Correcting for Unreliability and Partial Invariance: A Two-Stage Path Analysis Approach

Mark H. C. Lai[1], Winnie Wing-Yee Tse[1], Gengrui Zhang[1], Yixiao Li[2], and & Yu-Yu Hsiao[3]

[1] Department of Psychology, University of Southern California

[2] Department of Computer Science, University of Southern California

[3] Department of Individual, Family, & Community Education, University of New Mexico

Author Note

Mark H. C. Lai https://orcid.org/0000-0002-9196-7406

Winnie Wing-Yee Tse https://orcid.org/0000-0001-5175-6754

Gengrui Zhang https://orcid.org/0000-0002-2062-6620

Yixiao Li https://orcid.org/0000-0002-8039-338X

Yu-Yu Hsiao https://orcid.org/0000-0001-9296-4517

This is an Accepted Manuscript of an article to be published by Taylor & Francis in Structural Equation Modeling: A Multidisciplinary Journal on September 14, 2022, to be available online: http://www.tandfonline.com/10.1080/10705511.2022.2125397

Simulation codes and data are openly available on the project's GitHub page (github.com/marklhc/2spa-inv-supp). This research is based on work supported by the National Science Foundation under Grant No. 2141790. We have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Mark H. C. Lai, Department of Psychology, University of Southern California, 3620 S McClintock Ave., Los Angeles, CA 90089-1061, United States. E-mail: hokchiol@usc.edu

Abstract

In path analysis, using composite scores without adjustment for measurement unreliability and violations of factorial invariance across groups leads to biased estimates of path coefficients. Although joint modeling of measurement and structural models can theoretically yield consistent structural association estimates, the estimation of a model with many variables is often impractical in small samples. A viable alternative is two-stage path analysis (2S-PA), where researchers first obtain factor scores and the corresponding individual-specific reliability coefficients, and then use those factor scores to analyze structural associations while accounting for their unreliability. The current paper extends 2S-PA to also account for partial invariance. Two simulation studies show that 2S-PA outperforms joint modeling in terms of model convergence, the efficiency of structural parameter estimation, and confidence interval coverage, especially in small samples and with categorical indicators. We illustrate 2S-PA by reanalyzing data from a multiethnic study that predicts drinking problems using college-related alcohol beliefs.

Keywords: two-stage path analysis, factorial invariance, partial invariance, measurement error, factor scores

Word count: 7,366

Correcting for Unreliability and Partial Invariance: A Two-Stage Path Analysis Approach

Over the past two decades, there has been a tremendous increase in research evaluating the measurement invariance of instruments in psychology. If measurement invariance—the condition that an instrument measures the same construct the same way across groups—is violated, the observed composite scores are not on the same metric across groups, and thus group comparisons using those scores are not meaningful. That said, when only part of the items in an instrument is noninvariant—meaning that the instrument is partially invariant—researchers can still obtain valid statistical results by jointly modeling partial invariance and the structural associations among the latent constructs see Hsiao and Lai, 2018. However, the joint modeling approach is computationally demanding as it requires including all measurement indicators in the analysis, even when researchers only have a relatively simple structural model. Also, when the sample size is relatively small, joint modeling often suffers from issues of convergence and nonadmissible solutions (Rosseel, 2020). As discussed later in this paper, in practice, researchers rarely use the joint modeling approach to adjust for partial invariance, but continue to use composite scores (e.g., sum scores or mean scores) following invariance analyses.

However, using composite scores without any adjustment is potentially problematic in two regards. First, the presence of noninvariant items can systematically bias analysis results, such as regression coefficients or mean comparisons. Second, using composite scores assumes that they do not contain measurement error, meaning they are perfectly reliable, which is rare, if possible, in behavioral and social sciences. It is well known in the literature that ignoring measurement unreliability leads to biased regression coefficients (e.g., Carroll et al., 2006; Cole & Preacher, 2014; Ledgerwood & Shrout, 2011).

As an alternative, recently, there has been a renewed interest in using psychometric-model-based factor scores (e.g., Estabrook & Neale, 2013; McNeish & Wolf, 2020), which adjust for partial invariance to put the latent variables on a common or approximately common metric (e.g., Curran & Hussong, 2009). However, like sum scores, factor scores are also not perfectly reliable, so using them in analyses without correction for measurement error will still lead to biased coefficients, with the magnitude of bias depending on the reliability of the factor scores (Croon, 2002; Levy, 2017). Also, as shown later, when partial invariance exists, not every way of computing factor scores results in scores on the same metric, so further adjustment is needed.

Two general and related approaches to account for measurement error when using estimated scores (i.e., composite or factor scores) are of interest. In the first approach, researchers first obtain naive path coefficients by treating the estimated scores as the true latent variable scores. Correction factors are obtained based on the relation between the estimated scores and the latent variable in the measurement model, and then applied to the naive coefficients to obtain corrected coefficients. The correction factors are usually functions of score reliability. Fan (2003) discussed an example with two latent variables, η_X and η_Y. When the two latent variables were measured by multi-item scales that give composite scores X and Y, respectively, one can estimate the true correlation between η_X and η_Y as $r_{XY}/\sqrt{\rho_X \rho_Y}$, where r_{XY} is the correlation between the composite scores and ρ_X and ρ_Y are the composite reliability of X and Y, respectively. Croon (2002) showed how this approach can be used when factor scores are used instead, with slightly more involved correction formulas that are functions of factor loadings and latent variances. The method of Croon was further elaborated in the method of factor score path analysis (Devlieger & Rosseel, 2017; Devlieger et al., 2016), which also includes corrected standard errors and inferences for the corrected path coefficients.

The second approach is the reliability adjustment method, which treats the composite scores or factor scores as single indicators of latent variables and constrains the reliability of these indicators to either known values or estimates from the data (e.g., Bollen, 1989; Hsiao et al., 2018, 2021; Kwok et al., 2016; Savalei, 2019). However, both the correction factor approach and the reliability adjustment method generally assume constant measurement error variance for the whole sample, which is likely violated when only partial invariance holds or when indicators are binary or ordinal. Thus, previous methods for handling measurement error only address parameter bias due to unreliability, and may still yield inconsistent estimates due to unadjusted partial invariance. A more general approach to reliability adjustment is the two-stage path-analysis (2S-PA) with definition variables method by M. H. C. Lai and Hsiao (2021), which accounts for the unreliability in factor scores even when reliability is not constant across observations. While previous studies have only focused on the reliability adjustment aspect of 2S-PA, the current paper shows how researchers can use 2S-PA to adjust for both partial invariance and unreliability for continuous and discrete indicators. We also report evidence from two simulation studies showing that 2S-PA has fewer convergence issues and more accurate estimation and inference in small samples than the joint structural equation modeling (SEM) approach.

Multiple-Group Joint SEM

In behavioral sciences, joint SEM modeling is the recommended approach for incorporating imperfect measurement when analyzing relations among latent variables (e.g., Cole & Preacher, 2014). In SEM, theoretical constructs, such as depression and cognitive ability, are represented as latent variables, ηs, and each of them is measured by one or more observed indicators. When both the measurement (between ηs and their indicators) and the structural (among the ηs) models are correctly specified, joint SEM modeling with maximum likelihood estimation yields consistent and asymptotically efficient structural coefficient estimates (e.g., Bollen, 1989). In the case of partial invariance, where some measurement parameters differ across a grouping variable G, one common approach is to use a multiple-group analysis that places equality constraints on only those measurement parameters found invariant across groups. Specifically, denote the measurement model among p observed indicators, $\mathbf{y}$, and q latent variables, $\boldsymbol{\eta}$, as $f(\mathbf{y}|\boldsymbol{\eta}, \boldsymbol{\omega})$ with measurement parameters $\boldsymbol{\omega}$, and assume that the structural model can be characterized as a linear model. Assuming that each observation $i = 1, \ldots, n_g$ in group g is independent, the multiple-group joint SEM can be described by the model

$$\begin{aligned} \text{Measurement: } & f(\mathbf{y}_{ig}|\boldsymbol{\eta}_{ig}, \boldsymbol{\omega}_g) \\ \text{Structural: } & \boldsymbol{\eta}_{ig} = \boldsymbol{\alpha}_g + \mathbf{B}\boldsymbol{\eta}_{ig} + \boldsymbol{\zeta}_{ig} \end{aligned}, \tag{1}$$

with equality constraints on a subset of $\boldsymbol{\omega}_g$ across some groups g_1, g_2, and so forth. In the structural model, $\boldsymbol{\alpha}$ contains the q latent regression intercepts, $\mathbf{B}$ is a $q \times q$ matrix of structural coefficients, and $\boldsymbol{\zeta}_{ig}$ is a vector of disturbances with the standard assumption $\mathrm{E}(\boldsymbol{\zeta}_{ig}) = \mathbf{0}$.[1]

As an example, consider a model with $\mathbf{y} = [y_1, \ldots, y_6]^\top$ and $\boldsymbol{\eta} = [\eta_X, \eta_2]$, where η_2 is observed (similar to Figure 1). Assuming a linear factor model linking $\mathbf{y}$ and η_1 such that $y_j = \nu_j + \lambda_j \eta_1 + \varepsilon_j$ with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Theta})$ and j indexing indicators, the measurement parameters are $\boldsymbol{\omega} = (\boldsymbol{\nu}, \boldsymbol{\lambda}, \boldsymbol{\Theta})$ and there is one structural path coefficient in $\mathbf{B} = \begin{bmatrix} 0 & 0 \\ \beta_1 & 0 \end{bmatrix}$. When there are multiple groups, a prerequisite to compare structural coefficients (e.g., $\boldsymbol{\alpha}$ and $\mathbf{B}$) across groups is that $\boldsymbol{\omega}$ is sufficiently invariant, which can be examined by analyzing factorial invariance (Meredith, 1964)—measurement invariance under a factor model—of the items. If the items all have the same number of underlying factors (one in this example), and the pattern of how the items and the underlying factors are linked is the same across groups, the condition of configural invariance (Horn & McArdle, 1992; Meredith, 1964) is met. Furthermore, the items are considered metric invariant if $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_2$, scalar invariant if, additionally, $\boldsymbol{\nu}_1 = \boldsymbol{\nu}_2$, and strict invariant if also $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_2$ so that all measurement parameters are equal (e.g., Widaman & Reise, 1997). If some elements of $\boldsymbol{\omega}$ are not invariant across groups, a partial invariance model should be specified so that only the invariant subset of $\boldsymbol{\omega}$ is constrained equal across groups in the joint SEM approach (e.g., Byrne et al., 1989; Hsiao & Lai, 2018).

Although the multiple-group joint SEM approach (hereafter JSEM) is very flexible, as discussed in the previous literature (e.g., Croon, 2002; Devlieger et al., 2016; M. H. C. Lai & Hsiao, 2021; McNeish &

[1] Note that we could allow $\mathbf{B}$ to be group-specific to represent $G\times$ $\boldsymbol{\eta}$ interactions; however, based on our small literature review (described later in the paper), researchers rarely specified such an interaction, so in the current paper we mainly focus on analyses with a common $\mathbf{B}$.

