Two-Stage Path Analysis With Definition Variables: An Alternative Framework to Account for Measurement Error

Mark H. C. Lai[1] & Yu-Yu Hsiao[2]

[1] Department of Psychology, University of Southern California

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

Author Note

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

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

Yu-Yu Hsiao was supported by the National Institute on Alcohol Abuse and Alcoholism under Grant R01 AA025539. Codes for simulations and the empirical demonstration, and other supplemental materials are openly available at the project's Open Science Framework page (https://osf.io/h95vx/). We would like to thank Winnie Tse for helping with the organization of the supplemental materials, and Hio Wa Mak, Stefan Schneider, and Roy Levy for thoughtful comments on earlier versions of the manuscript.

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

Abstract

When estimating path coefficients among psychological constructs measured with error, structural equation modeling (SEM), which simultaneously estimates the measurement and structural parameters, is generally regarded as the gold standard. In practice, however, researchers usually first compute composite scores or factor scores, and use those as observed variables in a path analysis, for purposes of simplifying the model or avoiding model convergence issues. Whereas recent approaches, such as reliability adjustment methods and factor score regression, has been proposed to mitigate the bias induced by ignoring measurement error in composite/factor scores with continuous indicators, those approaches are not yet applicable to models with categorical indicators. In this paper, we introduce the two-stage path analysis (2S-PA) with definition variables as a general framework for path modeling to handle categorical indicators, in which estimation of factor scores and path coefficients are separated. It thus allows for different estimation methods in the measurement and the structural path models and easier diagnoses of violations of model assumptions. We conducted three simulation studies, ranging from latent regression to mediation analysis with categorical indicators, and showed that 2S-PA generally produced similar estimates to those using SEM in large samples, but gave better convergence rates, less standard error bias, and better control of Type I error rates in small samples. We illustrate 2S-PA using data from a national data set, and show how researchers can implement it in Mplus and OpenMx. Possible extensions and future directions of 2S-PA are discussed.

Keywords: measurement error, SEM, path analysis, reliability adjustment, item response theory, definition variable

Word count: 8,907

Two-Stage Path Analysis With Definition Variables: An Alternative Framework to Account for Measurement Error

In social and behavioral sciences, researchers are usually interested in estimating structural relations (i.e., path coefficients) among constructs that cannot be directly observed and can only be measured by noisy indicators (Kline, 2016). Traditionally, researchers have been using computed variables—such as composite scores (Hsiao et al., 2018) or factor scores (e.g., Skrondal & Laake, 2001)—as proxies of the latent constructs of interest. However, because these computed variables are generally not measurement error free, their use can result in biased estimates of structural relations (e.g., Cole & Preacher, 2014) that are usually of substantive interest to researchers. Two common approaches to reduce such bias due to measurement error are (a) full structural equation modeling (SEM; Figure 1) that simultaneously estimates measurement models for the latent constructs and a structural model specifying their relations (Jöreskog, 1970), and (b) two-step analyses that adjust the estimated path (structural) coefficients obtained using observed scores for measurement error (Devlieger et al., 2016). Whereas full SEM is generally regarded as the gold standard, in practice it usually requires a large sample size to get stable parameter estimates, especially when the numbers of latent variables and of observed variables are large (Savalei, 2019).

On the other hand, given their relative simplicity compared with full SEM, recently there has been a renewed interest in observed score regression and path analysis methods with measurement error adjustment, which are based on concepts found in much earlier literature in econometrics (e.g., Caroll et al., 2006; Reiersøl, 1950; Wansbeek & Meijer, 2000) and in SEM (Hayduk, 1987). Examples include factor score regression (Devlieger et al., 2016; Hoshino & Bentler, 2013), factor score path analysis (Devlieger & Rosseel, 2017; Kelcey, 2019), and reliability-adjustment for latent interactions (Hsiao et al., 2018) and mediation analyses (Savalei, 2019). When the assumptions of the underlying measurement models are met, these methods have been shown to produce estimates very similar to those with full SEM (Devlieger et al., 2016; Hsiao et al., 2018), have better small sample properties (Kelcey, 2019; Savalei, 2019), and be more robust to misspecifications in the measurement models (Devlieger & Rosseel, 2017).

Despite the promising results of these measurement error adjustment methods, each of them have certain limitations. Most notably, these methods assume that the observed indicators are continuous and normally distributed so that the measurement error variance for each observation is constant. In psychological measurement, however, indicators usually have discrete response options, which results in measurement error with nonconstant variance at the observed score level across different levels of the latent variable (Embretson, 1996). To address this limitation, in this paper we aim to (a) introduce the two-stage path analysis (2S-PA) with definition variables, a general framework for adjusting measurement error in regression and path analyses, (b) compare the performance of 2S-PA with observed score path analysis, full SEM, and other measurement error adjustment methods in a series of simulation studies with categorical indicators, and (c) demonstrate the use of 2S-PA in a public data set. Potential benefits and limitations of 2S-PA and possible extensions are discussed.

A Two-Stage Approach for Handling Measurement Error

Consider a general path model for the relations among a set of q constructs, represented by a variable vector $\boldsymbol{\eta}_i = [\eta_{i1}, \eta_{i2}, \dots, \eta_{iq}]^\top$ for the ith observation (i = 1, 2, …, N):

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\zeta}_i \tag{1}$$

where $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_q]^\top$ contains the regression intercepts, $\mathbf{B}$ is a $q \times q$ matrix with each element β_{mn} representing the regression coefficient of η_m regressed on η_n, and $\boldsymbol{\zeta}_i$ is a vector of length q of disturbances, with the standard assumption that $\boldsymbol{\zeta}_i = \mathbf{0}$.[1]

For simplicity, and as a common practice, in this paper we assume that the components of $\boldsymbol{\zeta}_i$ are independently and identically distributed following a multivariate normal distribution with a covariance matrix $\boldsymbol{\psi}$, and that they are independent to the exogenous components in $\boldsymbol{\eta}$. Equation (1) is commonly referred to as the *structural model* linking the constructs (ηs) of interest.

In practice, the ηs are usually unobserved, latent variables and so the parameters in the above equation cannot be directly estimated. When each η is measured by multiple observed indicators, researchers usually compute a sum score or factor score, denoted as $\tilde{\eta}$, as a single indicator to represent each η. Such practice is not uncommon, as Cole and Preacher (2014) reported that 11.7% of published articles in seven major psychology journals in 2011 involved path analysis with observed single indicators, and the prevalence would be much higher if articles using multiple regression (which is a special case of path analysis) were also included. However, researchers rarely adjust for measurement error in observed single indicators despite recommendations from the SEM literature (e.g., Bollen, 1989; Hayduk, 1987; Hsiao et al., 2018; Savalei, 2019) and also in econometrics (e.g., Murphy & Topel, 1985) and statistics (e.g., Caroll et al., 2006), which showed that ignoring measurement error led to biased structural coefficient estimates, with unpredictable bias in small samples (Loken & Gelman, 2017) and in moderately complex path models (Cole & Preacher, 2014).

[1] We follow the "all-y" notation system by Jöreskog and Sörbom (2001), except using Σ_ε later to indicate the measurement error variance of the factor scores.

In the present paper, we propose a two-stage alternative approach to full SEM by first obtaining factor scores (which include the special case of sum scores), $\tilde{\boldsymbol{\eta}}$, and the corresponding estimated *standard error of measurement* for each factor score, using appropriate psychometric analyses, and then accounts for measurement error in the second-stage analysis of factor scores using definition variables. Given space limitations we only discuss the use of the expected a posteriori (EAP) method for computing factor scores and do not compare other alternatives, but readers can get a good overview of some common factor score options in Estabrook and Neale (2013).

Specifically, the two-stage approach estimates the measurement and the structural models separately:

$$\text{Measurement: } \tilde{\boldsymbol{\eta}}_i|\boldsymbol{\omega}, \mathbf{y} \tag{2}$$

$$\text{Structural: } \begin{cases} \boldsymbol{\eta}_i = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\zeta}_i \\ \tilde{\boldsymbol{\eta}}_i = \boldsymbol{\Lambda}_i\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \\ \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\varepsilon i}) \end{cases}, \tag{3}$$

where $\tilde{\eta}_i$ is the q-vector of factor scores for the ith person obtained from a measurement model of observed item scores $\mathbf{y}$ with parameters $\boldsymbol{\omega}$, and $\boldsymbol{\Sigma}_{\varepsilon i}$ is the $q \times q$ covariance matrix of measurement error for the factor scores, typically obtained from the first stage. When separate measurement models are fitted to separate sets of items, $\boldsymbol{\Sigma}_{\varepsilon i}$ is diagonal with elements $[\sigma^2_{\varepsilon 1i}, \sigma^2_{\varepsilon 2i}, \ldots, \sigma^2_{\varepsilon qi}]$. The loading matrix $\boldsymbol{\Lambda}$ is a known diagonal matrix to standardize $\boldsymbol{\eta}$, so that elements of $\mathbf{B}$ are standardized coefficients. The above model is a special case of the broader class of multivariate nonlinear models with classical measurement error in the statistics and econometrics literature (e.g., Caroll et al., 2006; Fuller, 1987; Wansbeek & Meijer, 2000). However, instead of assuming that $\boldsymbol{\Sigma}_{\varepsilon i}$ is given, it is estimated using psychometric methods that are familiar to SEM researchers. While the above model can be estimated using maximum likelihood as discussed in Caroll et al. (2006, chapter 8); because the estimated standard error of measurement is not constant across observations, in the SEM framework it requires the use of definition variables to fix the error variance to individual-specific values.

Two-Stage Path Analysis With Definition Variables

In SEM, definition variables are "observed variables used to fix model parameters to individual specific data values" (Mehta & Neale, 2005, p. 259) and were originally developed in the Mx program (see e.g., Neale, 2000). In conventional SEM, definition variables are not needed because the model parameters, such as factor loadings, path coefficients, and the measurement error variance parameters, are assumed constant across individuals, which implies that the likelihood function for each observation is the same. This is obviously not the case for the model in equation (3), as the likelihood function depends on the standard error of measurement, $\Sigma_{\varepsilon i}$, which is not constant across observations. Using definition variables, on the other hand, allows estimation with non-identical likelihood functions across observations.

