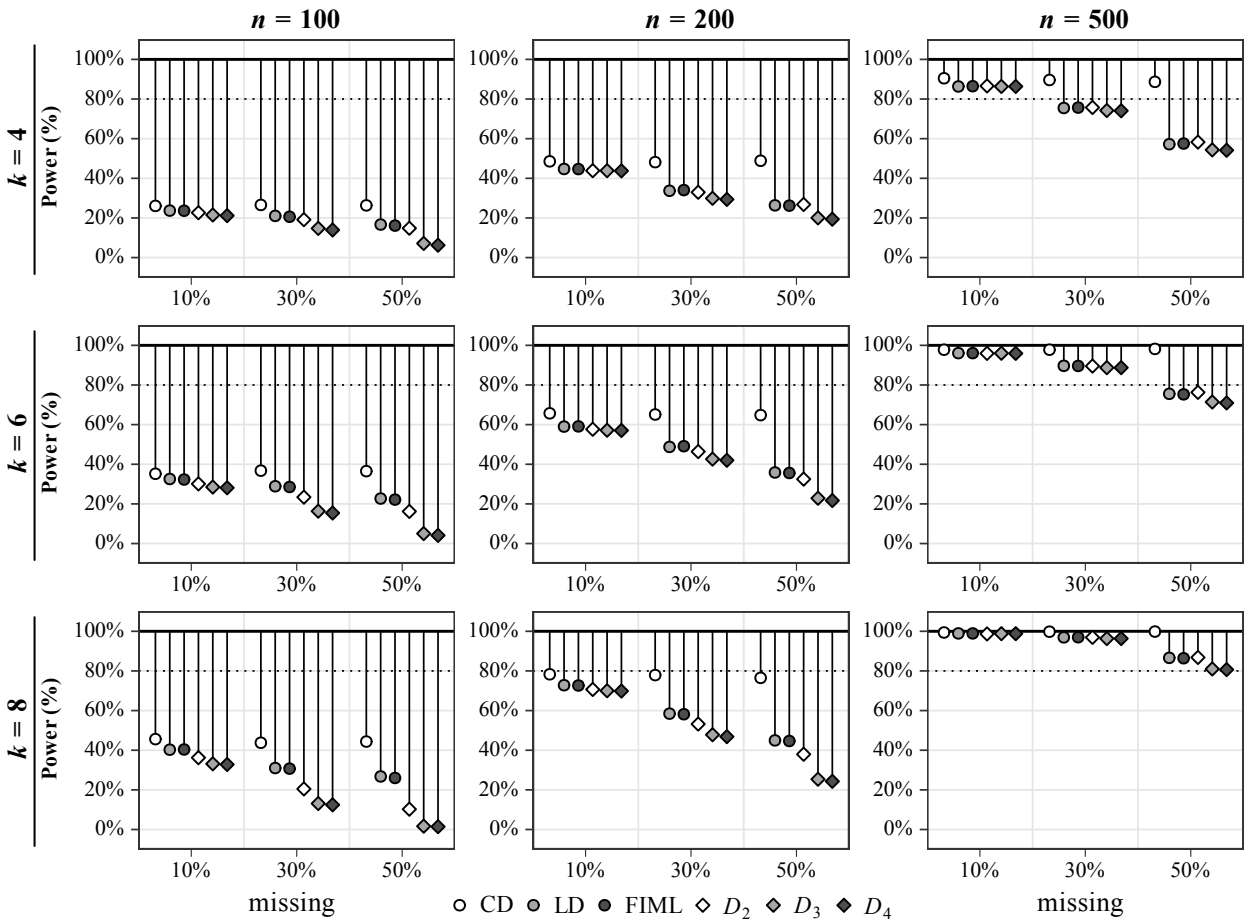

Figure 6

Power (in %) in conditions with medium correlations between items (ρ = .10) in Study 3. n = sample size; k = number of items per factor; CD = complete data; LD = listwise deletion; FIML = full-information maximum likelihood; D_2, D_3, D_4 = pooling methods.