Wolf, 2020; Rosseel & Loh, 2021), it does present several challenges, both in terms of computation and usage in practice. First, JSEM requires specifying a large model even when the structural model is relatively small. Consider a structural model with three latent constructs, each measured by 10 observed indicators, resulting in a total of 30 observed variables. Although a researcher may only be interested in three or four structural coefficients, JSEM would need estimations of hundreds of measurement parameters. Not only is the large model size difficult for researchers to keep track of and identify misfits, but it also makes JSEM more prone to convergence failures in optimization algorithms and inflated Type I error rates of the structural coefficients (Devlieger & Rosseel, 2017; Kelcey, 2019; M. H. C. Lai & Hsiao, 2021; Rosseel, 2020). Second, for models not assuming a multivariate normal likelihood function, such as when the observed indicators are ordinal or categorical, estimating JSEM models (e.g., with numerical integration) is computationally demanding and not feasible even with only a few latent dimensions (Pritikin et al., 2018). Third, with joint modeling, misspecifications in the measurement models can affect parameter estimates in the structural model. Compared to multistage methods like 2S-PA, structural parameters with JSEM are more susceptible to misspecifications in the measurement models (Devlieger & Rosseel, 2017; M. H. C. Lai & Hsiao, 2021). The other side of the same coin—that misspecifications in the structural model can affect parameter estimates in the measurement models—is equally troubling. It leads to *interpretational confounding* (Bollen & Maydeu-Olivares, 2007; Burt, 1976; Levy, 2017), where the operationalization of the latent construct is different in different structural models, even with the same data.

A Brief Review on the Use of Joint Modeling Following Invariance Evaluation

Given the conceptual and computational challenges of SEM, researchers often use composite or factor scores to analyze structural models, even after they conduct extensive psychometric explorations such as measurement invariance analysis. For example, in a review of articles published in the *Journal of Applied Psychology* and *Personality and Individual Differences* in 2020, we identified 30 articles that either tested measurement invariance ($n = 26$) or cited external evidence for measurement invariance ($n = 4$). Among them, 26 articles concluded with configural ($n = 2$), metric ($n = 13$), scalar ($n = 10$), or strict ($n = 1$) invariance; the remaining articles either reported noninvariance ($n = 3$) or provided insufficient information about test results ($n = 1$).

Even though many of the articles we reviewed already performed invariance analyses, the majority ($n = 16$) still used composite scores for subsequent statistical analyses, while others used either factor scores ($n = 1$) or JSEM ($n = 8$). For the articles that used composite scores, only four supported scalar or strict invariance (i.e., the minimum requirement for composite scores to be comparable across groups; Putnick & Bornstein, 2016), whereas eight established only metric invariance, two showed only configural invariance, and two concluded with measurement noninvariance.

When comparing the sample sizes of articles using different methods, those using composite scores had a median of two groups with median $\bar{n} = 377$ per group, whereas the ones using JSEM had a median of three groups with median $\bar{n} = 451$ per group. While JSEM requires a relatively large sample size (e.g., $n > 500$ or 2000; Rosseel, 2020), studies with fewer samples might adopt alternative methods for group comparisons or regression analyses.

Thus, this brief review shows that researchers commonly used composite scores in subsequent analyses, even when measurement invariance was violated. Ignoring violations of measurement invariance and imperfect reliability may result in biased statistical results. Therefore, alternative methods that are easy to specify while still producing consistent estimates are desirable. The current paper will focus on 2S-PA as one of those methods.

Two-Stage Path Analysis (2S-PA) With Definition Variables

Building on the literature of errors-in-variables models (e.g., Carroll et al., 2006; Meijer et al., 2021), M. H. C. Lai and Hsiao (2021) proposed 2S-PA as an alternative to JSEM. In the first stage of 2S-PA, researchers obtain factor scores for each observation i on each latent construct m, $\tilde{\eta}_{mi}$, and their estimated reliability, $\tilde{\rho}_{\tilde{\eta}mi}$. In addition, in order to account for noninvariant measurement parameters, the factor scores should be obtained from partial invariance models where η is calibrated to be on the same metric. Unlike JSEM, where the same software and estimation method are used for all measurement and structural models, with 2S-PA, one can use different software for obtaining factor scores for different constructs, as long as consistent estimates of factor score reliability can be obtained for each observation. For example, one can use a specialized item response model for factor scores of one construct and a network model for centrality scores for another construct, as long as they are appropriate models to operationalize variables in their hypothesized model. In the second stage of 2S-PA, full-information maximum likelihood is used to estimate the structural model:

$$\begin{aligned} \text{Measurement: } \tilde{\boldsymbol{\eta}}_i &= \Lambda_i^* \boldsymbol{\eta}_i^* + \boldsymbol{\varepsilon}_i^* \\ \text{Structural: } \boldsymbol{\eta}_i^* &= \boldsymbol{\alpha}^* + \mathbf{B}^* \boldsymbol{\eta}_i^* + \boldsymbol{\zeta}_i^* \end{aligned}, \tag{2}$$

where Λ_i^* is a loading matrix and is assumed diagonal when each factor score variable is an indicator of only one latent variable, $\boldsymbol{\zeta}_i^* \sim N(\mathbf{0}, \boldsymbol{\Psi}^*)$ and $\boldsymbol{\varepsilon}_i^* \sim N(\mathbf{0}, \boldsymbol{\Theta}_i^*)$. When the factor scores are calibrated to the same metric across individuals and groups, one can set $\Lambda_i^* = \mathbf{I}$ for identification; however, when they are not calibrated, $\Lambda_i^* \neq \mathbf{I}$ should be specified so that $\boldsymbol{\eta}^*$ is on the same metric across groups, as discussed below for composite scores and factor scores obtained with the regression method. The 2S-PA model further accounts for the unreliability of $\tilde{\eta}_m$ by setting the ratio of true score variance ($\lambda^{*2}_{mi}\,\text{Var}[\eta^*_m]$) and the total variance (i.e., true score variance + error variance) to the reliability value estimated in stage 1, such that

$$\frac{\lambda^{*2}_{mi}\,\text{Var}(\eta^*_m)}{\theta^*_{mi} + \lambda^{*2}_{mi}\,\text{Var}(\eta^*_m)} = \tilde{\rho}_{\tilde{\eta}mi}. \tag{3}$$

Note that we use η^* in equation (2), as they can be on a different metric than η in the first stage estimation; thus, the unstandardized parameter estimates from JSEM and 2S-PA are generally not comparable. M. H. C. Lai and Hsiao (2021) showed that in single-group analyses, one should compare the standardized coefficients; as discussed later, when the analyses involve multiple groups, additional adjustments on the standardized coefficients are needed to place the parameter estimates from multiple-group JSEM and single-group 2S-PA on approximately the same unit.

The constraints in 2S-PA are similar to those discussed in the reliability adjustment literature (e.g., Hsiao et al., 2018; Meijer et al., 2021; Savalei, 2019), except that it allows the reliability to be observation-specific, which accommodates ordered categorical items and violations of strict factorial invariance. It thus requires software programs that support observation-specific constraint variables, such as OpenMx (via definition variables; Neale et al., 2016) and Mplus (via constraint variables; Muthén & Muthén, 1998–2017).

2S-PA With Various Estimated Scores

Below, we consider how 2S-PA can be applied to three commonly computed scores for continuous indicators under a factor model: regression factor scores (Thomson, 1935), Bartlett factor scores (Bartlett, 1937), and sum scores. In each case, the estimated scores are linear combinations of the observed item scores such that $\tilde{\eta}_{ig} = \mathbf{A}_g \mathbf{y}_{ig}$, where $\mathbf{A}_g$ is the factor score matrix. For simplicity, we drop the mean structure in the discussion as mean differences across groups do not affect the path coefficients when the group membership is included as a covariate in the second stage analysis (Curran et al., 2018). We also assume that the items are unidimensional, so only one latent variable is involved. As the factor model implies $\mathbf{y}_{ig} = \boldsymbol{\lambda}_g \eta_{ig} + \boldsymbol{\varepsilon}_{ig}$, we have $\tilde{\eta}_{ig} = \mathbf{A}_g \boldsymbol{\lambda}_g \eta_{ig} + \mathbf{A}_g \boldsymbol{\varepsilon}_{ig} = \lambda^*_g \eta_{ig} + \varepsilon^*_i$.

As such, the reliability of the estimated scores is $(\mathbf{A}_g \boldsymbol{\lambda}_g)^2 \psi_g / [(\mathbf{A}_g \boldsymbol{\lambda}_g)^2 \psi_g + \mathbf{A}_g \boldsymbol{\Theta}_g \mathbf{A}_g^\top]$, where ψ_g is the variance of η_g and $\boldsymbol{\Theta}_g$ is the unique factor covariance matrix. When factorial invariance does not hold across groups, generally $\mathbf{A}_g \boldsymbol{\lambda}_g$ is different for different gs, so the estimated scores are on different metrics across groups. Therefore, the second stage of 2S-PA needs to incorporate information of $\mathbf{A}_g \boldsymbol{\lambda}_g$ when setting the loading of $\tilde{\eta}$ on η^* so that η^* is calibrated to the same metric.

As shown in Table 1, for both the regression factor scores and the sum scores, the loading (λ^*) of $\tilde{\eta}$ on η depends on $\boldsymbol{\lambda}_g$, so the scores are on different metrics when metric invariance is violated. Therefore, in 2S-PA, λ^* needs to be group-specific by setting the loading parameter as a definition/constraint variable. On the other hand, the Bartlett scores are calibrated to be on the same unit as the latent variable with $\lambda^* = 1$, so group-specific loading is not needed for 2S-PA. Also, regression scores are shrinkage estimates, meaning they have a smaller variance when reliability is low (note that $\lambda^* = \rho_{\tilde{\eta}}$), whereas Bartlett scores are not. For sum scores, $\rho_{\tilde{\eta}}$ is the familiar ω reliability for composite scores (McDonald, 1999; Raykov, 1997). In Study 1, we evaluate the performance of 2S-PA with these three types of estimated scores for continuous items.

For categorical items, sum scores are generally not appropriate as the items are not intervally scaled. As discussed in Hoshino and Bentler (2013), the expected a posteriori (EAP) scores are analogous to the regression factor scores, whereas maximum likelihood estimates of η are analogous to the Bartlett factor scores.

Within-Group Standardization and Grand Standardization

Structural parameter estimates depend on the assigned metrics of the latent variables. Because JSEM and 2S-PA use observed variables on different units, the unstandardized parameter estimates are generally not comparable. One solution is to look at the standardized coefficients, namely, the transformed **B** coefficients when all ηs have unity variance. However, in a multiple-group analysis (e.g., multiple-group SEM), coefficients are often standardized using the within-group *SD* for η_m ($\sigma_{\eta mg}$), whereas in single-group analysis with groups pooled into one analytic sample (e.g., in 2S-PA), coefficients are standardized using the grand, or total, *SD* ($\sigma_{\eta m}$). Let $\mu_{\eta m1}, \ldots, \mu_{\eta mG}$ be the latent means of η_m across groups and $n_1, \ldots, n_G$ be the respective sample sizes with $\sum_{g=1}^{G} n_g = N$, then one can show that the grand *SD* is related to the within-group *SD* in the equation (dropping the η_m subscript for better readability)

$$\sigma^2 = \frac{1}{N}\sum_{g=1}^{G} n_g[\sigma_g^2 + (\mu_g - \mu)^2], \tag{4}$$

where $\mu = \sum_{g=1}^{G} n_g \mu_g / N$ is the grand mean. While researchers may prefer one way of standardization or the other in applied research, in simulation studies or research syntheses where different methods are compared, the coefficients are comparable only when converted to the same *SD* unit.