Applications of definition variables include multilevel models with random slopes (Mehta & Neale, 2005), models with heterogeneous measurement error (B. O. Muthén & Asparouhov, 2002), and meta-analysis (Cheung, 2013). A path diagram involving definition variables for a regression model of η_2 on η_1, with η_1 indicated by $\tilde{\eta}_1$ with heterogeneous error variance, is shown in Figure 2b. In the diagram, both the loading of $\tilde{\eta}_1$ on η_1, $\tilde{\lambda}_1$, and the error variance, $\tilde{\sigma}^2_{\varepsilon 1}$, are fixed as definition variables, represented in diamonds.

In the proposed two-stage path analysis (2S-PA) with definition variables, in stage 1 the factor score variables ($\tilde{\eta}$s) can be obtained with any appropriate psychometric analyses (e.g., using Figure 2a), as long as the individual-specific factor score and standard error of measurement estimates can be obtained. For example, item response models can be used for binary or ordered categorical variables using maximum likelihood with the expected a posteriori (EAP) method. When one or more indicators in $\mathbf{y}$ is categorical, the standard error of measurement generally varies across individuals (Lord, 1984, also see Appendix A for an illustration).[2]

Because latent variables generally do not have an intrinsically meaningful unit, when fitting a measurement model, it is common to set the variance of the latent variables to unity. Let $\hat{\sigma}_{\tilde{\eta}_1 i}$ be the estimated standard error of the factor score $\tilde{\eta}_1$ for person i. Then the true score variance of $\tilde{\eta}_{1i}$ is $1 - \hat{\sigma}^2_{\tilde{\eta}_1 i}$, which is also the estimated individual-specific reliability of the factor score. As shown in Figure 2b, in the second stage, $\tilde{\eta}_1$ is modeled as an indicator of η_1 with unit variance, with the factor loading set to be $\lambda_{1i} = 1 - \hat{\sigma}^2_{\tilde{\eta}_1 i}$ and the error variance set to $\sigma^2_{\varepsilon 1i} = \hat{\sigma}^2_{\tilde{\eta}_1 i}(1 - \hat{\sigma}^2_{\tilde{\eta}_1 i})$, so that the reliability of each observation is fixed to $1 - \hat{\sigma}^2_{\tilde{\eta}_1 i}$.

The second stage of 2S-PA can be easily performed on SEM software that supports the use of definition variables, including Mplus (L. K. Muthén & Muthén, 2017) and OpenMx (Neale et al., 2016), as demonstrated in the supplemental materials (https://osf.io/h95vx/).

[2] Although the distribution of $\tilde{\eta}$ is usually not exactly normal with categorical indicators, it quickly converges to a normal distribution as the number of items increases (Bock & Mislevy, 1982) so that equation (3) is a good approximation.

Comparing 2S-PA and Other Measurement Error Adjustment Methods

If the indicators are continuous and normally distributed, 2S-PA is similar to other approaches for adjusting for measurement error. For example, Hsiao et al. (2018) and Savalei (2019) discussed the use of composite scores in the context of interaction and path analyses by fixing the factor loading for each latent variable, λ, to be 1.0 and constraining the uniqueness (i.e., measurement error variance) to be $s_y^2(1 - \rho_{yy})$ where s_y^2 is the sample variance of the composite score and ρ_{yy} is the composite reliability (which can be an estimate or a fixed/known value).[3] It is thus obvious that path analysis with composite scores and reliability adjustment is a special case of 2S-PA with $\tilde{\eta}$ being the composite scores and $\sigma^2_{\varepsilon i}$ set to $s_y^2(1 - \rho_{yy})$, which is constant for all observations. We expect this procedure to be biased when measurement error varies across observations, such as in the case of categorical indicators.

Factor score regression and factor score path analysis (Devlieger et al., 2016; Devlieger & Rosseel, 2017; Kelcey, 2019), on the other hand, directly use factor scores as observed variables for parameter estimation in regression and path analysis, and then correct for the biases in the estimated path coefficients and standard error estimates based on the method by Croon (2002), which generalized the results on the effects of measurement error in regression (e.g., Fuller, 1987; also Hardin, 2002; Murphy & Topel, 1985) to path analysis. These methods share the same idea as in 2S-PA by treating the estimated factor scores as indicators of true latent variables with known measurement error variances. It, however, requires involved calculations of the adjustment factor, although the current version of the `lavaan` R package (Rosseel, 2012; Rosseel et al., 2020) has automated the computation. Also, unlike reliability adjustment methods, it currently does not support estimation of interaction and non-linear effects. More importantly, like the reliability adjustment approach, it assumes a constant covariance matrix for the estimated factor scores, and so may not be appropriate for heterogeneous measurement error variance, which is more the norm than the exception for psychological measurement as binary and Likert-type items are particularly common.[4] As shown in Greene (2003, chapter 11), unmodeled heterogeneous error variance may lead to inefficient estimators and inadequate standard error estimates when the nonconstant variance is correlated with the predictor, but it is not clear how unmodeled heterogeneity in measurement error variance affects estimation in a path model.

Another estimation approach is the model-implied instrumental variable estimator (Bollen, 1996,

[3] An alternative way to identify the same model is to fix the latent factor variance to 1.0, and impose the constraint $\lambda^2/(\lambda^2 + \sigma^2_\varepsilon) = \rho_{yy}$.

[4] Croon and van Veldhoven (2007) discussed how to incorporate heterogeneous error variance for two-stage estimation in the context of multilevel modeling; Hardin (2002) discussed a sandwich estimator for two-stage models for heterogeneous disturbances. These are limited information maximum likelihood approaches with corrections on parameter and covariance estimates, while 2S-PA uses joint modeling that incorporates the heterogeneous measurement error in the likelihood function.

2019), with the extension of the polychoric instrumental variable (PIV) estimator (Bollen & Maydeu-Olivares, 2007) for binary and ordered categorical data. PIV is a two-stage equation-by-equation estimation method using instrumental variables that are implied from the model structure, which is less susceptible to convergence issues. It has also been shown to be more robust to model misspecification (e.g., Jin et al., 2016; Nestler, 2013). We include PIV in our simulation Study 2, which evaluates the performance of various methods under model misspecifications.

Comparing 2S-PA and Full SEM

Although full SEM is commonly regarded as the gold standard to account for measurement error in estimating structural relations, previous studies have suggested that single indicator methods with adjustment have several advantages over full SEM, including more precise estimates of the path coefficients as measured by the root mean squared error (RMSE) in small samples (Kelcey, 2019; Savalei, 2019) and robustness to misspecification in the measurement model (Devlieger & Rosseel, 2017) when factor scores were estimated in separate models. As will be demonstrated and discussed in a series of simulation studies in this paper, by reducing model complexity, the proposed 2S-PA approach also provides better control of Type I error rates and smaller RMSEs for the structural coefficients, as well as drastically improved convergence rates. Besides, on a more conceptual level, we argue that the 2S-PA approach has the following two advantages over full SEM.

Separate Estimation of Measurement and Structural Models

The first advantage of 2S-PA is that it allows for separate estimation processes for the measurement and the structural models. In a full SEM model, usually there are many more variables involved in the measurement model than in the structural model. In the presence of ordered categorical data, estimation methods under full SEM generally fall into two categories: weighted least squares (WLS) and maximum likelihood (ML). Whereas WLS estimators were shown to have reasonable performance with sufficient sample size (Asparouhov & Muthén, 2012), some research found they produced biased structural coefficients (e.g., Li, 2016) and, contrary to ML estimators, WLS estimators do not automatically handle missing data under the missing at random mechanism (as illustrated in Pritikin et al., 2018). On the other hand, ML estimators for categorical data generally require the use of numerical integration by conditioning on the latent variables (Embretson & Reise, 2000), and estimating models with more than a few latent variables is computationally challenging.[5]

Instead, with 2S-PA, researchers can fit a separate measurement model for each latent variable in

[5] See Wirth and Edwards (2007) for a more comprehensive comparison of different estimation choices.

the overall model, which solves the dimensionality problem. By doing so, it allows the use of the most appropriate estimation method for each measurement model. Researchers are also free to choose state-of-the-art psychometric models that are available only in specialized software, and estimate the structural model on SEM software that supports definition variables. For example, one can fit the monotonic polynomial generalized partial credit model (Falk & Cai, 2016) with the Metropolis-Hastings Robbins-Monro algorithm (Cai, 2010) in the *mirt* package in R (Chalmers, 2012), obtain factor scores via the EAP method, and use Mplus or OpenMx to estimate structural relations together with other observed variables. Such an option, however, is currently limited with full SEM as it requires that the SEM software directly supports the advanced psychometric models. Indeed, many of the recent development in psychometrics, such as IRT tree models (De Boeck & Partchev, 2012), network psychometrics (Epskamp et al., 2017), and so forth, are not based on the conventional SEM framework and thus may not be available in some current SEM software. Similarly, the structural model may contain nonnormal or discrete observed outcome variables that require different intensive estimation methods, and putting the measurement model and the structural model with all variables together may not be feasible. By separately estimating the measurement and the structural models, 2S-PA allows researchers to combine the best from both worlds.

Apply Diagnostic Tools Commonly Used in Regressions

Another advantage of 2S-PA is that, by explicitly obtaining the factor scores, it allows researchers to use diagnostic tools that are commonly deployed for regression models to assess problems such as nonlinearity and outliers. As Hallgren et al. (2019) pointed out, none of the 37 articles they reviewed in addiction research journals that used SEM provided scatterplots or other diagnostic plots commonly used in regression analyses, and a main reason was that the latent variables were not realized values. Therefore, Hallgren et al. (2019) recommended obtaining factor scores and used them to provide diagnostic plots for structural relations in SEM. Although factor scores are not the same as error-free latent variables and different options for computing factor scores can sometimes produce substantially different scores (Skrondal & Laake, 2001), by estimating and saving them in the first stage, researchers are more equipped to evaluate the validity of the specified functional form and the distributional assumption for each path in the structural model, which are often masked when using full SEM and cannot be detected with significance tests of path coefficients and goodness-of-fit indices. Figure 3, which is based on the empirical example presented later in this paper, shows that the normality assumption is violated at the factor score level.

In the following sections, we report the results of a series of Monte Carlo studies comparing the performance of 2S-PA with full SEM and several alternative methods. In Study 1, we use a latent regression model with measurement error in the predictor. In Study 2, both the predictor and the outcome in the model have measurement error, and we examine the robustness of 2S-PA and other approaches to misspecification in the measurement model. In Study 3, we examine a path model with three latent variables, with a focus on estimating an indirect effect.

Study 1: Measurement Error in a Single Predictor

In Study 1, we examine the performance of 2S-PA as compared to full SEM and alternative measurement error adjustment methods when there is measurement error on the predictor.

Data Generating Model

The data generating model was similar to the one shown in Figure 1, where each indicator for η_1, the latent predictor, has K categories. The indicators were generated from a graded response model (Samejima, 1969) with different loadings, and parameterized as an item factor analysis model (Wirth & Edwards, 2007) with a cumulative logit link:

$$y^*_{ij} = \lambda_j \eta_i + \epsilon_{ij}, \tag{4}$$

$$y_{ij} = \begin{cases} 0 & \text{if } y^*_{ij} < \tau_{j1} \\ k & \text{if } \tau_{jk} \leq y^*_{ij} < \tau_{j(k+1)} \\ K-1 & \text{if } y^*_{ij} \geq \tau_{j(K-1)}, \end{cases} \tag{5}$$

where y^*_{ij} is the score of the ith person on the latent continuous response variate for indicator j, ϵ_{ij} is the realized value of the unique factor following a standard logistic distribution, and τ_{j1}, ..., $\tau_{j(K-1)}$ are the threshold parameters for the jth indicator.

We used R 3.6.1 (R Core Team, 2019) to first generate η_1 from a standard normal distribution, and then computed η_2, the observed outcome variable, as $\eta_{2i} = \beta_0 + \beta_1 \eta_{1i} + \zeta_i$, where ζ_i was also normally distributed with mean 0 and variance 1 - β_1^2 so that the total variance of η_2 was also 1. The indicators were then generated according to the graded response model as previously discussed.

We simulated the threshold levels so that the observed indicators had skewed distributions. Specifically, when $K = 2$, the thresholds were generated as $\boldsymbol{\tau}^* = \{-2.20, -1.39, -0.95, -0.41, 0\}$ on the logit scale so that the indicators had success probabilities of 0.9, 0.8, 0.7, 0.6, 0.5. When $K = 4$, the first thresholds $\boldsymbol{\tau}_1$ corresponded to -1 + $\boldsymbol{\tau}^*$, the second thresholds $\boldsymbol{\tau}_2$ corresponded to $\boldsymbol{\tau}^*$, and the third thresholds $\boldsymbol{\tau}_3$ corresponded to $1 + \boldsymbol{\tau}^*$, respectively.

Design Factors

Number of Categories (K)

The number of categories were chosen to be 2 or 4 for each indicator. This covers a range of commonly used response formats in the behavioral and social sciences. More categories were not studied as we expected the results to be at least as good as when $K = 4$, as discussed in Rhemtulla et al. (2012).

Sample Size per Indicator (N/p)

In full SEM a general recommendation is to have a sample size of 100 or more for a simple model like this one (e.g., Kline, 2016), so we would like to examine whether 2S-PA performs better than SEM in small samples, as Savalei (2019) found some evidence that reliability adjustment methods with fixed reliability outperformed SEM. As sample size recommendations in SEM were usually based on the relative N per indicator (e.g., MacCallum et al., 1999), in Study 1 we chose N/p = 6, 25, 100, which covered common situations with small to large sample sizes. As a result, the maximum sample size was 2,000 and the smallest was 30.

Average Factor Loading ($\bar{\lambda}$)

We simulated data with varying loadings with either $\bar{\lambda} = 1$ or 2.5. With unit variance for the latent predictor, the average standardized loadings for the latent response variates were approximately 0.48 and 0.81. The loadings sequentially decreased in equally-spaced intervals across indicators, with the maximum being $1.5 \times \bar{\lambda}$ and the minimum being $0.5 \times \bar{\lambda}$. For example, in conditions with $\bar{\lambda} = 2.5$ and with 10 indicators, the maximum loading was 3.75 and the minimum was 1.25. The combination of $\bar{\lambda} = 1$ and small p resulted in low composite reliability (e.g., $\omega^{\text{NL}} \approx 0.47$ when $p = 5$ and $K = 2$), whereas $\bar{\lambda} = 2.5$ coupled with large p resulted in high composite reliability (e.g., $\omega^{\text{NL}} \approx 0.93$ when $p = 10$ and $K = 4$).

In addition, we manipulated the number of indicators for the latent predictor to be p = 5, 10, 20, and the regression (structural) coefficient of η_1 predicting η_2 to be either $\beta_1 = 0$ (null effect) or $\beta_1 = 0.5$ (medium effect).

Analytic Approaches

We compared six analytic approaches in Study 1, which includes (a) linear regression/path analysis (PA), (b) full SEM (SEM), (c) 2S-PA, and reliability adjustment with (d) coefficient alpha (RA-α), (e) coefficient omega (RA-ω), and (f) coefficient omega for categorical indicators (RA-ω^{NL}). For PA, the predictor is a composite score of the five indicators of η_1. Mplus 8.3 (L. K. Muthén & Muthén, 2017) was used for all approaches. For SEM, the diagonally weighted least squares (DWLS) estimator with robust standard errors (`ESTIMTOR=WLSMV` in Mplus) was used.[6] [7] For 2S-PA, we first fit a one-factor model to the five categorical indicators using maximum likelihood estimation with numerical integration with adaptive quadrature and 15 integration points.[8] [9] [10], and then obtained the factor scores and the corresponding standard errors with the EAP method. For the three RA methods, we obtained the composite reliability estimates using R (with the *psych* package, Revelle, 2019, for α; and the *MBESS* package, Kelley, 2020, for ω and ω^{NL}).

For all models, we obtained the sample point and standard error estimates of β_1, denoted as $\hat{\beta}_1$ and $\hat{\text{SE}}(\hat{\beta}_1)$. For all structural models, the measurement part of η_1 was identified by constraining the latent factor variance to be 1 and the uniqueness of X to be 0, so that the latent predictor was standardized to ensure fair comparison to the population β_1 parameter. In other words, the analytic approaches were compared on the standardized $\hat{\beta}_1$ coefficient, consistent with previous simulation studies (e.g., Cole & Preacher, 2014; Savalei, 2019).

The Monte Carlo simulation was structured using the R package `SimDesign` (Chalmers, 2020), which automatically collected warning and error messages during the simulation. For replications where one or more analyses returned an error, the package automatically resimulated a new data set until convergence was obtained for all analyses, but for each attempt we also saved information on which analyses encountered convergence issues so that we could properly compute convergence rates. For each condition, we obtained 5,000 complete replications. The R code for all simulation studies can be found in the supplemental materials.

[6] The DWLS estimator first estimates the polychoric correlation matrix by assuming an underlying standard normal latent response variate for each indicator as well as the asymptotic covariance matrix of the polychoric correlations. The diagonal elements of the asymptotic covariance matrix is then used as the weight matrix in weighted least square estimation of model parameters.

[7] Assuming an underlying normal distribution for an observed categorical indicator corresponds to the probit link, which is different from the logit link used to generate the data. In practice, probit and logit usually give very similar results other than a scaling difference on the measurement parameters (Paek et al., 2018), as the standard normal distribution has a variance of 1 and the standard logistic distribution has a variance of $\pi^2/3$. To examine the sensitivity to this choice, in Study 2 we generated data using a probit link.

[8] With ML, the logit link is used as the default in Mplus in the first stage of 2S-PA.

[9] We did not include a version of 2S-PA that used DWLS for factor score estimation in the first stage, as it did not perform well based on our preliminary simulation results. The poor performance is likely due to the computation of the factor scores and the associated standard errors based on the maximum a posteriori (MAP) method.

[10] We also included a variant of 2S-PA that used the R package mirt for factor score estimation in the first stage, but because the results were very similar to using Mplus, we only presented results of 2S-PA using Mplus. The full results can be found in the supplemental materials (https://osf.io/h95vx/).

Evaluation Criteria

For each method in each replication, we computed the convergence rate, bias, the root mean squared error (RMSE), the relative standard error (*SE*) bias, the empirical Type I error rate (for $\beta_1 = 0$ conditions), and the empirical power (for $\beta_1 > 0$ conditions).

Convergence Rate

The convergence rate was computed as the proportion of replications without an error, including replications where the program gave a warning (e.g., variance estimates < 0), out of all replication attempts (including the failed ones that did not go into the complete replications). Major reasons for nonconvergence included empirical underidentification due to simulated indicators having close to zero correlations (mostly for full SEM) and negative sample estimates of overall reliability (for RA methods) or individual-specific reliability (for 2S-PA).

For some converged conditions, Mplus still gave extreme parameter and standard error estimates (e.g., $SE > 500$ in some small samples). To avoid the influence of extreme outliers, we computed robust versions of bias, RMSE, and *SE* bias, as explained below, while the raw bias, RMSE, and *SE* bias can be found in the supplemental materials.[11]

Bias

The bias was computed as $\bar{\hat{\beta}}_1 - \beta_1$, where $\bar{\hat{\beta}}_1 = \sum_{i=1}^{R} \hat{\beta}_{1i}/R$ with R = 5,000 replications is the 20% trimmed mean (Wilcox, 2016) of the $\hat{\beta}_{1i}$ estimates across replications. The 20% trimmed mean was suggested to be a good compromise between the arithmetic mean (or 0% trimmed mean), which is highly sensitive to outliers, and the median (or 100% trimmed mean), which is robust but inefficient for normally distributed data. For conditions with $\beta_1 \neq 0$, we also computed the relative bias = bias / β_1.

RMSE (Ratio)

The robust RMSE was computed as $\sqrt{\text{Bias}^2 + [MAD(\hat{\beta}_1)]^2}$, where $MAD(\hat{\beta}_1)$ was the sample median absolute deviation (from the median with a scale factor of 1.4826) of the 5,000 $\hat{\beta}_1$ estimates. The RMSE indicated the typical distance of the sample estimated value from the true value of β_1, the standardized regression coefficient. As RMSE was heavily dependent on sample size and the magnitude of

[11] The full SEM method generally suffered more from extreme parameter estimates, especially in small samples. For example, in one small sample condition, the usual RMSE for SEM was 0.42, versus 0.25 for the robust RMSE. In larger samples, the robust and non-robust versions of the evaluation criteria were almost identical. We also reported the proportion of outliers for each method in the supplemental materials.

β_1, we computed the RMSE ratio relative to PA (denoted as RR) as RR = $\text{RMSE}_{\text{PA}}(\hat{\beta}_1)/\text{RMSE}_M(\hat{\beta}_1)$ for method M, with RR > 1 indicating the method M is more efficient than PA.