Current Studies

In the remainder of the current paper, we report two simulation studies comparing JSEM and 2S-PA in the presence of partial invariance. In Study 1, we use a simple latent regression model in which only the predictor contains measurement error and partial invariance. In Study 2, we use a more complex mediation model involving three constructs, wherein both the mediator and the outcome contain measurement error and partial invariance. In addition, Study 2 also involves data with binary indicators. The two simulation studies cover balanced and unbalanced sample sizes across two groups. We then provide an example using data from a published paper illustrating how researchers can use 2S-PA following evidence of partial invariance. We conclude with some future research directions for 2S-PA.

Study 1

In Study 1, we compare two approaches without reliability adjustment: sum-score path analysis (PA) and factor score path analysis (FS-PA, with regression factor scores), with five approaches that adjust for unreliability: Croon's correction (Croon), JSEM, and three 2S-PA methods, for estimating a regression coefficient. We examined three variations of 2S-PA that use (a) regression factor scores, (b) Bartlett factor scores, and (c) sum scores. In the data generating model, the latent predictor, X, is measured by six indicators with partial invariance across two groups ($G = 1$ and 2). We generate data with four levels of sample size for Group 1 ($n_1 = 50$, 100, 500, 1000), and the data is either balanced ($n_2 = n_1$) or unbalanced ($n_2 = 0.6n_1$). The average loading for Group 1 has two levels to represent situations of low reliability (average loading = 0.7; composite reliability = .49 and .61 for Groups 1 and 2) and moderate reliability (average loading = 1.0; composite reliability = .71 and .77 for Groups 1 and 2).[2] Figure 1 shows the data generating values of the model parameters, where Group 2 has larger loadings on items 2 and 5. In addition, items 4 and 5 have different intercepts, and items 4 and 6 have different unique variances. The two groups also have different means and variances of η_X. To resemble minor misspecification in the measurement model, we follow the suggestion by MacCallum and Tucker (1991) to add minor common variances among the indicators, which results in covariances of magnitudes between -0.356 and 0.356 (i.e., 10% of the observed indicator variance in Group 1). For each condition, we simulated 2,500 replications using R.

The unstandardized regression coefficient b_1 is manipulated to either 0 or 0.5 for both groups (i.e., no $\eta_X \times G$ interaction). To account for the above-mentioned metric incomparability issue in the estimated coefficients, we obtained the regression coefficient with η_X standardized using the grand *SD* of X. When b_1

[2] The composite reliability for sum scores is computed using the same formula as presented in Table 1 (see also Raykov, 1997).

= 0.5, the standardized coefficient is $\beta_1 \approx 0.54$ (unbalanced samples) and 0.56 (balanced samples), whereas when $b_1 = 0$, β_1 is also zero. All data generation is carried out in R. The package OpenMx (Neale et al., 2016) is used to obtain the grand-standardized regression coefficient using composite scores of X (i.e., PA, which ignores noninvariance and unreliability), factor scores of X (i.e., FS-PA, which adjusts for partial invariance but not unreliability), Croon's correction (see supplemental material for implementation details), JSEM, and 2S-PA methods. For each condition and method, we evaluated the raw bias and root mean squared error (RMSE) in estimating β_1, as well as the relative standard error bias and the coverage of the 95% confidence interval (CI). The full R script for the simulation can be found in the supplemental materials (https://github.com/marklhc/2spa-inv-supp/).

For all conditions, the convergence rates were 99% or above; we only observed some estimation issues in small-sample, low-reliability conditions for JSEM and 2S-PA; in some replications, there were problems obtaining likelihood-based CI for 2S-PA. As shown in Tables 2 and 3, the impact of unequal ns across groups was small. The results for PA and FS-PA were highly similar, except that FS-PA had higher RMSEs and larger SE biases in small-sample, low-reliability conditions. Croon's correction performed slightly worse than 2S-PA methods for most conditions in terms of bias. When $b_1 = \beta_1$ (standardized coefficient) = 0, the estimates for all methods were essentially unbiased (with $|\text{bias}| \leq 0.01$). When $b_1 = 0$ and $\beta_1 = 0$, JSEM and 2S-PA with sum scores (no bigger than 0.02 in absolute values) had the least bias, while PA and FS-PA generated larger biases up to -0.15. 2S-PA with regression scores and Bartlett scores showed downward bias in small samples (-0.07 for regression scores; -0.06 for Bartlett scores), but improved with larger samples. For RMSE, PA performed the best when estimating a zero coefficient as there was no attenuation due to unreliability; 2S-PA methods performed slightly better than JSEM for estimating a zero coefficient and were virtually identical to JSEM across other conditions. In small-sample, low-reliability conditions, 2S-PA with sum scores performed best in terms of RMSE.

When the reliability was relatively high, all methods gave acceptable standard errors, and all methods except PA and FS-PA had acceptable CI coverage; the latter two had suboptimal coverage when $\beta_1 \neq 0$, because their estimated coefficients were attenuated due to unreliability. When reliability was low, the standard errors with JSEM were severely underestimated (up to -71.91%), and coverage was suboptimal ($< 92\%$) when the sample size was small; FS-PA and 2S-PA with regression and with Bartlett scores had substantial bias in the estimated standard errors and undercoverage for nonzero true coefficients in small samples, probably due to some instability in factor score estimation; Croon's correction performed better than 2S-PA in terms of SE bias, but had worse coverage rates for low-reliability, small-sample conditions. Overall, 2S-PA with sum scores performed well for all conditions; JSEM and 2S-PA were similar and performed best for conditions with large sample sizes and relatively high reliability. When comparing 2S-PA methods with regression scores and with Bartlett scores, they were generally similar, with the former giving slightly better coverage rates overall.

The results of Study 1 show that 2S-PA and JSEM are both effective in accounting for both partial invariance and unreliability when sample sizes are large or when score reliability is above .70. Also, 2S-PA seems to give more efficient estimates and control the Type I error rate better. In small-sample or low-reliability situations with continuous indicators, Study 1 shows that 2S-PA with sum scores can be used for valid inferences and better estimation efficiency. The difference between 2S-PA and JSEM may be more prominent with more complex models, as shown in Study 2.

Study 2

As 2S-PA fits a simpler model in each step, compared to JSEM, we expect that it shows more benefits in a more complex model, particularly when some of the indicators for the latent variables are categorical. In Study 2, we consider a mediation model with a binary treatment variable X and with both the mediator η_M and the outcome η_Y variables measured with errors. Both η_M and η_Y showed noninvariance with respect to a grouping variable G. As shown in Figure 2, there were six indicators for η_M and 16 indicators for η_Y. For the indicators of η_M, the population values of loadings, intercepts, and unique variances were the same as those in the high-reliability condition in Study 1. For the indicators of η_Y, we simulated them to be binary items following a 2-parameter normal ogive item response model such that

$$y_j^* = \lambda_{Yj}\eta_Y + \varepsilon_{Yj}$$

$$y_j = \begin{cases} 1 & y_j^* > \tau_j \\ 0 & \text{otherwise} \end{cases},$$

with $\varepsilon_{Yj} \sim N(0, 1)$; λs are the loading parameters analogous to those in the factor model, and τs are the thresholds. The population values of the measurement parameters were taken from a real-data abstract reasoning test example in Embretson and Reise (2000, Table 4.2, p. 69), with loadings between 0.465 to 0.958 and item difficulties between -2.118 to 1.061 for Group 1. Items 1, 5, 9 were simulated to have noninvariant loadings (magnitude = 0.118 to 0.294), and items 2, 5, 8 were simulated to have noninvariant thresholds (magnitude = 0.3 to 0.5). The exact values can be found in the simulation code. The test information for the η_Y indicators was above 1.81 for η_Y between -2 and 2, with peak information of 4.29 for Group 1; it was similar for Group 2 (above 1.74 for $\eta_Y \in$ [-2, 2], peak = 4.42).

Similar to Study 1, we added minor common variances among the indicators of η_M and η_Y, resulting in unique correlations in the range [-0.1, 0.1]. The sample sizes were equal across the two levels of G, with conditions of n = 50, 100, 300, 1,000 per group.

For both groups, we had the following structural model:

$$\eta_M = \alpha_M + \beta_1 X + \zeta_M$$
$$\eta_Y = \alpha_Y + \beta_2 X + \beta_3 \eta_M + \zeta_Y.$$

We allowed α_M and α_Y to be group-specific to represent main effects of G on η_M and η_Y, but there were no group-related interactions. The population values were $\alpha_M = 0$ and 0.2, and $\alpha_Y = 0$ and 0.3, respectively for $G = 1$ and 2; also, for all conditions, we fixed $\beta_2 = 0.3$. There were four conditions for the values β_1 and β_3, including (a) $\beta_1 = \beta_3 = 0$, (b) $\beta_1 = 0.5$, $\beta_3 = 0$, (c) $\beta_1 = 0$, $\beta_3 = 0.3$, and (d) $\beta_1 = 0.5$, $\beta_3 = 0.3$. Note that the indirect effect of X on η_Y was $\beta_1\beta_3 = 0$ for (a), (b), and (c), and was 0.15 for (d). The values of $\text{Var}(\zeta_M)$ and $\text{Var}(\zeta_Y)$ were chosen such that the grand variances of η_M and η_Y are both one, so that the grand-standardized coefficients (i.e., using the total *SD* without group memberships) and the unstandardized coefficients were the same.

We compared multiple-group JSEM, path analysis with factor scores (FS-PA; without reliability adjustment), and 2S-PA. For JSEM, we used the *lavaan* package (Rosseel, 2012) in R to fit a full SEM model with partial invariance using all indicators, with identification constraints such that the grand variances of η_M and η_Y were unity. Diagonally weighted least squares (DWLS) was used as the model included both continuous and binary indicators. For FS-PA and 2S-PA, we first used *lavaan* and a multiple-group CFA (with maximum likelihood estimation) to obtain the regression factor scores for η_M, denoted as $\tilde{\eta}_M$, and then used the *mirt* R package (Chalmers, 2012) and a multiple-group two-parameter logistic item response model (with maximum likelihood estimation) to obtain the expected a posteriori (EAP) scores for η_Y, denoted as $\tilde{\eta}_Y$. For 2S-PA, reliability estimates, $\tilde{\rho}_{\tilde{\eta}_Y i}$ and $\tilde{\rho}_{\tilde{\eta}_M i}$, were computed using $1 - SE^2(\tilde{\eta}_i)/\text{Var}(\eta)$, where $SE(\tilde{\eta}_i)$ is the case-specific standard error of the EAP score, available from *mirt*. Similar to multiple-group SEM, in the item response models, the loadings and thresholds were constrained equal for the invariant items but free for the noninvariant items, so the latent factor was on the same metric. In both JSEM and the first stages of FS-PA and 2S-PA, we specified the correct partial invariance models (but without the unique covariances). We deliberately used two separate programs for 2S-PA to demonstrate its flexibility. In the second stage, we used *OpenMx* with the measurement model

$$\begin{aligned}\tilde{\eta}_{Mi} &= \eta_{Mi} + e_{\tilde{\eta}_{Mi}} \\ \tilde{\eta}_{Yi} &= \eta_{Yi} + e_{\tilde{\eta}_{Yi}}\end{aligned},$$

the constraints

$$\begin{aligned}\tilde{\rho}_{\tilde{\eta}_M i}\,\mathrm{Var}(e_{\tilde{\eta}_M i}) &= (1 - \tilde{\rho}_{\tilde{\eta}_M i})\,\mathrm{Var}(\eta_M)\\ \tilde{\rho}_{\tilde{\eta}_Y i}\,\mathrm{Var}(e_{\tilde{\eta}_Y i}) &= (1 - \tilde{\rho}_{\tilde{\eta}_Y i})\,\mathrm{Var}(\eta_Y)\end{aligned},$$

and the structural model

$$\begin{aligned}\eta_M &= \alpha_M + \alpha_1 G + \beta_1 X + \zeta_M\\ \eta_Y &= \alpha_Y + \alpha_2 G + \beta_2 X + \beta_3 \eta_M + \zeta_Y.\end{aligned}$$

The inclusion of $\alpha_1 G$ and $\alpha_2 G$ accounted for the intercept differences across groups.