Relative SE *Bias*

The robust relative standard error bias (RSB) was computed as $\overline{\hat{SE}}(\hat{\beta}_1)/MAD(\hat{\beta}_1) - 1$, where $\overline{\hat{SE}}(\hat{\beta}_1)$ was the 20% trimmed mean of the estimated standard error of $\hat{\beta}_1$, and $MAD(\hat{\beta}_1)$ was used as an estimate of the empirical *SE*. We considered the bias acceptable if its absolute value is within 10% (Hoogland & Boomsma, 1998).

Empirical Type I Error Rate/Power

The empirical Type I error rate (α^*) was defined as the proportion of replications where the Wald test statistic exceeded the critical value at .05 significance level for conditions with $\beta_1 = 0$; empirical power was similar defined but for conditions with $\beta_1 \neq 0$.

Results

Convergence Rate

For all methods, when either $N/p = 100$ or $p \geq 10$, the convergence rate was ≥ 99.41%. For almost all conditions, RA-α and 2S-PA showed the highest convergence rates, especially for low reliability conditions ($p = 5$, $N/p = 6$, $\bar{\lambda} = 1$), where the mean convergence rate was 98.60% for RA-α, 98.31% for 2S-PA, 94.11% for SEM, 76.40% for RA-ω, and 91.53% for RA-ω^{NL}.

Bias

When $\beta_1 = 0$, the estimates were essentially unbiased for all methods (with absolute values < 0.004). Table 1 shows the relative bias when $\beta_1 = 0.5$. Across conditions, full SEM provided the best estimates in terms of bias as the relative bias was less than 7.94% in absolute value. The three reliability adjustment methods also performed reasonably with no more than 10% of bias in all but one condition; however, the biases were higher for conditions with larger $\bar{\lambda}$, and did not decrease with a larger sample size. The 2S-PA method demonstrated substantial biases when $\bar{\lambda} = 1$, $p = 5$, where the relative bias was -25.16% when $K = 2$ and -19.60% when $K = 4$. The bias was within 10% when there were at least 10 indicators, $N/p \geq 25$, or $\bar{\lambda} = 2.5$.

RMSE Ratio

In general, the RMSE ratio (RR) relative to PA was smaller than 1 for all methods when $\beta_1 = 0$ (RRs between 0.66 and 0.98) or when $N/p = 6$ (RRs between 0.66 and 1.13), so PA was generally more efficient in small samples and when estimating a zero coefficient. When $\beta_1 = 0.5$ and $N/p \geq 25$, adjusting for measurement error generally produced better estimates than PA, with larger RR when $\bar{\lambda} = 1$ and $\beta_1 = 0.5$ (RRs between 1.33 and 3.02). There was little variation in RR across the different analytic approaches.

Relative SE *Bias*

Table 2 shows the RSB values of the different methods across conditions of N/p, K, and p. All methods showed acceptable RSB except for SEM with downward bias of around 15% when the sample size was small and $p = 5$.

Empirical Type I Error Rate/Power

For conditions with $\beta_1 = 0$, SEM showed the largest inflation in α^*, especially when $N/p = 6$ and $\bar{\lambda} = 1$ (α^* up to 0.14). PA and the RA methods generally performed best (with α^* up to 0.07); 2S-PA had α^* slightly worse than PA (with α^* up to 0.08), but improved with larger N/p, $\bar{\lambda}$, and p. As for power, there was little difference across methods, except that SEM had larger power in low reliability and small sample conditions; however, the increased power in those conditions was largely driven by the inflated α^* of SEM.

Discussion

In Study 1, we compared the performance of 2S-PA with full SEM and other reliability adjustment methods when there was measurement error on the latent predictor measured by categorical indicators. Overall, it was found that 2S-PA gave slightly smaller path coefficient estimates with small sample sizes, and otherwise performed similarly to SEM and had better convergence rates and control of *SE* bias and Type I error rates. We also examined the effect of number of indicators, which increased the reliability of the composite scores and the estimated factor scores. When $p = 20$, generally all methods that accounted for measurement error performed similarly.

Given the downward bias of 2S-PA, some small sample adjustment might be beneficial. Incorporating Bayesian priors in the first stage of 2S-PA largely reduced the bias, as further shown and discussed in Study 2.

Study 2: Robustness Against Misspecifications in the Measurement Model

So far, we have shown that 2S-PA performed favorably as compared with SEM and other reliability adjustment methods (other than RA-α), especially in small samples, in terms of convergence rates and bias of standard error estimates. However, one potential benefit of SEM is that it allows indicators to load on more than one latent construct. Although with 2S-PA, one can still obtain factor scores from a q-dimensional measurement model, the errors in the obtained factor scores are usually correlated, and theoretically such covariances need to be incorporated into the definition variable step to obtain unbiased path coefficients. In other words, one would need to obtain a $q \times q$ covariance matrix for the factor score estimates for each individual, which is not always available in standard software.[12]

Instead, in Study 2, we evaluate an approach that fits a separate unidimensional measurement model to each latent factor to obtain factor score estimates, similar to what Devlieger and Rosseel (2017) studied in the context of factor score path analysis with continuous indicators. While this approach can lead to bias due to omitted cross-loadings or unique factor covariances across latent factors, it also reduces the model size in the measurement model and the structural model, and Devlieger and Rosseel (2017) found that this approach was more robust to misspecification in the measurement model part compared to full SEM. We also include the polychoric instrumental variable (PIV) estimator, which was found robust to misspecification in previous research (Jin et al., 2016; Nestler, 2013). Like in Study 1, we compare the methods on the standardized β_1 coefficient.

Data Generating Model

The data generating model was similar to the one in Study 1, except that the latent outcome, η_2, was measured by five binary indicators (i.e., $K = 2$), as Study 1 found relatively small impact of K. Also, the probit link was used such that the unique factors, ϵ_{ij} in equation (4), followed a standard normal distribution. In addition, in some conditions, the third indicators for η_1 and for η_2 were predicted by an unobserved confounding variable, so that they had a residual unique factor covariance of δ.

Design Factors

We manipulated sample size (N), population regression coefficient (β_1), average standardized factor loading ($\bar{\lambda}^s$), and the residual unique factor correlation (δ). Similar to Study 1 we chose $N/p = 6$, 25, 100, and with five indicators, $N = 30$, 125, and 500. β_1 was set to 0 (null effect) or 0.5 (medium effect). Under the probit link, we set the average loading to 0.707 and 1.789, which corresponded to standardized factor loadings ($\bar{\lambda}^s = \sqrt{\bar{\lambda}^2/[\bar{\lambda}^2 + 1]}$) of .5 and .8 for the latent responses and were similar to those of Study 1 after the scale adjustment of probit/logit link. The first indicator had a loading of $1.2 \times \bar{\lambda}$, and the loading sequentially decreased to $0.8 \times \bar{\lambda}$ for the fifth indicator, for both the latent predictor and the latent outcome. For δ, the correlation between the latent continuous response variates of the third indicators of

[12] To our knowledge Mplus does not output individual covariance matrices for factor score estimates, but they can be obtained in R packages such as OpenMx and mirt.

η_1 and of η_2, the manipulated levels were -0.16, 0, 0.16, and 0.64.

Analytic Approaches

We compared SEM (omitting the unique factor covariances), SEM-cov (which correctly modelled the unique factor covariances), RA-α, 2S-PA, 2S-PA with Bayes (see Appendix B for details of our implementation), and PIV (see Appendix C for more details). Results for PA were not reported as it substantially underestimated the population coefficient, as demonstrated in Study 1, although it was still used as a baseline to compute the RMSE ratios.

Results

Convergence Rate

For conditions with $\beta_1 = 0$, $N = 30$, and $\bar{\lambda}^s = .5$, the convergence rates for SEM and SEM-cov (medians = 90.44% and 90.52%) were substantially lower than those for RA-α, the 2S-PA methods, and PIV (all of which had median convergence rates > 98.16%).[13]

Bias

When $\beta_1 = 0$, the 2S-PA methods, RA-α, and PIV showed only small biases (between -0.02 and 0.09), despite the model misspecification. On the other hand, SEM gave biased estimates of β_1 (bias = 0.07 to 0.19) when $\bar{\lambda}^s = .5$ and $N = 30$. Surprisingly, even the correctly specified model, SEM-cov, also demonstrated similar upward bias (0.07 to 0.11) when $\bar{\lambda}^s = .5$ and $N = 30$.

Figure 4 shows the *relative bias* on the estimates of β_1 across different methods when $\beta_1 = 0.5$. Generally, all methods except SEM-cov and PIV were affected by model misspecification. When $N = 30$, SEM and SEM-cov showed the largest upward biases (up to 51.97%), whereas PIV showed the largest downward biases (up to -41.26%). Similar to the results in Study 1, 2S-PA showed smaller but still substantial downward biases when reliability was low (i.e., $\bar{\lambda}^s = .5$), but 2S-PA with Bayes removed that bias and performed the best in terms of bias in small samples. For larger samples, SEM-cov yielded estimates with negligible bias only when $N = 500$ or when $N = 125$ and $\bar{\lambda}^s = .8$, whereas the bias for PIV did not go away until $N = 500$ and $\bar{\lambda}^s = .8$. RA-α performed reasonably well in low reliability conditions (except when $\delta = .64$) but consistently yielded coefficients that were too small in high reliability conditions. On the other hand, 2S-PA methods generally gave estimates with relative bias < 5%, except for conditions with strong misspecification ($\delta = .64$) and $\bar{\lambda}^s = .5$.

[13] In 4.12% to 13.76% of the replications for conditions with $N = 30$, standardized coefficients were not obtainable for PIV due to negative variance estimates of the latent predictor.

RMSE Ratio

When $\beta_1 = 0$, 2S-PA, 2S-PA with Bayes, and RA-α were relatively more efficient than SEM and SEM-cov in small samples. PIV was generally the least efficient with RRs = 0.30 to 0.67. When $\beta_1 = 0.5$, the 2S-PA methods had better RMSE for conditions with $\delta = .64$ and $\bar{\lambda}^s = .5$ (RRs = 0.77 to 1.80, compared to 0.75 to 1.32 for SEM and SEM-cov). In other conditions, the differences among the 2S-PA methods, SEM, and SEM-cov were negligible. Again, PIV generally had the worst RR ratio.