For all three approaches, we obtained the standardized coefficients for the β_1, β_2, β_3 paths, as well as the product term $\beta_1\beta_3$ (i.e., the standardized indirect effect). With JSEM, the corresponding 95% CIs were obtained using the delta method for β_1 to β_3, and the Monte Carlo method (MacKinnon et al., 2004) for $\beta_1\beta_3$; with 2S-PA, CIs were obtained using the profile likelihood method (Pek & Wu, 2015). The analytic approaches were compared based on the convergence rate, bias, RMSE, and 95% CI coverage. We also evaluated the statistical power based on the proportion of replications where the 95% CI excludes zero for conditions with nonzero indirect effects.

Results

Convergence

Convergence was 100% for all conditions with FS-PA. When $n = 50$, convergence was substantially better for 2S-PA (89.80%) than for JSEM (8.29%). When $n \geq 100$, 2S-PA had 100% convergence, but JSEM still had convergence issues (32.56%). JSEM had 82.49% convergence when $n = 300$, and 99.66% when $n = 1{,}000$. The main reason for nonconvergence in 2S-PA was failures in computing factor scores or the corresponding reliability in the first stage due to negative latent or error variance estimates, whereas it was empirical unidentifiability due to near-perfect or near-zero associations among indicators for JSEM.

Bias

Figure 3 shows the bias in estimating the βs and the indirect effect. All three methods estimated coefficients that are truly zero with little bias. When the true coefficients were nonzero, FS-PA, ignoring measurement error, produced biased estimates for virtually all coefficients (bias between -0.117 and 0.014). Both 2S-PA and JSEM performed better with a larger n; with a small $n = 50$, 2S-PA (bias between -0.066 and -0.026) generally performed better than JSEM (bias between -0.240 and 0.116), especially for β_2 and β_3.

RMSE

Figure 4 shows the RMSE of the different methods, which combines both bias and (in)efficiency of the estimates. The RMSEs for FS-PA were the smallest for small-sample conditions, especially when there could be little attenuation due to measurement error; however, FS-PA performed worst in larger samples for nonzero coefficients. For all of the β coefficients and the indirect effect, 2S-PA generally provided better RMSEs than JSEM, especially in small samples. When n reaches 300, the RMSEs were comparable for 2S-PA and JSEM.

Coverage

Figure 5 shows the coverage of 95% CI for 2S-PA and JSEM; coverage for FS-PA was bad for nonzero coefficients due to parameter bias (close to 0.00 for β_3 and $\beta_1\beta_3$ when $n = 1{,}000$), and was excluded from the graph. 2S-PA showed coverage close to 95% for almost all conditions and parameters, except for some undercoverage when estimating zero β_3 in small samples. JSEM generally had worse coverage than 2S-PA, which also corresponded to severely inflated Type I error rates (i.e., 1 - coverage rate when true coefficient = 0) of up to 0.34 when $n = 50$ for β_3, whereas 2S-PA had Type I error rates < 0.06 for all conditions and coefficients.

Power

Figure 6 shows the empirical power, calculated as the rates in which the 95% CI excluded zero when making inferences on coefficients that are truly nonzero. Power was generally similar for FS-PA and 2S-PA, while JSEM had higher power for β_1, β_2, and β_3 with small samples (but at the cost of higher Type I error rates). When $n \geq 100$, the empirical power was similar for all three approaches.

In summary, with a more complex data generating model, we found 2S-PA to have substantially fewer convergence issues than JSEM, and it mostly outperforms JSEM in parameter estimation and inference, especially in small samples.

Empirical Example

In this section, we demonstrate 2S-PA as well as PA, FS-PA, and the JSEM approaches, using empirical data made publicly available by Lui (2019) on the Open Science Framework (https://osf.io/93qpt/). Data were collected in 2018 from 1,148 undergraduate students, aged 18 or older, in a private university. Lui evaluated measurement invariance of the College Life Alcohol Salience Scale (CLASS; Osberg et al., 2010), which measures individuals' college-related alcohol beliefs, across different sociodemographic subgroups, including ethnicity. Subsequently, CLASS was used to predict students' alcohol consumption and drinking problems, measured by the Alcohol Use Disorders Identification Test (AUDIT; Saunders et al., 1993). While meeting scalar invariance across most grouping variables, CLASS showed partial scalar invariance across ethnicity. For pedagogical purposes, we focus on analyzing the relationship between college-related alcohol beliefs and drinking problems across ethnic groups in this demonstration.

CLASS contains fifteen 5-point Likert items (1 = *strongly disagree* and 5 = *strongly agree*). Seven of the ten items of AUDIT measure negative alcohol-related consequences, i.e., drinking problems, on a variety of 3-to-5-point scales.[3] Study participants were domestic students of European American (44.9%), Asian American (19.9%), African American (10.3%), Latinx American (16.7%), and mixed or other ethnic backgrounds (8.3%).

We assess configural, metric, and scalar invariance of CLASS and AUDIT, respectively, using *lavaan* (Rosseel, 2012) with maximum likelihood estimation. If a more constrained model has a worse fit than a less constrained model, indicating invariance violations, we use sequential specification search (Yoon & Kim, 2014) to identify and free noninvariant parameters, until arriving at a partial invariance model. After establishing scalar or partial scalar invariance, we predict drinking problems with college alcohol beliefs using five approaches: (a) PA, (b) FS-PA, (c) JSEM, and (d) 2S-PA with regression scores, and (e) 2S-PA with Bartlett scores. With (a), we model the relationship between CLASS and AUDIT with their sum scores and ethnicity as a covariate. Sum score PA does not account for measurement noninvariance nor unreliability. With (b), we first obtain the regression factor scores of CLASS and AUDIT from a multigroup CFA. We then use the regression factor scores in a path model with ethnicity as a covariate. Measurement noninvariance, if identified, is adjusted in the first step, whereas measurement unreliability is not accounted for in FS-PA. With (c), we perform multiple-group SEM that includes a structural path between the two latent factors, with scalar or partial scalar models for CLASS and AUDIT. Thus, JSEM accounts for both measurement unreliability and noninvariance in one model. With (d) and (e), in the first stage, we obtain the factor scores from the scalar or partial scalar models and compute the reliability of the factor scores as shown in Table 1. Partial invariance is accounted for in the first stage. In the second stage, we treated the factor scores as indicators of the latent variables with known reliability. We compare the standardized path coefficients using the grand *SD* among the five approaches.

Details of the measurement invariance test results are provided in the supplemental materials. We replicated the findings of Lui (2019) for CLASS and concluded with a partial scalar model by freeing 10

[3] As reported in Lui (2019), items 4 to 10 of AUDIT measure drinking problems; items 4, 6, and 8 are on a scale of 0-4, items 5 and 7 are on a scale of 0-3, and items 9 and 10 consists of three response categories (0, 2, and 4).

intercept equality constraints across four ethnic groups (European American, Asian American, African American, Latinx American). For AUDIT, we first established partial metric invariance by freeing four loading equality constraints and concluded with a partial scalar model by additionally freeing four intercept equality constraints. The reliability of the composite and factor scores was similarly high for CLASS ($\tilde{\rho}$ = .92, .92, .87, .91) and satisfactory for AUDIT ($\tilde{\rho}$ = .79, .79, .87, .78), using formulas from Table 1.

Consistent with the results in Lui (2019), we found that higher college alcohol beliefs predicted more drinking problems in all three approaches (all ps $< .001$). Among the five approaches, FS-PA yielded the smallest standardized coefficient of AUDIT on CLASS ($\hat{\beta} = 0.49$, 95% CI [0.44, 0.54]), followed by sum-score PA ($\hat{\beta} = 0.54$, 95% CI [0.49, 0.58]). 2S-PA with regression scores ($\hat{\beta} = 0.59$, 95% CI [0.53, 0.65]) and 2S-PA with Bartlett scores ($\hat{\beta} = 0.59$, 95% CI [0.52, 0.65]) resulted in a similar standardized path coefficient as JSEM ($\hat{\beta} = 0.60$, 95% CI [0.54, 0.65]).

As shown in this example, consistent with our simulation results, using composite or factor scores without adjusting for unreliability resulted in a smaller standardized path coefficient. On the other hand, both 2S-PA and JSEM yielded a larger coefficient as well as wider CIs.

Discussion

In behavioral sciences, measured variables are prone to random and systematic errors. To account for these errors, the methodological literature generally regards joint modeling of measurement and structural models as the gold standard. While joint modeling is flexible, it is not always the most convenient for applied researchers, who usually treat construct operationalization and statistical analyses as two separate processes. Furthermore, joint modeling usually means dealing with many variables simultaneously, even when researchers have a relatively simple conceptual model, which presents many computational and practical challenges. As a result, while joint modeling is a gold standard *in theory*, applied researchers still use composite scores when analyzing their conceptual models *in practice*.

A salient example of the above problem, which is also the focus of the current paper, can be found in analyses involving composite scores that are potentially noninvariant across groups. While methodological guidelines are clear that joint modeling should be used if measures show only partial invariance across groups, from our observation and a small literature review, applied researchers continue to use composite scores following measurement invariance analysis. However, as is well known in the methodological literature, using composite scores ignores random and systematic errors and thus leads to biased parameter estimates and invalid inferences.

As an alternative to joint SEM modeling, we suggest that researchers use 2S-PA to analyze their conceptual models by obtaining factor scores and then adjusting for measurement errors using estimates of observation-specific reliability of those factor scores. We recommend using 2S-PA with factor scores over JSEM in analysis with discrete indicators, moderate sample size ($< 1,000$), and moderate reliability of the factor scores (similar to the values in our Study 2). For analysis with continuous indicators, we recommend using 2S-PA with sum scores when the sample size is small (e.g., < 400 per group) and when the composite reliability is low (e.g., $< .70$ in any groups). Results of two simulation studies show that 2S-PA gives comparable estimates as JSEM in relatively simple models and large sample sizes, has better control of Type I error rates, and has substantially fewer convergence problems in complex models with categorical indicators. While the most complex model in our studies only has three latent variables, we expect the advantage of 2S-PA over joint modeling to be even more striking for models with more latent variables.

Although the current paper focuses solely on applying 2S-PA for adjusted inferences following multiple-group measurement invariance analyses, we also want to acknowledge other developed two-stage approaches that tackle similar problems. For example, when all indicators are continuous with homogeneous measurement error variances within a group, the within-group reliability of composite or factor scores is constant. One can thus use a multiple-group version of the reliability adjustment method discussed in Hsiao et al. (2018) and Savalei (2019) in any SEM software without constraint/definition variables, which is similar to 2S-PA with composite scores in Study 1 but uses a multi-group model. Another promising line of research is the Structural After Measurement (SAM) approach (Rosseel & Loh, 2021). With SAM, one obtains measurement parameter estimates (e.g., loadings and intercepts, instead of factor scores) from separate measurement models of the latent constructs and uses those measurement parameters to obtain corrected estimates of structural coefficients. It subsumes two-stage methods such as factor score regression and path analysis with Croon (2002)'s corrections and was recently added to the R package *lavaan*. At the time of writing, however, SAM supports neither equality constraints of structural coefficients across groups nor analyses with categorical indicators, so we could not include it for comparisons in our simulation studies. As 2S-PA, SAM, and other two-stage methods continue to evolve, future research can compare and integrate these approaches.

When using 2S-PA and other two-stage estimation methods, one consideration is whether one can obtain factor scores in separate measurement models for different constructs in the structural model. In the current paper, as in M. H. C. Lai and Hsiao (2021), we assume that the indicators follow an independent cluster structure, meaning that each indicator is directly associated with only one latent construct, which allows us to separate the measurement models into chunks. When there are cross-loadings or unique covariances between indicators of different constructs, the separation strategy is more robust as it reduces the influence of omitting these crossed paths on the structural parameter estimation, compared to joint modeling that omits these crossed paths (M. H. C. Lai & Hsiao, 2021). However, neither the separation strategy nor omitting the cross paths in JSEM gives consistent structural parameter estimates (Hayes & Usami, 2020); instead, a theoretically valid approach is to use a JSEM model that correctly specifies the cross-loadings and unique covariances. An extension of 2S-PA for handling cross paths in measurement models would obtain factor scores from models with multiple latent constructs. In addition to computing case-specific reliability estimates, one also needs the case-specific loadings and covariances of the factor scores, and in the second stage, the factor scores are treated as indicators of the latent constructs but with loadings and error covariances constrained based on the values obtained in the first stage. Such an approach can be further explored in future studies.

The current paper also shows that obtaining standardized coefficients for analyses involving multiple samples or subgroups is not trivial. When researchers use multiple-group analyses, popular SEM software such as *OpenMx*, *Mplus*, and *lavaan* performs standardization using the group-specific *SD*s. However, researchers can also use single-group analyses on the pooled data with dummy-coded grouping variables for group membership, as is the case in the 2S-PA methods we examined in this paper and in multiple-indicator multiple-cause models (e.g., Bauer, 2017). As we illustrated, the grand standard deviation is typically used to obtain standardized coefficients with the single-group approach, which is not comparable to those in multiple-group analyses. In our opinion, grand standardization is more appropriate as it preserves ordering and equality constraints on the unstandardized coefficients; standardization using group-specific *SD*s generally leads to unequal coefficients even when the path coefficients are constrained to equal in the model. An alternative is to use the pooled within-group *SD*, which also preserves ordering and equality constraints as each coefficient is scaled by the same number across groups.[4] Both applied and methodological work should be mindful that different analytic approaches and standardization strategies may yield incomparable coefficients across studies, and future research can further explore the pros and cons of different standardization options.

Given that 2S-PA is relatively new, many opportunities exist to address its current limitations in future studies. We highlight a few major ones here. First, in the current implementation of 2S-PA, the second-stage likelihood function assumes that the measurement error of the factor scores is normally distributed. Such an assumption holds when normality is assumed in the measurement models, as in factor analysis assuming normality; however, the sampling distribution of factor scores only approaches normality in large samples for measurement models with categorical indicators. Even though the current simulation

[4] This is commonly done when computing Cohen's *d* effect size.

results show 2S-PA to still perform reasonably well in small samples with categorical indicators, future research can (a) investigate situations with more complex first-stage measurement models, which may take larger samples to achieve asymptotic normality, and (b) extend the likelihood functions in the second stage of 2S-PA to accommodate nonnormality. One specific direction is to examine the performance of robust standard errors (e.g., with sandwich estimators or resampling methods; see K. Lai, 2019, for an overview).

Second, while our simulation Study 2 only focused on expected a posteriori factor scores, future research should explore the performance of 2S-PA with other types of factor scores for categorical indicators (e.g., maximum a posteriori scores, maximum likelihood estimates, etc; see Estabrook & Neale, 2013). Based on the theory of 2S-PA, the estimated scores should be consistent estimates for the latent variables, have an approximately normal sampling distribution, and have consistent estimates of sampling variability available. Third, although 2S-PA uses a simplified structural model, users still need to specify the required constraints to set the reliability of factor scores and obtain standardized coefficients. We are currently working on providing R scripts to automate some of these steps. Fourth, future research can extend 2S-PA to models researchers routinely use, such as models with latent interactions and multilevel models. Finally, for the second-stage estimation, alternative estimators, such as Bayesian, least squares, and generalized method of moments estimators, can be explored.

In conclusion, the current paper shows how researchers can account for measurement quality—both measurement invariance and measurement reliability—using two-stage path analysis with each construct operationalized by a factor score variable. We show that two-stage path analysis can be a viable option, especially in small samples or when the number of measurement indicators is too big to deal with practically. While it is good to see more empirical research reporting on measurement invariance and reliability, we recommend researchers take the necessary next step: incorporate both partial invariance and unreliability in their main statistical analyses to obtain more valid results.

References

Bartlett, M. S. (1937). The statistical conception of mental scores. *British Journal of Psychology. General Section*, *28*(1), 97–104. https://doi.org/10.1111/j.2044-8295.1937.tb00863.x

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), 507–526. https://doi.org/10.1037/met0000077

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley. https://doi.org/10.1002/9781118619179

Bollen, K. A., & Maydeu-Olivares, A. (2007). A polychoric instrumental variable (PIV) estimator for structural equation models with categorical variables. *Psychometrika*, *72*(3), 309–326. https://doi.org/10.1007/s11336-007-9006-3

Burt, R. S. (1976). Interpretational confounding of unobserved variables in structural equation models. *Sociological Methods & Research*, *5*(1), 3–52. https://doi.org/10.1177/004912417600500101

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective* (2nd ed.). Chapman & Hall/CRC. https://doi.org/10.1201/9781420010138

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*(2), 300–315. https://doi.org/10.1037/a0033805

Croon, M. (2002). Using predicted latent scores in general latent structure models. In G. Marcoulides & I. Moustaki (Eds.), *Latent Variable and Latent Structure Models* (pp. 195–224). Lawrence Erlbaum.

Curran, P. J., Cole, V. T., Bauer, D. J., Rothenberg, W. A., & Hussong, A. M. (2018). Recovering predictor–criterion relations using covariate-informed factor score estimates. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(6), 860–875. https://doi.org/10.1080/10705511.2018.1473773

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, *14*(2), 81–100. https://doi.org/10.1037/a0015914

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement*, *76*(5), 741–770. https://doi.org/10.1177/0013164415607618

Devlieger, I., & Rosseel, Y. (2017). Factor score path analysis: An alternative for SEM? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *13*, 31–38. https://doi.org/10.1027/1614-2241/a000130

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates Publishers.

Estabrook, R., & Neale, M. (2013). A comparison of factor score estimation methods in the presence of missing data: Reliability and an application to nicotine dependence. *Multivariate Behavioral Research*, *48*(1), 1–27. https://doi.org/10.1080/00273171.2012.730072

Fan, X. (2003). Two approaches for correcting correlation attenuation caused by measurement error: Implications for research practice. *Educational and Psychological Measurement*, *63*(6), 915–930. https://doi.org/10.1177/0013164403251319

Hayes, T., & Usami, S. (2020). Factor score regression in the presence of correlated unique factors. *Educational and Psychological Measurement*, *80*(1), 5–40. https://doi.org/10.1177/0013164419854492

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3-4), 117–144. https://doi.org/10.1080/03610739208253916

Hoshino, T., & Bentler, P. M. (2013). Bias in factor score regression and a simple solution. In *Analysis of Mixed Data*. Chapman and Hall/CRC.

Hsiao, Y.-Y., Kwok, O.-M., & Lai, M. H. C. (2018). Evaluation of two methods for modeling measurement errors when testing interaction effects with observed composite scores. *Educational and Psychological Measurement*, *78*(2), 181–202. https://doi.org/10.1177/0013164416679877

Hsiao, Y.-Y., Kwok, O.-M., & Lai, M. H. C. (2021). Modeling measurement errors of the exogenous composites from congeneric measures in interaction models. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(2), 250–260. https://doi.org/10.1080/10705511.2020.1782206

Hsiao, Y.-Y., & Lai, M. H. C. (2018). The impact of partial measurement invariance on testing moderation for single and multi-level data. *Frontiers in Psychology*, *9*, Article 740. https://doi.org/10.3389/fpsyg.2018.00740

Kelcey, B. (2019). A robust alternative estimator for small to moderate sample SEM: Bias-corrected factor score path analysis. *Addictive Behaviors*, *94*, 83–98. https://doi.org/10.1016/j.addbeh.2018.10.032

Kwok, O.-M., Im, M., Hughes, J. N., Wehrly, S. E., & West, S. G. (2016). Testing statistical moderation in research on home–school partnerships: Establishing the boundary conditions. In S. M. Sheridan & E. Moorman Kim (Eds.), *Family-School Partnerships in Context* (pp. 79–107). Springer International Publishing. https://doi.org/10.1007/978-3-319-19228-4_5

Lai, K. (2019). More robust standard error and confidence interval for SEM parameters given incorrect model and nonnormal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(2), 260–279. https://doi.org/10.1080/10705511.2018.1505522

Lai, M. H. C., & Hsiao, Y.-Y. (2021). Two-stage path analysis with definition variables: An alternative framework to account for measurement error. *Psychological Methods*. https://doi.org/10.1037/met0000410

Ledgerwood, A., & Shrout, P. E. (2011). The trade-off between accuracy and precision in latent variable models of mediation processes. *Journal of Personality and Social Psychology*, *101*(6), 1174–1188. https://doi.org/10.1037/a0024776

Levy, R. (2017). Distinguishing outcomes from indicators via Bayesian modeling. *Psychological Methods*, *22*(4), 632–648. https://doi.org/10.1037/met0000114

Lui, P. P. (2019). College alcohol beliefs: Measurement invariance, mean differences, and correlations with alcohol use outcomes across sociodemographic groups. *Journal of Counseling Psychology*, *66*(4), 487–495. https://doi.org/10.1037/cou0000338

MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin*, *109*(3), 502–511. https://doi.org/10.1037/0033-2909.109.3.502

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*(1), 99–128. https://doi.org/10.1207/s15327906mbr3901_4

McDonald, R. P. (1999). *Test theory: A unified treatment*. Psychology Press.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Meijer, E., Oczkowski, E., & Wansbeek, T. (2021). How measurement error affects inference in linear regression. *Empirical Economics*, *60*(1), 131–155. https://doi.org/10.1007/s00181-020-01942-z

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, *29*(2), 177–185. https://doi.org/10.1007/BF02289699

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Muthén & Muthén. https://www.statmodel.com

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., Estabrook, R., Bates, T. C., Maes, H. H., & Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, *81*(2), 535–549. https://doi.org/10.1007/s11336-014-9435-8