Relative SE *Bias*

Consistent with Study 1, 2S-PA and RA methods outperformed SEM in terms of the accuracy of *SE* estimates, especially in small samples. When $N = 30$, RA-α and 2S-PA performed the best (RSB = -15.81% to -5.50%), followed by 2S-PA with Bayes (-21.38% to -6.81%); SEM and SEM-cov showed substantial biases (-62.43% to -18.98%). The *SE* bias improved for all methods when $N \geq 125$ and were generally within the 10% benchmark, except for SEM and SEM-cov (e.g., -35.51% when $N = 125$ and -19.05% when $N = 500$) and PIV (which had extremely large relative SE bias of up to 1,074.77% when $N = 30$ and 172.03% when $N = 125$).

Empirical Type I Error Rate

The empirical power was very similar across analytic approaches except for conditions where SEM and SEM-cov showed inflated α^* levels. As shown in Figure 5, SEM and SEM-cov showed the largest α^* when $N \leq 125$, especially when $\bar{\lambda}^s = 0.5$ (α^* between 0.15 and 0.48 for SEM and 0.14 and 0.43 for SEM-cov). Although still inflated, RA-α and the two 2S-PA methods generally had α^* closer to the nominal level even under model misspecification (except with small N and small $\bar{\lambda}^s$). Consistent with previous studies, PIV was conservative and had α^* below nominal level except when $\bar{\lambda}^s = .8$ and $N = 500$.

Discussion

In Study 2, we found that when both the latent predictor and the latent outcome were measured with error, 2S-PA—even when omitting some misspecification in the measurement model—outperformed full SEM that omits or correctly models the unique factor covariance in terms of convergence rates, bias, efficiency, and control of Type I error rates. This holds not just with both low reliability and small sample size, but also with medium or even large sample size and with high reliability conditions. In addition, although RA-α performed better than SEM, it was generally inferior to 2S-PA methods in terms of convergence and robustness to misspecification, but provided better control of Type I error rates. When the sample size is small and bias is a concern, we recommend the use of 2S-PA with Bayes to obtain factor scores in the first stage, whereas 2S-PA with maximum likelihood estimation is suitable for situations with high reliability or large sample size.

Study 3: Mediation Model

In the previous two studies we have shown that 2S-PA is mostly a good alternative to SEM when there is measurement error in the predictor and/or the outcome in a regression model. Given that 2S-PA can also handle multivariate analyses as in SEM, following Savalei (2019), in Study 3 we compare the performance of 2S-PA with SEM using a mediation model with three variables, a model commonly used in psychological research (see e.g., MacKinnon et al., 2007).

Data Generating Model

The data generating model is shown in Figure 6, where each of the latent variables, η_1 (the predictor), η_2 (the mediator), and η_3 (the outcome), was measured by 5 binary indicators. There were no unique factor covariances among any pairs of indicators. The structural model was:

$$\begin{aligned}\eta_2 &= a\eta_1 + \zeta_2 \\ \eta_3 &= b\eta_2 + c\eta_1 + \zeta_3.\end{aligned}$$

Different from Studies 1 to 3, here there were three path coefficients instead of one. In addition, the indirect effect of the latent constructs, defined as the product of the two coefficients ab, was also of interest, but none of the previous simulation studies on measurement error adjustment specifically studied the estimation of the indirect effect. Therefore, in Study 3 we evaluated the estimation of the individual a, b, and c coefficients, as well as the ab indirect effect. All coefficients were obtained with the latent variables standardized.

Design Factors

Following previous simulation studies (e.g., Fairchild et al., 2009), we manipulated each of a and b to be either 0 (null effect) or 0.39 (medium effect). The population coefficient of c was fixed to be .15 (small effect). Therefore, there were in total four configurations of the coefficients $\{a, b, c, ab\}$: {0, 0, .15, 0}, {0, .39, .15, 0}, {.39, 0, .15, 0}, {.39, .39, .15, .1521}.

The other design factors were similar to Studies 1 and 2: N = 30, 125, 500, and $\bar{\lambda}$ = 1 or 2.5 (under a logit link as in Study 1). The analytic approaches included 2S-PA, 2S-PA with Bayes, full SEM, RA-α, and path analysis (PA; using sum scores without accounting for measurement error). For 2S-PA and 2S-PA with Bayes, we obtained factor scores separately for η_1, η_2, and η_3, in three separate measurement models. For each approach, the estimate of the indirect effect ab was computed as the product of the estimated a and b coefficients, and we evaluated the convergence rate and the bias of each coefficient. In addition, because it is common in practice to use a 95% confidence interval (CI) for statistical inference of the indirect effect ab (MacKinnon et al., 2002), for each method we also computed the 95% CI using the Monte Carlo method (MacKinnon et al., 2004; Preacher & Selig, 2012), and obtained the empirical CI coverage for ab, defined as the proportion of replications in which the 95% CI contained the population value of ab. Note that for conditions where $ab = 0$, the empirical coverage was the same as $1 - \alpha^*$.

Results

Convergence Rate

Similar to Studies 1 and 2, SEM had poor convergence rate for conditions with $N = 30$ and $\bar{\lambda} = 1$ (min = 72.58%) as compared to RA-α (min = 85.91%), 2S-PA with Bayes (min = 92.22%), and 2S-PA (min = 95.52%). When $N \geq 125$, all methods had convergence rates above 95%, although 2S-PA still yielded better convergence when $\bar{\lambda} = 1$.

Bias

When the population values of coefficients a and b were zero, only SEM tended to overestimate the zero coefficients (bias between 0.09 and 0.15 when $N/p = 6$ and when $\bar{\lambda} = 1$), while all other methods gave close to unbiased estimates in all conditions (bias between 0.00 and 0.04). Figure 7 showed the relative bias for estimating non-zero coefficients a, b, and c. Consistent with Study 2, 2S-PA underestimated the non-zero coefficients when $N = 30$ and when $\bar{\lambda} = 1$, but the bias was mostly corrected in 2S-PA with Bayes. On the other hand, SEM overestimated the true coefficients not only when $N = 30$ and $\bar{\lambda} = 1$ (up to 121.69%), but also when $N = 125$ and $\bar{\lambda} = 1$ (up to 20.12%) as well as when $N = 30$ and $\bar{\lambda} = 2.5$ (up to 35.70%). RA-α also showed upward bias when $N = 30$ and $\bar{\lambda} = 1$ (up to 41.27%). The biases were negligible with $N = 500$.

For the estimates of the indirect effect (ab), when $a = b = 0$, all methods had bias with absolute value less than 0.02. When either $a = 0.39$ or $b = 0.39$ but the true $ab = 0$, only SEM had some upward bias when $\bar{\lambda} = 1$ (with bias up to 0.04), while all other methods were unbiased. When $a = b = 0.39$, as shown in Figure 7, 2S-PA showed downward bias when $\bar{\lambda} = 1$ (-73.18% when $N = 30$; -34.38% when $N = 125$), and 2S-PA with Bayes could not fully correct the small sample bias (-39.81% when $N = 30$; -29.04% when $N = 125$). With larger $\bar{\lambda}$ or N, the estimates of ab under the 2S-PA method were close to the population values. RA-α showed smaller small sample bias (up to -18.98%), but did not provide consistent estimates as the bias was still large in high reliability and large sample size conditions (-13.76%). SEM showed upward bias when $N = 30$ (up to 43.37%). Therefore, whereas 2S-PA showed less bias on the individual coefficients, it seemed to yield more biased indirect effect estimates in small samples. When either $\bar{\lambda} = 2.5$ or $N = 500$, both 2S-PA methods and SEM yielded virtually unbiased estimates of non-zero indirect effects.

Empirical Coverage for the Indirect Effect

As shown in Table 3, the coverage for *ab* for 2S-PA was generally 92% or above except for two conditions for 2S-PA and one condition for 2S-PA with Bayes (with non-zero *ab*, $\bar{\lambda} = 1$, and $N \leq 125$). For SEM, coverage < 92% for five conditions with $N \leq 125$, and overall had inflated Type I error rates when either *a* or *b* was zero (up to 10.6%), as compared to other methods. RA-α had coverage above 92% except for conditions with non-zero *ab* and low measurement error.

Discussion

In Study 3, we found that the 2S-PA methods generally yielded consistent estimates and inferences for indirect effects, but might produce negatively biased estimates of path coefficients in small samples, compared to overestimates in SEM. Overall, 2S-PA methods provided better control on Type I error and coverage rates, and had convergence rates superior to those of SEM.

Empirical Demonstration

Here we demonstrate 2S-PA methods as well as path analysis with composite scores, full SEM (with DWLS), and reliability adjustment methods with alpha (RA-α) using an empirical path model comparable to the model studied by Jang et al. (2008). Data were collected from the Midlife Development in the United States project from 1995 to 1996 (MIDUS I). The total number of participants recruited in MIDUS I was 7,108. We selected participants aged 45 to 74 based on the criterion in Jang et al. (2008) and excluded those missing in all the variables in the model for the following analyses. The final sample size for analyses ranged from 3,440 to 3,574.[14]

The latent predictor, Perceived Discrimination (PD), was tapped by nine Likert-type items (1 = *Often* to 4 = *Never*) assessing the frequency of maltreatment or disrespects by others in daily life. The latent mediator, Sense of Control (SC), was measured by twelve items (1 = *agree strongly* to 7 = *disagree*

[14] The sample sizes were smaller for the 2S-PA methods and path analysis, as they removed cases that had missing responses on all items for one or more of the three constructs.

strongly) capturing one's sense of mastery and perceived constraints within 30 days. The latent outcome, Positive Affect (PA), was assessed by six items on 5-point scales measuring the frequency of feeling cheerful, good spirits, extremely happy, calm and peaceful, satisfied, and full of life within 30 days. See the supplemental materials for the full set of items. For all constructs, we reverse-coded some items in the analyses so that higher item scores indicated higher levels of PD/SC/PA, and the score reliability was high (α = .926 and ω = .932 for PD; α = .850, ω = .858 for SC; α = .910, ω = .912 for PA).

We hypothesized that PD would be negatively related to SC and that SC would be positively related to PA (Jang et al., 2008), and tested a path model similar to the one used in Study 3. R and Mplus were used to perform reliability estimations and parameter estimations of four analytic approaches in the same way as in Study 3. These approaches were compared in terms of point and CI estimates of the indirect effect.

Table 4 listed the path coefficients and the product of coefficients for the path model across the four approaches, and significant indirect effects were observed for all approaches. As hypothesized, we found that higher PD was associated with lower SC (all ps $< .001$), and individuals with lower SC had lower PA (all ps $< .001$). Using the product of coefficient method to calculate the indirect effect (MacKinnon et al., 2002), we found evidence for the indirect effect of higher PD on lower PA with all four approaches, based on the 95% Monte Carlo CIs. In terms of the magnitude of the indirect effect, the two 2S-PA methods, full SEM, and RA-α yielded comparable estimates, ranging from -0.089 to -0.087. On the other hand, the indirect effect yielded from the conventional path model was the smallest in magnitude among the four approaches (-0.069). The *SE* estimates were also similar across the four approaches.

In addition, as shown in Figure 3, the estimated factor scores of PD had a strong floor effect as a majority of the participants responded with a "1" for all items of Perceived Discrimination. Such assessment of distributional assumptions was rarely reported when using SEM,[15] but can be easily obtained using 2S-PA and RA methods. Looking at the distribution of PD, it might be sensible for researchers to estimate separate models for participants with all "1"s on Perceived Discrimination items and the remaining ones, or consider alternative analytic approaches that take into account the nonnormality of the latent predictor, a step we would argue is usually ignored when using SEM, based on our experiences. Moreover, with 2S-PA and RA methods one can easily obtain robust *SE*s (e.g., with the `ESTIMATOR=MLR` option in Mplus and the `imxRobustSE()` function in OpenMx) in the second stage, which

[15] Strictly speaking, given that PD was an exogenous variable, the normality assumption was only made when PD was modelled as a latent variable but not when it was treated as observed as in path analysis.

should give inference that is more robust to nonnormality of the latent predictor and disturbances.[16]

To compare the small sample performance of the four analytic approaches, we randomly sampled 100 participants from the whole sample and reran the analyses on the subset. The detail can be found in the supplemental materials, together with the Mplus and R codes for running the analyses. It was found that, whereas the indirect effect was not significant for all four approaches due to the small sample size, the estimate was largest with full SEM (-0.106) compared to the other approaches (-.086 for RA-α and -.092 for 2S-PA methods), and the *SE* estimates were smallest with full SEM. As a result, SEM yielded a narrower 95% CI for the indirect effect, [-0.232, 0.013], as compared to that with 2S-PA, [-0.239, 0.050]. These were consistent with the results of Study 3 that CIs under full SEM had undercoverage in small samples.

General Discussion

In this paper, we propose a two-stage path analysis with definition variables framework and report findings from three simulation studies comparing it with conventional SEM and other methods that account for measurement error, when constructs are measured by ordered categorical indicators. We also illustrate the 2S-PA method using real data from a public data set, and provide software code in both Mplus and in R (using the OpenMx and the mirt packages) for implementing 2S-PA. Here we summarize the findings from the three studies, discuss the pros and cons of 2S-PA and the implications for research, and explore future extensions of the method.

Summary of Findings

Results of Study 1 show that for data generated with equal loadings, 2S-PA with maximum likelihood estimation generally yields estimates with negligible biases for the standardized path coefficient and the corresponding standard error and acceptable control of Type I error rates. It performs similarly as SEM in large sample and high reliability conditions, but is better than SEM in small sample and low reliability conditions in terms of *SE* bias, Type I error rate, and convergence rates. 2S-PA tends to yield underestimated path coefficients in small sample (N = 30) and low reliability conditions; the bias, however, can be reduced with the use of weakly informative priors with Bayesian estimation of factor scores.

Although the reliability adjustment method RA-α is not a main focus of this research, we also find that it performs reasonably well in most simulation conditions, especially in small samples. Indeed, with small samples, it is slightly better than both 2S-PA and SEM in terms of *SE* bias, Type I error rate control, and convergence rates, despite making the assumption of homogeneous standard error of

[16] See the supplemental materials for the Mplus and OpenMx syntaxes that compute robust *SE*s in the second stage of 2S-PA.

measurement across participants. Therefore, for data similar to the small sample conditions in Study 1, we conclude that RA-α is also a good alternative to SEM for data with small to medium sample size and with moderate reliability. On the other hand, the homogeneous measurement error variance assumption leads to inconsistent estimates of the path coefficients with categorical indicators, as the estimated coefficients from RA-α did not converge to the population coefficient and had lower RMSE than those from SEM and 2S-PA when sample size is large and reliability is high, where the bias dominates the sampling variance. We also expect that the unmet assumption of homogeneous measurement error may have a bigger impact for data with more extreme values on the latent variable distributions than a normal distribution, as extreme values generally resulted in higher standard errors for the composite scores.

From Study 2, 2S-PA still performs well when both the latent predictor and the latent outcome are measured with error and with minor misspecification in the measurement model. It is more robust than full SEM, produces more accurate standard error estimates of the path coefficients in small sample sizes, and gives better control of Type I error. On the other hand, with small samples full SEM yields highly biased coefficient estimates and has highly inflated Type I error rates (as much as 50%), even with a correctly specified model. Study 3 shows that 2S-PA tends to yield negatively biased estimates of path coefficients in small samples, as opposed to overestimates by SEM, but both 2S-PA and SEM give consistent estimates and inferences for indirect effects. Overall, 2S-PA has higher convergence rate and better control of *SE* bias and Type I error rates.

Implications for Practice

With the introduction of 2S-PA and the simulation results, we now offer several recommendations for conducting path analysis using error-prone psychological measurement. First, as more journals are encouraging researchers to share their data, we suggest researchers to also compute the estimated factor scores *and* the corresponding standard errors of those scores for each latent variable when they are using 2S-PA or SEM, and append them to the data they share. We think such a practice is advantageous for two reasons. First, the estimated factor scores can be visualized to examine whether standard assumptions such as linearity and normality are appropriate, which are rarely checked in SEM analyses (Hallgren et al., 2019). Second, these scores make replications and secondary analyses easier: rather than refitting a full SEM model with many indicators from scratch, researchers can use 2S-PA with only the factor scores and the corresponding standard errors to get mostly the same (and sometimes more accurate) results. Item-level data, however, are still important as they allow examination of alternative measurement models that may fit the data better, and analyses that require cross-sample comparisons of items such as measurement invariance (e.g., Millsap, 2011).

Although the present studies examined only ordered categorical indicators, the recommendations above also applies to measurement models for continuous variables, such as confirmatory factor analysis (CFA), which is usually used for indicators with five or more categories (Rhemtulla et al., 2012). With CFA, measurement error is assumed to be constant across trait levels, so the 2S-PA model will be reduced to one where the loadings and unique factor variances of the factor scores are constrained with constants, which is equivalent to the reliability adjustment method (except that factor scores, instead of composite scores, are used). However, even with continuous indicators, the assumption of constant measurement error will not hold in the presence of missing item responses or differential item functioning (Millsap, 2011), whereas 2S-PA will have no problem handling measurement error with nonconstant variance. Therefore, in our opinion, 2S-PA represents a widely applicable approach for handling measurement error and producing reproducible results.

Although we have preliminary evidence as shown in Study 2 that 2S-PA may be more robust than regular SEM against misspecification in measurement models, consistent with the findings in Devlieger and Rosseel (2017), the path coefficient estimates still depend on whether the measurement models are specified correctly (at least approximately). Therefore, it is important that researchers assess the fit of the measurement models in the first stage, either using regular SEM fit indices for CFA for continuous indicators (cf. Kline, 2016), or fit indices based on item response theory (e.g., M_2, Maydeu-Olivares & Joe, 2006). In the supplemental materials, we also provide modified software syntax for the empirical demonstration where unique factor covariances are added based on improvement of model fit, and the fit indices of the measurement model for each construct.

Limitations

Like other statistical methods, 2S-PA has its limitations. First, because it requires different likelihood functions for each individual, to our knowledge, currently it can be implemented only in Mplus and OpenMx among the general purpose SEM software. It also requires additional specification, but future development can simplify these steps, as has been done with factor score regression in `lavaan`. Second, whereas fit indices can still be obtained for the separate measurement models in the first stage of 2S-PA, as with other models using individual likelihood (e.g., random slope models, IRT with maximum likelihood), conventional SEM fit indices could not be obtained for the structural model. It is however still possible to compare models using the likelihood ratio test. On the same note, it should be pointed out that existing cutoffs on fit indices for SEM models were mostly based on simulation studies on the measurement models (e.g., Hu & Bentler, 1998, 1999), whereas other studies have shown that fit indices performed differently for misspecification in the path coefficients (e.g., Fan & Sivo, 2007). In the structural model, even though constraining some paths or covariances to be zero may give better fit indices due to an increase in degrees of freedom of the model, those constraints may cause misspecification that leads to biased estimates of structural coefficients of interest. Therefore, we recommend that researchers use a saturated structural model except for paths that should be constrained based on theoretical and conceptual reasons (see Kenny et al., 2015).

In addition, the simulation studies in this paper do not capture the diversity of models that researchers use in SEM, such as growth curve analyses, latent interactions, and so forth. Therefore, future studies are needed to further extend the 2S-PA method to these models. Also, we considered only one type of misspecification where indicators of two latent variables have unmodeled association, so future studies are needed to examine the performance of 2S-PA under other types of misspecification in the measurement models and its sensitivity to misspecification in the structural model.

Like other reliability adjustment methods such as factor score regression (Devlieger et al., 2016) and reliability adjustment for interaction effects (Hsiao et al., 2018), the proposed 2S-PA approach does not fully take into account the uncertainty in the estimated standard errors of measurement in the first stage as they are assumed known when used in the second stage (cf. Cole & Preacher, 2014). Although, as demonstrated in Yang et al. (2012) and our simulations, the impact of omitting that uncertainty is generally minimal with moderate to large sample sizes, it is likely responsible for the biases of 2S-PA in small samples, even though 2S-PA mostly still outperformed full SEM based on our results. Future research effort to develop small-sample corrections would greatly improve 2S-PA. Although we propose an ad hoc Bayesian solution in Mplus with weakly informative priors to mitigate the bias, the standard error of the factor scores are obtained as a separate step with plausible value imputation and limited iterations; future research can explore alternative priors and the use of more general Bayesian programs such as STAN (Stan Development Team, 2020). Alternatively, a Bayesian approach that takes the uncertainty of these estimates into account by assigning a prior probability on the estimated standard errors of measurement may further improve the approach discussed in this paper (see Levy, 2017, for a recently proposed Bayesian solution with continuous indicators). Another reason for the bias observed in 2S-PA in small samples and low-reliability conditions is that, for extreme factor scores, their sampling distributions may be highly skewed so that the normal approximation is not reasonable. Possible solutions for future explorations include using the width of asymmetric confidence intervals to quantify the measurement error, relaxing the normality assumption with a skewed distribution, and Bayesian methods that directly use the full posterior distributions of factor scores.