Osberg, T. M., Atkins, L., Buchholz, L., Shirshova, V., Swiantek, A., Whitley, J., Hartman, S., & Oquendo, N. (2010). Development and validation of the College Life Alcohol Salience Scale: A measure of beliefs about the role of alcohol in college life. *Psychology of Addictive Behaviors*, *24*(1), 1–12. https://doi.org/10.1037/a0018197

Pek, J., & Wu, H. (2015). Profile likelihood-based confidence intervals and regions for structural equation models. *Psychometrika*, *80*(4), 1123–1145. https://doi.org/10.1007/s11336-015-9461-1

Pritikin, J. N., Brick, T. R., & Neale, M. C. (2018). Multivariate normal maximum likelihood with both ordinal and continuous variables, and data missing at random. *Behavior Research Methods*, *50*(2), 490–500. https://doi.org/10.3758/s13428-017-1011-6

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*(2), 173–184. https://doi.org/10.1177/01466216970212006

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

Rosseel, Y. (2020). Small sample solutions for structural equation modeling. In R. van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners* (pp. 226–238). Routledge. https://doi.org/10.4324/9780429273872-19

Rosseel, Y., & Loh, W. W. (2021). *The "Structural-After-Measurement" (SAM) approach to SEM.* PsyArXiv. Retrieved November 19, 2021, from https://osf.io/pekbm/

Saunders, J. B., Aasland, O. G., Babor, T. F., De La Fuente, J. R., & Grant, M. (1993). Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption: II. *Addiction*, *88*(6), 791–804. https://doi.org/10.1111/j.1360-0443.1993.tb02093.x

Savalei, V. (2019). A comparison of several approaches for controlling measurement error in small samples. *Psychological Methods*, *24*(3), 352–370. https://doi.org/10.1037/met0000181

Thomson, G. H. (1935). Definition and measurement of general intelligence. *Nature*, *135*(3413), 509–509. https://doi.org/10.1038/135509b0

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research.* (pp. 281–324). American Psychological Association. https://doi.org/10.1037/10222-009

Yoon, M., & Kim, E. S. (2014). A comparison of sequential and nonsequential specification searches in testing factorial invariance. *Behavior Research Methods*, *46*(4), 1199–1206. https://doi.org/10.3758/s13428-013-0430-2

Table 1

Three Types of Estimated Scores and the Corresponding Reliability.

Estimated scores	Scoring matrix ($\mathbf{A}_g$)	Loading on latent variable (λ_g^*)	Var($\tilde{\eta}$)	Reliability ($\rho_{\tilde{\eta}}$)
Regression	$\psi_g \boldsymbol{\lambda}_g^\top \boldsymbol{\Sigma}_g^{-1}$	$\psi_g \boldsymbol{\lambda}_g^\top \boldsymbol{\Sigma}_{yg}^{-1} \boldsymbol{\lambda}_g$	$\psi_g^2 \boldsymbol{\lambda}_g^\top \boldsymbol{\Sigma}_{yg}^{-1} \lambda_g$	$\psi_g \boldsymbol{\lambda}_g^\top \boldsymbol{\Sigma}_g^{-1} \boldsymbol{\lambda}_g$
Bartlett	$(\boldsymbol{\lambda}_g^\top \boldsymbol{\Theta}_g^{-1} \boldsymbol{\lambda}_g)^{-1} \boldsymbol{\lambda}_g^\top \boldsymbol{\Theta}_g^{-1}$	1	$\psi + (\boldsymbol{\lambda}_g^\top \boldsymbol{\Theta}_g^{-1} \boldsymbol{\lambda}_g)^{-1}$	$\frac{\psi}{\psi + (\boldsymbol{\lambda}_g^\top \boldsymbol{\Theta}_g^{-1} \boldsymbol{\lambda}_g)^{-1}}$
Sum score	$\mathbf{1}^\top$	$\mathbf{1}^\top \boldsymbol{\lambda}_g$	$(\mathbf{1}^\top \boldsymbol{\lambda}_g)^2 \psi + \mathbf{1}^\top \Theta_g \mathbf{1}$	$\frac{(\mathbf{1}^\top \boldsymbol{\lambda}_g)^2 \psi}{(\mathbf{1}^\top \boldsymbol{\lambda}_g)^2 \psi + \mathbf{1}^\top \Theta_g \mathbf{1}}$

Table 2

Results of Study 1 (Bias and RMSE)

			Bias							RMSE						
β_1	ω	n_1, n_2	PA	FS-PA	Croon	JSEM	2S-PA$_1$	2S-PA$_2$	2S-PA$_3$	PA	FS-PA	Croon	JSEM	2S-PA$_1$	2S-PA$_2$	2S-PA$_3$
0	.49, .61	50, 30	-0.001	-0.001	0.000	-0.001	**0.000**	-0.001	0.000	**0.115**	0.118	0.139	0.163	0.143	0.148	0.152
		100, 60	**0.004**	0.005	0.005	0.006	0.007	0.006	0.006	0.079	**0.079**	0.096	0.103	0.098	0.101	0.103
		500, 300	**0.001**	0.002	0.002	0.002	0.002	0.002	0.002	**0.035**	0.035	0.044	0.044	0.044	0.045	0.045
		1000, 600	**0.000**	0.000	0.000	0.001	0.001	0.001	0.000	**0.025**	0.025	0.032	0.032	0.032	0.032	0.033
	.71, .77	50, 30	-0.003	-0.005	-0.005	-0.006	-0.005	-0.005	-0.004	**0.115**	0.116	0.128	0.137	0.130	0.133	0.133
		100, 60	**0.005**	0.006	0.006	0.007	0.007	0.007	0.006	**0.080**	0.080	0.090	0.092	0.090	0.092	0.092
		500, 300	**0.001**	0.002	0.002	0.002	0.002	0.002	0.002	0.035	**0.035**	0.039	0.040	0.040	0.040	0.040
		1000, 600	**0.000**	0.000	0.000	0.000	0.000	0.000	0.000	0.025	**0.025**	0.028	0.028	0.028	0.029	0.029
0.54	.49, .61	50, 30	-0.122	-0.151	-0.087	0.015	-0.073	-0.061	-0.010	0.166	0.197	0.172	0.773	0.196	0.178	**0.146**
		100, 60	-0.119	-0.129	-0.043	-0.001	-0.033	-0.019	-0.001	0.142	0.153	0.107	0.102	0.107	0.112	**0.098**
		500, 300	-0.120	-0.121	-0.018	-0.005	-0.013	-0.002	0.003	0.125	0.126	0.047	0.043	0.044	0.046	**0.042**
		1000, 600	-0.121	-0.121	-0.017	-0.007	-0.012	-0.001	0.002	0.123	0.124	0.035	0.032	0.033	0.033	**0.031**
	.71, .77	50, 30	-0.076	-0.082	-0.035	-0.010	-0.030	-0.020	-0.009	0.134	0.141	0.130	0.131	0.131	0.134	**0.125**
		100, 60	-0.065	-0.066	-0.011	**0.002**	-0.007	0.002	0.004	0.102	0.103	0.087	0.088	**0.087**	0.090	0.087
		500, 300	-0.068	-0.066	-0.005	-0.002	-0.003	0.004	0.002	0.076	0.074	0.038	**0.037**	0.037	0.039	0.037
		1000, 600	-0.068	-0.066	-0.005	-0.003	-0.003	0.004	**0.002**	0.073	0.070	0.027	**0.027**	0.027	0.028	0.027
0	.49, .61	40, 40	-0.002	0.000	0.000	0.000	-0.002	0.000	0.000	**0.114**	0.117	0.136	0.157	0.153	0.144	0.145
		80, 80	**0.003**	0.003	0.004	0.004	0.004	0.003	0.003	**0.081**	0.081	0.098	0.104	0.103	0.102	0.103
		400, 400	**0.002**	0.002	0.002	0.002	0.002	0.002	0.002	**0.035**	0.036	0.044	0.044	0.044	0.045	0.045
		800, 800	0.000	**0.000**	0.000	0.000	0.000	0.000	0.001	**0.025**	0.025	0.031	0.031	0.031	0.032	0.032
	.71, .77	40, 40	-0.005	-0.005	-0.005	-0.006	-0.005	-0.005	-0.006	**0.113**	0.115	0.126	0.134	0.128	0.130	0.130
		80, 80	**0.004**	0.004	0.005	0.005	0.005	0.005	0.005	**0.080**	0.080	0.089	0.091	0.089	0.091	0.091
		400, 400	**0.001**	0.001	0.001	0.001	0.001	0.001	0.001	0.035	**0.035**	0.040	0.040	0.040	0.040	0.040
		800, 800	**0.000**	0.000	0.000	0.000	0.000	0.000	0.000	0.025	**0.025**	0.028	0.028	0.028	0.028	0.029
0.56	.49, .61	40, 40	-0.120	-0.143	-0.082	-0.006	-0.064	-0.056	-0.018	0.163	0.186	0.160	0.150	0.163	0.165	**0.149**
		80, 80	-0.114	-0.123	-0.039	0.000	-0.027	-0.017	-0.003	0.140	0.148	0.106	0.101	0.112	0.109	**0.098**
		400, 400	-0.115	-0.114	-0.016	-0.004	-0.011	**0.000**	0.002	0.120	0.120	0.045	0.043	0.044	0.046	**0.042**
		800, 800	-0.115	-0.114	-0.014	-0.005	-0.009	**0.001**	0.002	0.118	0.117	0.033	0.030	0.031	0.032	**0.030**
	.71, .77	40, 40	-0.075	-0.080	-0.036	-0.011	-0.030	-0.021	-0.012	0.134	0.138	0.127	0.127	0.128	0.130	**0.123**
		80, 80	-0.063	-0.064	-0.011	0.002	-0.007	**0.001**	0.003	0.101	0.102	0.087	0.087	0.087	0.089	**0.087**
		400, 400	-0.065	-0.063	-0.005	-0.002	-0.003	0.004	0.002	0.074	0.071	0.038	**0.038**	0.038	0.039	0.038
		800, 800	-0.066	-0.063	-0.004	-0.002	-0.003	0.004	**0.002**	0.070	0.067	0.027	**0.027**	0.027	0.028	0.027

Note. The best performing method for each condition was indicated in bold fonts. 2S-PA$_1$ = two-stage path analysis with regression factor scores. 2S-PA$_2$ = two-stage path analysis with Bartlett factor scores. 2S-PA$_3$ = two-stage path analysis with sum scores.