Finally, it should also be pointed out that the 2S-PA approach is similar to the recent development in mixture modeling for adjusting for measurement error in the assignment of class membership (Asparouhov & Muthén, 2014; Bolck et al., 2004; Vermunt, 2010). Future studies can explore the possibility of a unifying framework for reliability adjustment that accommodates continuous and categorical latent variables.

References

Asparouhov, T., & Muthén, B. (2012). *Comparison of computational methods for high dimensional item factor analysis* (tech. rep.) [Unpublished manuscript retrieved from https://www.statmodel.com].

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 329–341. https://doi.org/10.1080/10705511.2014.915181

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: one-step versus three-step estimators. *Political Analysis*, *12*(1), 3–27. https://doi.org/10.1093/pan/mph001

Bollen, K. A. (1989). *Structural equations with latent variables* (Vol. 8). John Wiley & Sons, Inc.

Bollen, K. A. (1996). An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika*, *61*(1), 109–121. https://doi.org/10.1007/BF02296961

Bollen, K. A. (2019). Model implied instrumental variables (MIIVs): An alternative orientation to structural equation modeling. *Multivariate Behavioral Research*, *54*(1), 31–46. https://doi.org/10.1080/00273171.2018.1483224

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

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*(1), 33–57. https://doi.org/10.1007/s11336-009-9136-x

Caroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective*. Chapman & Hall/CRC.

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

Chalmers, R. P. (2020). *Simdesign: Structure for organizing monte carlo simulation designs* [R package version 2.0.1]. https://CRAN.R-project.org/package=SimDesign

Cheung, M. W.-L. (2013). Multivariate meta-analysis as structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(3), 429–454. https://doi.org/10.1080/10705511.2013.797827

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. A. (2002). Using predicted latent scores in general latent structure models. In G. A. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure models* (pp. 195–244). Lawrence Erlbaum.

Croon, M. A., & van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, *12*(1), 45–57. https://doi.org/10.1037/1082-989X.12.1.45

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*, 1–28. https://doi.org/10.18637/jss.v048.c01

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*, *13*(Supplement 1), 31–38. https://doi.org/10.1027/1614-2241/a000130

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*(4), 341–349. https://doi.org/10.1037/1040-3590.8.4.341

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

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika*, *82*(4), 904–927. https://doi.org/10.1007/s11336-017-9557-x

Estabrook, R., & Neale, M. C. (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

Fairchild, A. J., MacKinnon, D. P., Taborga, M. P., & Taylor, A. B. (2009). R^2 effect-size measures for mediation analysis. *Behavior Research Methods*, *41*(2), 486–498. https://doi.org/10.3758/BRM.41.2.486

Falk, C. F., & Cai, L. (2016). Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis. *Psychometrika*, *81*(2), 434–460. https://doi.org/10.1007/s11336-014-9428-7

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, *42*(3), 509–529. https://doi.org/10.1080/00273170701382864

Fisher, Z., Bollen, K., Gates, K., & Rönkkö, M. (2020). *Miivsem: Model implied instrumental variable (miiv) estimation of structural equation models* [R package version 0.5.5]. https://CRAN.R-project.org/package=MIIVsem

Fuller, W. A. (1987). *Measurement error models*. Wiley.

Greene, W. H. (2003). *Econometric analysis* (5th ed.). Prentice Hall.

Hallgren, K. A., McCabe, C. J., King, K. M., & Atkins, D. C. (2019). Beyond path diagrams: Enhancing applied structural equation modeling research through data visualization. *Addictive Behaviors*, *94*, 74–82. https://doi.org/10.1016/j.addbeh.2018.08.030

Hardin, J. W. (2002). The robust variance estimator for two-stage models. *The Stata Journal: Promoting communications on statistics and Stata*, *2*(3), 253–266. https://doi.org/10.1177/1536867X0200200302

Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances*. Johns Hopkins University Press.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling an overview and a meta-analysis. *26*(3), 329–367. https://doi.org/10.1177/0049124198026003003

Hoshino, T., & Bentler, P. (2013). Bias in factor score regression and a simple solution. In A. R. de Leon & K. C. Chough (Eds.), *Analysis of mixed data: Methods & applications* (pp. 43–61). Chapman & 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

Hu, L.-t., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453. https://doi.org/10.1037/1082-989X.3.4.424

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jang, Y., Chiriboga, D. A., & Small, B. J. (2008). Perceived discrimination and psychological well-being: The mediating and moderating role of sense of control. *The International Journal of Aging and Human Development*, *66*(3), 213–227. https://doi.org/10.2190/AG.66.3.c

Jin, S., Luo, H., & Yang-Wallentin, F. (2016). A simulation study of polychoric instrumental variable estimation in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 680–694. https://doi.org/10.1080/10705511.2016.1189334

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika*, *57*(2), 239–251. https://doi.org/10.1093/biomet/57.2.239

Jöreskog, K. G., & Sörbom, D. (2001). *LISREL 8: User's reference guide* (2nd ed.). Scientific Software International.

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

Kelley, K. (2020). *MBESS: The MBESS R package* [R package version 4.7.0]. https://CRAN.R-project.org/package=MBESS

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods and Research*, *44*(3), 486–507. https://doi.org/10.1177/0049124114543236

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford.

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

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Lord, F. M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement*, *21*(3), 239–243. https://doi.org/10.1111/j.1745-3984.1984.tb01031.x

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, *58*, 593–614. https://doi.org/10.1146/annurev.psych.58.110405.085542

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*(1), 83–104. https://doi.org/10.1037/1082-989X.7.1.83

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

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713. https://doi.org/10.1007/s11336-005-1295-9

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*(3), 259–284. https://doi.org/10.1037/1082-989X.10.3.259

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Murphy, K. M., & Topel, R. H. (1985). Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, *3*(4), 370. https://doi.org/10.2307/1391724

Muthén, B. O., & Asparouhov, T. (2002). *Modeling of heteroscedastic measurement errors* (tech. rep.). https://www.statmodel.com/download/webnotes/mc3.pdf

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Neale, M. C. (2000). Individual fit, heterogeneity, and missing data in multigroup structural equation modeling. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches and specific examples* (pp. 249–267). Lawrence Erlbaum.

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

Nestler, S. (2013). A Monte Carlo study comparing PIV, ULS and DWLS in the estimation of dichotomous confirmatory factor analysis: *Dichotomous confirmatory factor analysis. British Journal of Mathematical and Statistical Psychology*, *66*(1), 127–143. https://doi.org/10.1111/j.2044-8317.2012.02044.x

Paek, I., Cui, M., Öztürk Gübeş, N., & Yang, Y. (2018). Estimation of an IRT model by Mplus for dichotomously scored responses under different estimation methods. *Educational and Psychological Measurement*, *78*(4), 569–588. https://doi.org/10.1177/0013164417715738

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo Confidence Intervals for Indirect Effects. *Communication Methods and Measures*, *6*(2), 77–98. https://doi.org/10.1080/19312458.2012.679848

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

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Reiersøl, O. (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica*, *18*(4), 375. https://doi.org/10.2307/1907835

Revelle, W. (2019). *Psych: Procedures for psychological, psychometric, and personality research* [R package version 1.9.12]. Northwestern University. Evanston, Illinois. https://CRAN.R-project.org/package=psych

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315

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

Rosseel, Y., Jorgensen, T. D., & Rockwood, N. (2020). lavaan: Latent variable analysis [R package version 0.6-5]. https://CRAN.R-project.org/package=lavaan

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*. https://doi.org/10.1002/j.2333-8504.1968.tb00153.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

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*(4), 563–575. https://doi.org/10.1007/BF02296196

Stan Development Team. (2020). Stan user's guide [Version 2.25]. https://mc-stan.org

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*(4), 450–469. https://doi.org/10.1093/pan/mpq025

Wansbeek, T., & Meijer, E. (2000). *Measurement error and latent variables*. North-Holland.

Wilcox, R. R. (2016). *Introduction to robust estimation and hypothesis testing* (2nd ed.). Academic Press.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58–79. https://doi.org/10.1037/1082-989X.12.1.58

Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in item response theory scale scores. *Educational and Psychological Measurement*, *72*(2), 264–290. https://doi.org/10.1177/0013164411410056

Table 1

Percentage Relative Bias of the Path Coefficient ($\beta_1 = 0.5$) in Study 1.

		PA		SEM		RA-α		RA-ω		RA-ω^{NL}		2S-PA	
N/p	p	$K = 2$	$K = 4$	$K = 2$	$K = 4$	$K = 2$	$K = 4$	$K = 2$	$K = 4$	$K = 2$	$K = 4$	$K = 2$	$K = 4$
						$\bar{\lambda} = 1$							
6	5	**-30.64**	**-26.20**	-4.30	**-7.94**	1.33	1.43	**-5.82**	-4.64	**-6.82**	-4.71	**-25.16**	**-19.60**
	10	**-21.63**	**-17.36**	-3.84	-3.43	-1.11	-1.13	-2.35	-2.32	-4.24	-3.26	**-8.90**	**-5.74**
	20	**-13.08**	**-10.01**	-1.80	-1.45	-1.56	-1.19	-2.08	-1.63	-2.86	-1.85	-2.11	-1.46
25	5	**-31.86**	**-26.22**	-1.29	-0.98	2.11	1.75	-2.58	-1.70	-3.56	-2.38	**-10.20**	**-6.15**
	10	**-20.74**	**-16.23**	-0.82	-0.55	-0.56	-0.26	-1.67	-1.22	-1.08	-0.75	-1.86	-1.11
	20	**-12.81**	**-9.70**	-0.70	-0.48	-1.33	-0.89	-1.76	-1.27	-0.79	-0.58	-0.63	-0.48
100	5	**-31.75**	**-26.02**	-0.46	-0.46	1.57	1.40	-1.55	-1.24	-1.97	-1.61	-2.60	-1.60
	10	**-20.70**	**-16.16**	-0.45	-0.29	-0.57	-0.22	-1.56	-1.10	-0.34	-0.30	-0.53	-0.38
	20	**-12.65**	**-9.51**	-0.39	-0.17	-1.18	-0.70	-1.59	-1.06	-0.22	-0.15	-0.17	-0.08
						$\bar{\lambda} = 2.5$							
6	5	**-17.28**	**-14.68**	-0.59	-4.22	**-5.04**	**-5.36**	**-6.81**	**-6.88**	**-8.75**	**-8.59**	**-6.13**	**-5.76**
	10	**-12.36**	**-9.70**	-2.39	-2.08	**-6.24**	**-5.23**	**-6.70**	**-5.65**	**-8.56**	**-6.65**	-2.46	-1.92
	20	**-8.48**	**-6.91**	-0.89	-0.84	**-5.39**	-4.67	**-5.56**	-4.83	**-6.63**	**-5.34**	-0.64	-0.80
25	5	**-15.92**	**-12.48**	-0.46	-0.61	-3.85	-3.36	**-5.19**	-4.58	**-6.28**	**-5.26**	-1.45	-1.19
	10	**-10.92**	**-8.59**	-0.48	-0.42	-4.85	-4.12	**-5.23**	-4.49	**-5.52**	-4.62	-0.36	-0.39
	20	**-8.32**	**-6.71**	-0.54	-0.45	**-5.24**	-4.47	**-5.39**	-4.62	**-5.50**	-4.61	-0.21	-0.36
100	5	**-15.72**	**-12.28**	-0.40	-0.43	-3.76	-3.22	-4.98	-4.37	**-5.51**	-4.78	-0.49	-0.44
	10	**-10.72**	**-8.43**	-0.24	-0.21	-4.65	-3.96	**-5.02**	-4.33	-4.92	-4.24	-0.03	-0.07
	20	**-8.14**	**-6.55**	-0.32	-0.26	**-5.06**	-4.31	**-5.21**	-4.45	**-5.10**	-4.32	-0.06	-0.04