Table 3

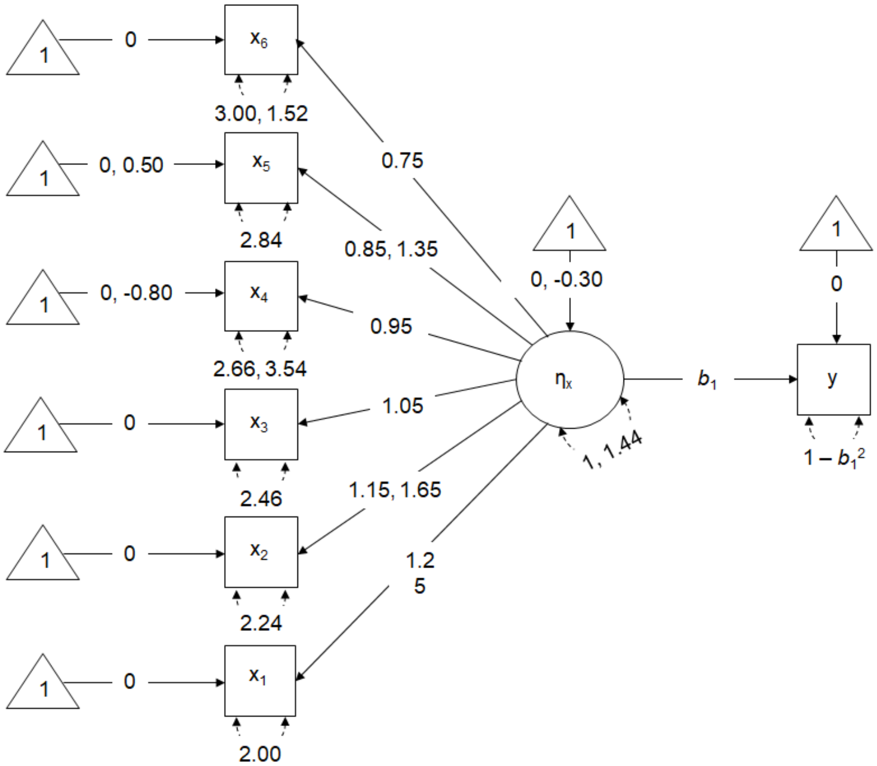
Results of Study 1 (SE Bias and Coverage)

			Relative SE Bias (%)							Coverage (%)						
β_1	ω	n_1, n_2	PA	FS-PA	Croon	JSEM	2S-PA$_1$	2S-PA$_2$	2S-PA$_3$	PA	FS-PA	Croon	JSEM	2S-PA$_1$	2S-PA$_2$	2S-PA$_3$
0	.49, .61	50, 30	-2.99	-3.53	-2.71	*-13.00*	-4.08	-4.89	-3.01	94.1	94.2	94.6	*90.72*	94.2	94.2	94.0
		100, 60	0.38	1.78	2.22	-2.37	1.64	0.93	0.17	95.0	95.2	95.1	94.0	95.2	95.2	94.9
		500, 300	2.22	2.60	2.06	1.92	2.51	2.47	2.06	94.8	94.7	94.6	94.4	94.5	94.5	95.0
		1000, 600	0.27	0.49	-0.16	0.01	0.50	0.42	0.32	95.3	95.6	95.4	95.4	95.5	95.5	95.4
	.71, .77	50, 30	-2.27	-3.02	-1.13	-7.30	-3.25	-3.80	-2.68	94.5	93.8	94.5	*92.08*	93.8	93.8	94.4
		100, 60	-0.13	0.03	0.69	-1.68	-0.08	-0.39	-0.35	94.9	94.8	95.1	94.3	94.9	94.9	95.1
		500, 300	1.81	2.25	2.35	1.92	2.21	2.17	1.84	95.0	95.0	95.0	94.8	94.9	94.8	95.0
		1000, 600	0.75	0.95	1.01	0.70	0.93	0.88	0.72	95.2	95.6	95.4	95.5	95.6	95.6	95.1
0.54	.49, .61	50, 30	-3.06	*-10.75*	*-13.14*	*-71.91*	*-25.83*	*-20.66*	-3.92	*81.04*	*73.20*	*85.36*	*90.60*	*87.64*	*85.40*	94.0
		100, 60	-1.32	-3.19	-5.49	-5.49	-5.03	*-13.17*	0.85	*69.12*	*64.12*	*90.92*	94.2	92.5	*91.16*	95.2
		500, 300	2.28	0.55	-2.14	0.91	2.87	-6.55	6.05	*7.12*	*7.16*	*92.36*	94.9	94.9	93.7	96.4
		1000, 600	-1.02	-2.37	-4.63	-2.19	-0.19	-8.31	2.80	*0.28*	*0.32*	*90.12*	94.0	93.7	92.9	95.3
	.71, .77	50, 30	-2.75	-5.34	-7.57	-7.89	-4.87	*-12.50*	-1.11	*89.40*	*87.92*	92.7	92.7	93.7	*91.92*	95.6
		100, 60	-2.12	-2.59	-5.62	-3.06	-0.61	-8.37	0.65	*85.96*	*86.64*	93.4	94.1	94.5	92.8	94.8
		500, 300	2.11	1.84	-1.85	2.02	4.47	-3.70	5.36	*51.08*	*53.92*	94.0	95.4	96.0	94.4	96.4
		1000, 600	-0.06	-0.50	-3.88	-0.21	1.83	-5.92	2.97	*21.72*	*25.04*	94.2	95.3	95.6	93.4	95.7
0	.49, .61	40, 40	-2.39	-3.16	-1.78	*-11.26*	-9.16	-4.14	-2.35	94.3	94.1	94.8	*91.36*	94.0	93.9	94.8
		80, 80	-1.68	-1.39	-1.27	-4.98	-3.58	-1.86	-1.92	94.8	95.1	95.1	93.6	95.1	95.1	94.9
		400, 400	0.71	1.01	0.50	0.32	0.93	0.86	0.62	95.5	95.2	95.0	95.0	95.1	95.0	95.3
		800, 800	1.28	1.40	1.07	1.07	1.43	1.42	1.20	95.8	95.5	95.8	95.4	95.5	95.6	95.7
	.71, .77	40, 40	-1.27	-1.95	-0.03	-5.71	-2.05	-2.58	-1.84	95.0	94.3	95.0	93.3	94.3	94.2	94.6
		80, 80	-0.58	-0.35	0.51	-1.93	-0.33	-0.57	-0.54	94.7	95.1	95.2	94.6	95.2	95.3	94.8
		400, 400	0.87	1.10	1.20	0.73	1.09	1.05	0.87	95.5	95.8	95.5	95.4	95.7	95.7	95.6
		800, 800	0.32	0.88	0.99	0.68	0.88	0.88	0.33	95.2	95.4	95.4	95.3	95.4	95.4	95.2
0.56	.49, .61	40, 40	-0.99	-5.43	-8.17	-9.51	-9.84	*-16.14*	-7.49	*81.68*	*75.28*	*87.04*	*92.40*	*89.64*	*87.44*	94.6
		80, 80	-3.68	-4.96	-7.88	-5.69	*-10.65*	*-13.27*	-0.41	*69.64*	*67.08*	*91.24*	93.4	92.6	*91.48*	94.6
		400, 400	0.37	-0.89	-3.74	-1.16	0.47	-8.12	3.21	*10.08*	*10.92*	*92.28*	94.5	94.6	93.0	96.0
		800, 800	0.95	-0.58	-3.06	0.18	2.05	-6.34	4.82	*0.64*	*0.60*	*92.20*	94.8	94.9	93.6	95.9
	.71, .77	40, 40	-2.29	-3.66	-6.21	-5.44	-3.08	*-10.43*	-0.40	*89.08*	*88.28*	92.6	92.8	93.7	*92.00*	94.7
		80, 80	-2.54	-3.24	-6.52	-3.18	-0.96	-8.79	0.37	*86.56*	*86.36*	93.3	94.6	94.4	92.5	95.5
		400, 400	0.87	0.70	-3.31	0.77	3.15	-4.74	4.03	*55.00*	*58.36*	93.8	95.1	95.6	93.9	96.0
		800, 800	-0.49	-0.32	-4.17	-0.10	2.19	-5.65	2.47	*25.44*	*28.60*	93.8	94.9	95.7	93.8	95.4

Note. Supotimal values are indicated in italic fonts (|RSB| > 10% or coverage < 92.5%). 2S-PA$_1$ = two-stage path analysis with regression factor scores. 2S-PA$_2$ = two-stage path analysis with Bartlett factor scores. 2S-PA$_3$ = two-stage path analysis with sum scores.

Figure 1

Data generating model for Study 1.



Note. Group-specific parameter values are separated by a comma. The loading values shown in the graph are for the moderate-reliability conditions; they were .45 to .95 for Group 1 in the low-reliability conditions.

Figure 2

Data generating model for Study 2.

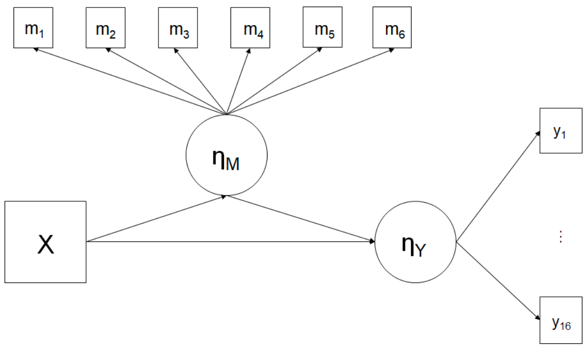
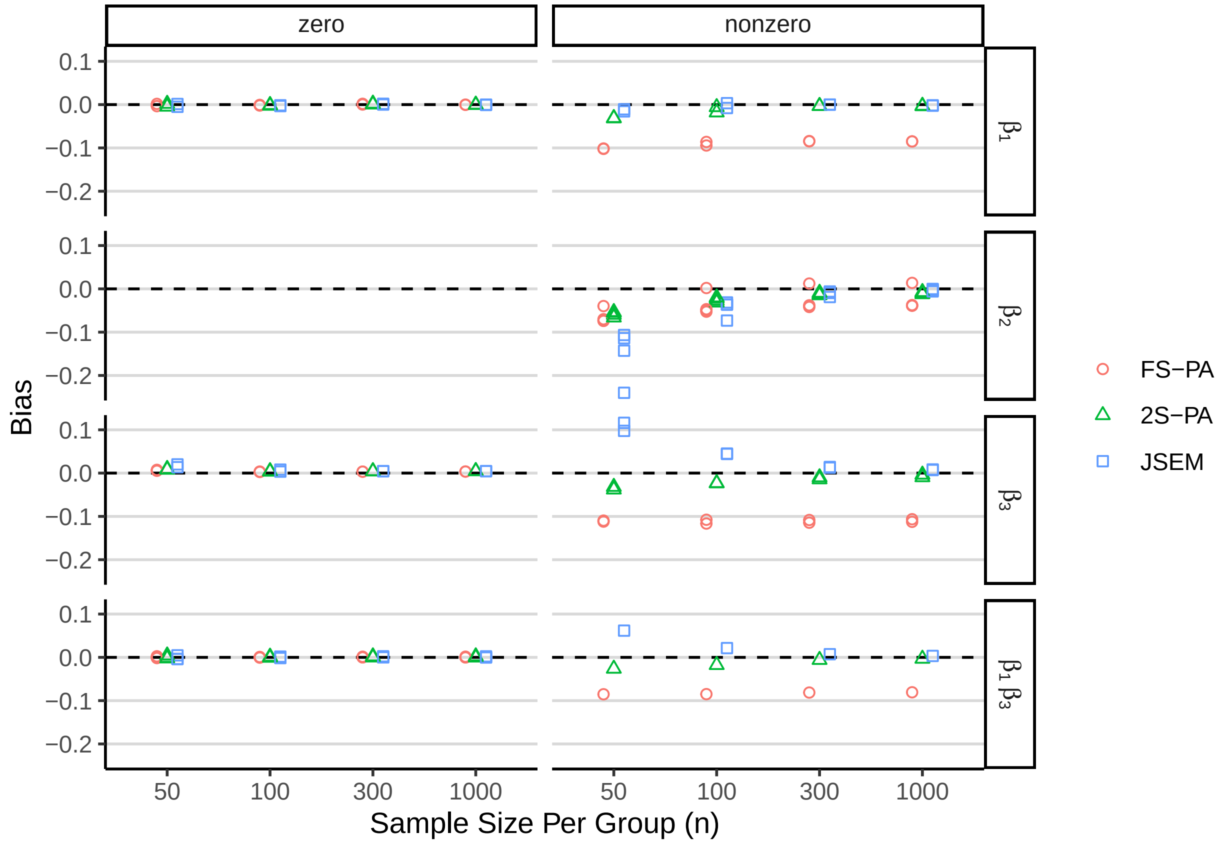


Figure 3

Bias in Parameter Estimates for Study 2



Note. Points represent values for all simulation conditions.

Figure 4

Root Mean Squared Error (RMSE) of Parameter Estimates for Study 2

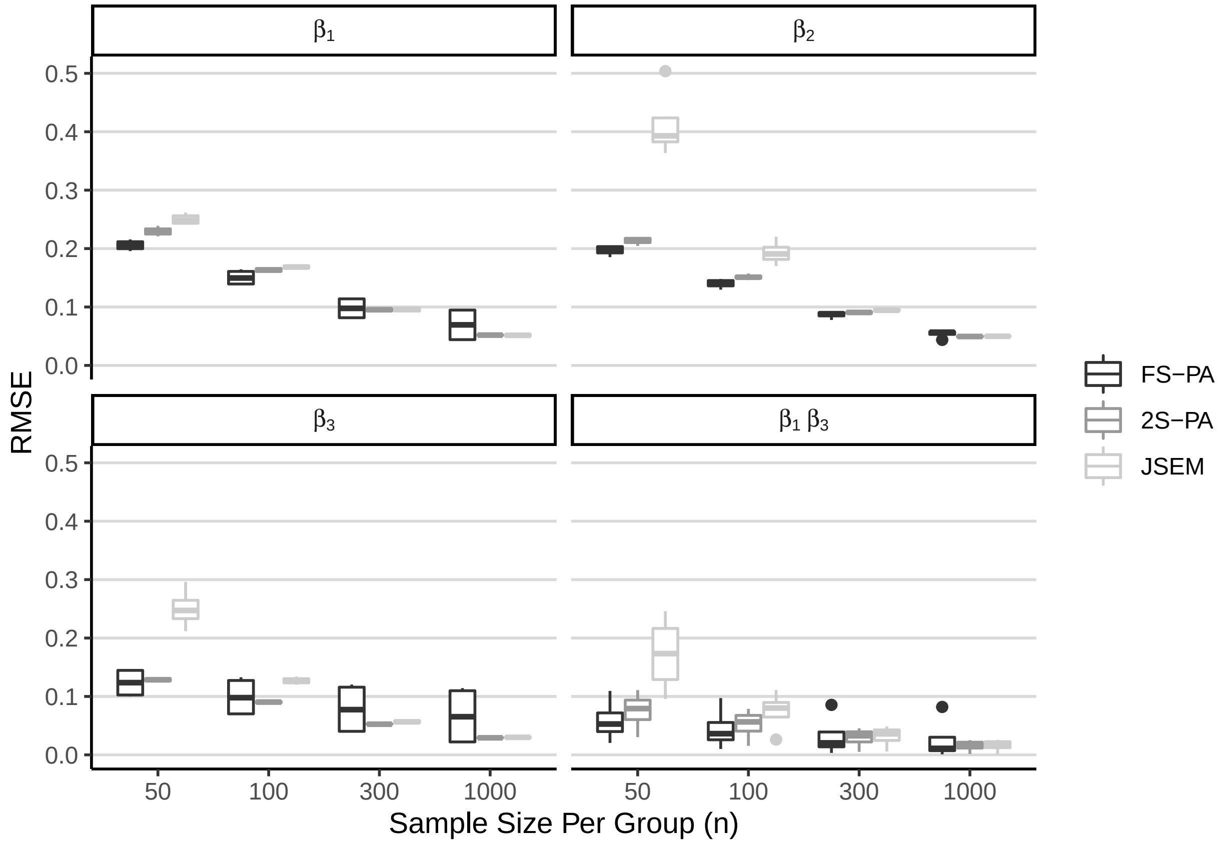
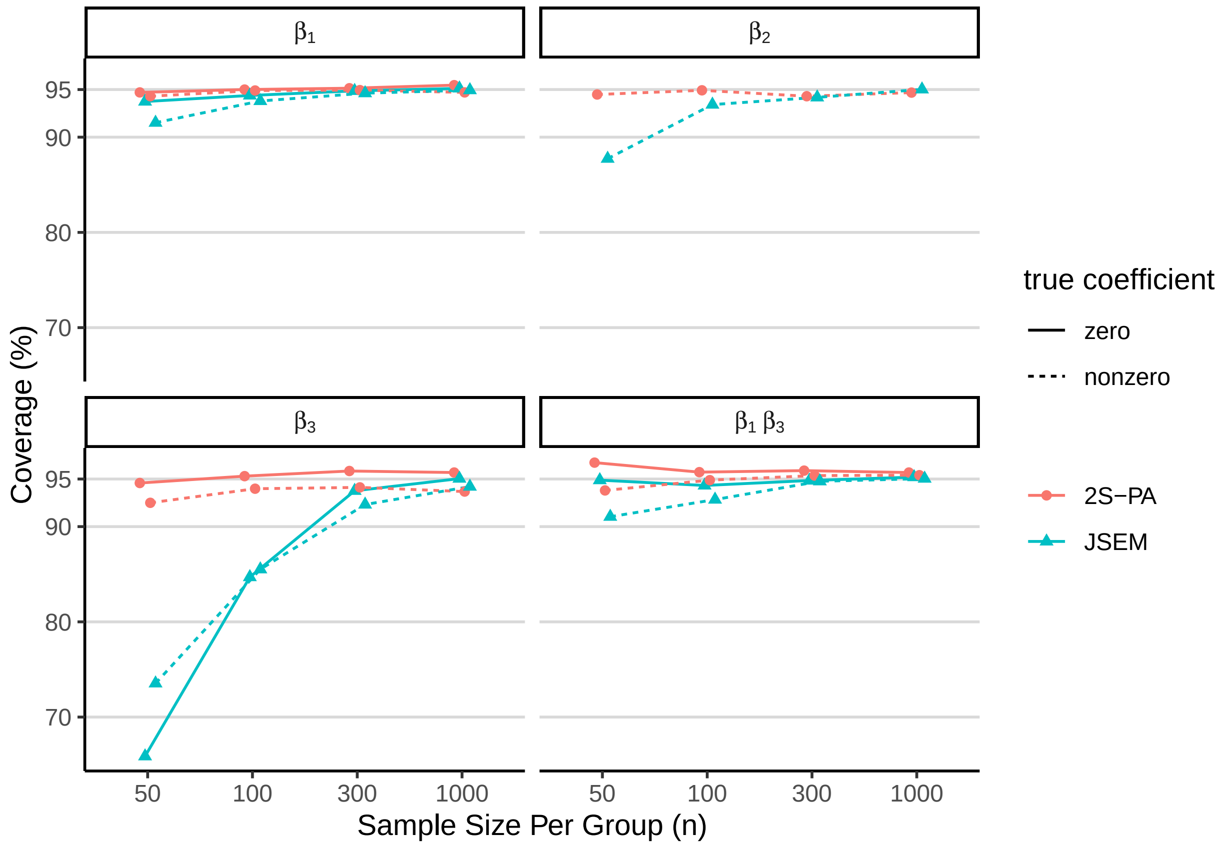


Figure 5

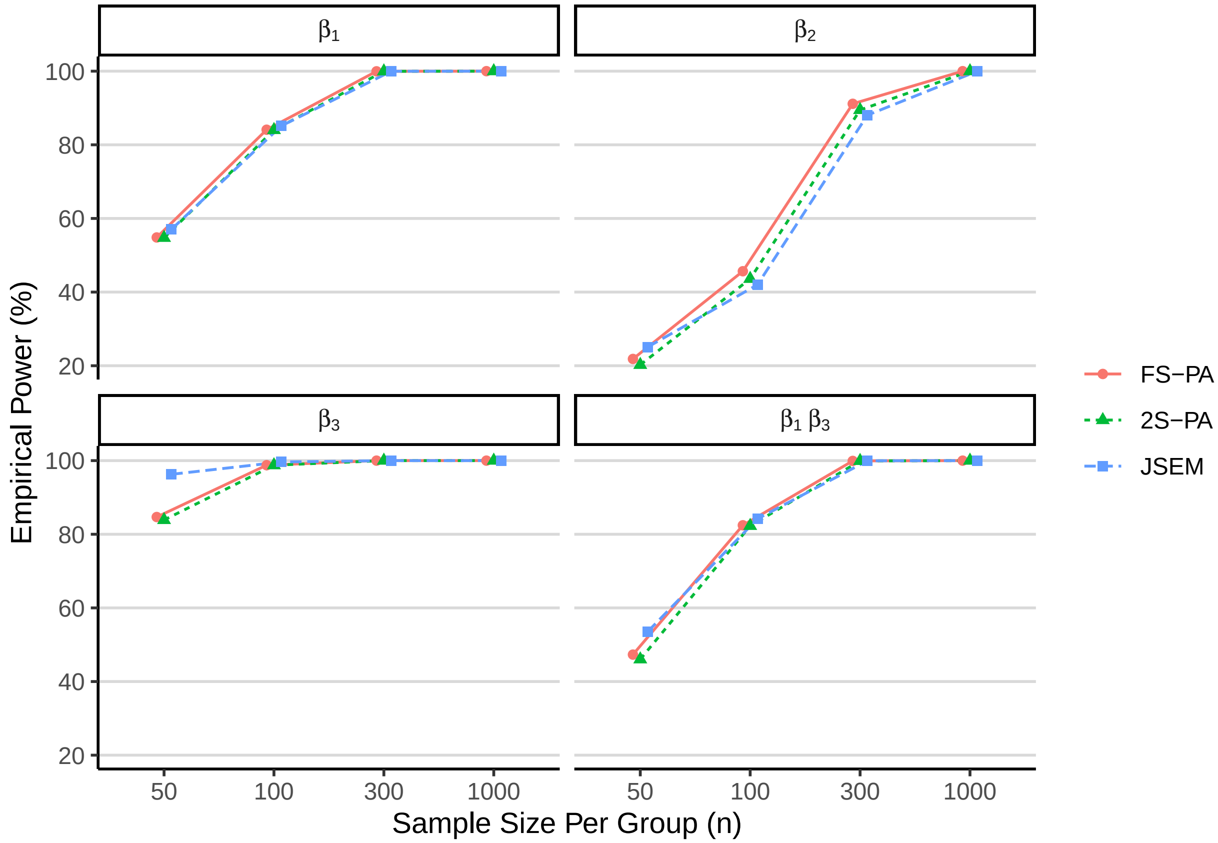
Coverage of 95% Confidence Intervals for Study 2



Note. The points for β_2 represent median values across conditions. The empirical Type I error rates can be obtained as 1 - coverage rate when the true coefficient is zero.

Figure 6

Empirical Power for Study 2



Note. The points for β_2 represent median values across conditions.