Note. p = number of indicators for the latent predictor K = number of indicator categories. $\bar{\lambda}$ = average factor loading. PA = linear regression/path analysis. SEM = structural equation model. RA = reliability adjustment method (with α, ω, and ω^{NL} coefficients). 2S-PA = two-stage path analysis with definition variable with maximum likelihood. The results represent averages across conditions. Numbers larger than 5 (in absolute values) are bolded.

Table 2

Percentage Relative Standard Error Bias of Path Coefficient in Study 1.

		PA		SEM		RA-α		RA-ω		RA-ω^{NL}		2S-PA	
N/p	p	$K = 2$	$K = 4$	$K = 2$	$K = 4$	$K = 2$	$K = 4$	$K = 2$	$K = 4$	$K = 2$	$K = 4$	$K = 2$	$K = 4$
6	5	-5.30	-5.11	**-15.75**	**-15.32**	-5.72	-4.98	-4.92	-5.28	-5.30	-4.89	-9.65	-6.95
	10	-2.12	-1.34	-7.43	-6.17	-0.85	-0.65	-1.03	-0.50	-0.97	-0.83	-2.96	-3.13
	20	-0.50	-0.27	-3.28	-3.29	-0.12	0.40	-0.28	0.09	0.13	0.93	-2.47	-3.20
25	5	-1.75	-1.58	-6.39	-5.52	-1.55	-1.31	-1.53	-1.73	-1.31	-1.04	-4.56	-5.03
	10	0.62	0.71	-1.69	-0.47	1.05	1.47	1.13	2.41	1.04	1.52	-0.95	-0.28
	20	3.03	1.05	1.42	1.87	2.64	3.22	3.15	3.37	2.79	2.48	0.23	0.66
100	5	2.82	0.81	0.15	-0.88	2.63	2.03	2.23	2.15	2.08	2.36	-0.42	-1.37
	10	1.54	1.31	0.09	1.34	2.37	3.54	2.60	3.43	2.33	3.34	-0.81	-0.10
	20	-0.84	-0.96	-2.05	-2.42	-0.73	-1.27	-0.90	0.97	-0.24	-2.46	-2.09	-2.47

Note. p = number of indicators for the latent predictor K = number of indicator categories. PA = linear regression. SEM = structural equation model. RA = reliability adjustment method (with α, ω, and ω^{NL} coefficients). 2S-PA = two-stage path analysis with definition variable using Mplus with maximum likelihood. The numbers are averages across multiple conditions. Numbers larger than 10 (in absolute values) are bolded.

Table 3

Empirical Coverage Percentages of Indirect Effect in Study 3.

a	b	N	PA	RA-α	SEM	2S-PA	2S-PA (Bayes)
$\bar{\lambda} = 1$							
.00	.00	30	99.6	100.0	96.6	99.4	99.9
		125	99.8	99.9	98.5	99.7	99.8
		500	99.8	99.9	99.7	99.8	99.8
.39	.00	30	99.0	100.0	96.0	99.0	99.7
		125	97.0	99.1	94.4	97.9	98.1
		500	93.6	95.5	93.5	94.9	94.5
.00	.39	30	98.7	99.9	94.0	99.3	99.6
		125	97.9	98.3	93.1	98.1	98.0
		500	95.3	95.4	93.0	94.8	94.5
.39	.39	30	**56.4**	97.0	**90.5**	**87.6**	96.3
		125	**6.0**	94.4	**91.5**	**86.6**	**89.1**
		500	**0.0**	95.3	95.0	95.3	94.6
$\bar{\lambda} = 2.5$							
.00	.00	30	99.4	99.5	96.3	99.2	99.3
		125	99.9	99.9	99.7	99.8	99.8
		500	99.9	99.9	99.9	99.8	99.8
.39	.00	30	96.1	97.2	**89.4**	96.3	97.1
		125	94.4	94.7	92.6	94.2	94.2
		500	93.9	94.6	94.3	94.6	94.6
.00	.39	30	97.3	97.7	**90.0**	97.4	97.5
		125	95.5	95.5	93.5	94.8	94.8
		500	94.4	94.4	94.0	93.5	93.5
.39	.39	30	**81.1**	92.0	**88.4**	92.7	94.2
		125	**58.2**	**91.8**	93.1	94.3	94.5
		500	**7.7**	**87.0**	94.5	94.2	94.3

Note. p = number of indicators per latent variable. a = population coefficient of predictor to mediator. b = population coefficient of mediator to outcome. $\bar{\lambda}$ = average factor loading. PA = path analysis. RA-α = reliability adjustment method with α. SEM = structural equation model. 2S-PA = two-stage path analysis with definition variable with maximum likelihood (Bayesian) estimation in the first stage. Values below 92% are bolded.

Table 4

Parameter Estimates of the Empirical Demonstration With Four Different Approaches.

	a (SE)	*b* (SE)	*c* (SE)	*ab* [95% CI]
PA	-0.156 (0.017)	0.445 (0.014)	-0.085 (0.015)	-0.069 [-0.085, -0.054]
SEM	-0.182 (0.020)	0.479 (0.013)	-0.095 (0.017)	-0.087 [-0.107, -0.068]
RA-α	-0.176 (0.019)	0.501 (0.016)	-0.081 (0.017)	-0.088 [-0.108, -0.069]
2S-PA	-0.189 (0.020)	0.472 (0.015)	-0.105 (0.018)	-0.089 [-0.109, -0.070]

Note. N = 3,547, The *a*-path was Perceived Discrimination to Sense of Control. The *b*-path was Sense of Control to Positive Affect. The *c*-path was Perceived Discrimination to Positive Affect. *ab* = indirect effect estimate. PA = Path analysis with composite scores as error-free observed variables. RA-α = reliability adjustment of PA with reliability coefficieint α. 2S-PA = two-stage path analysis with definition variables. The 95% CIs for *ab* were obtained with the Monte Carlo method.

Figure 1

Full SEM specification of linear regression with a latent predictor and an observed outcome.

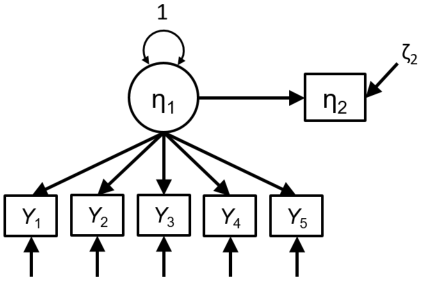
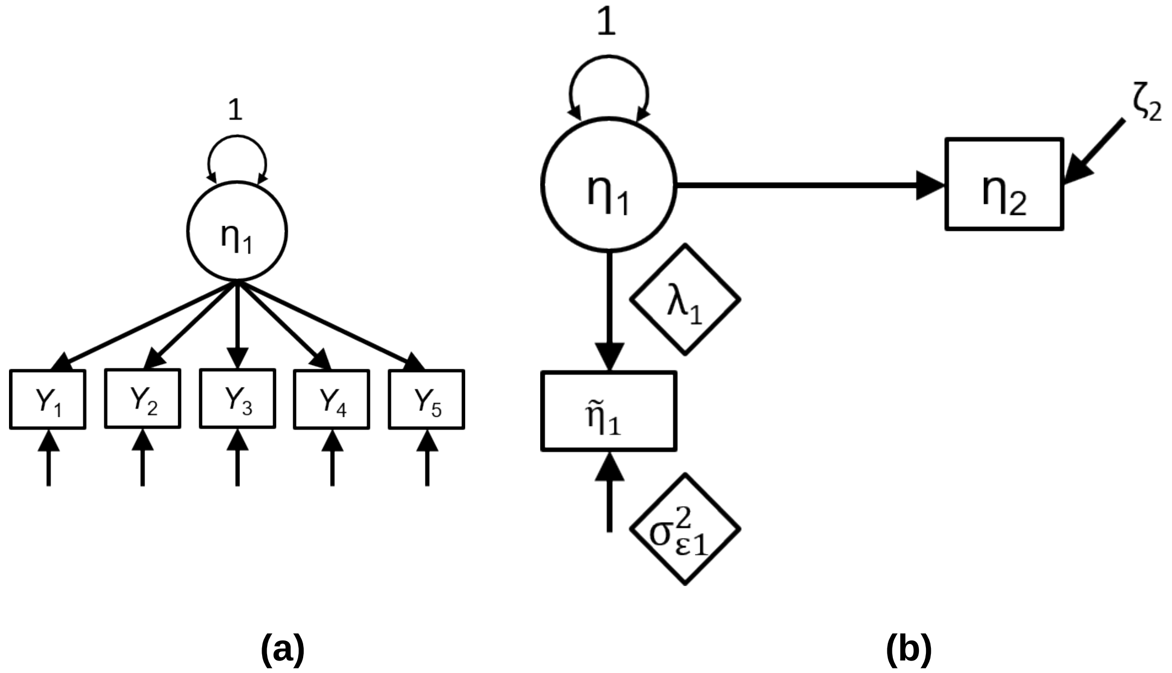


Figure 2

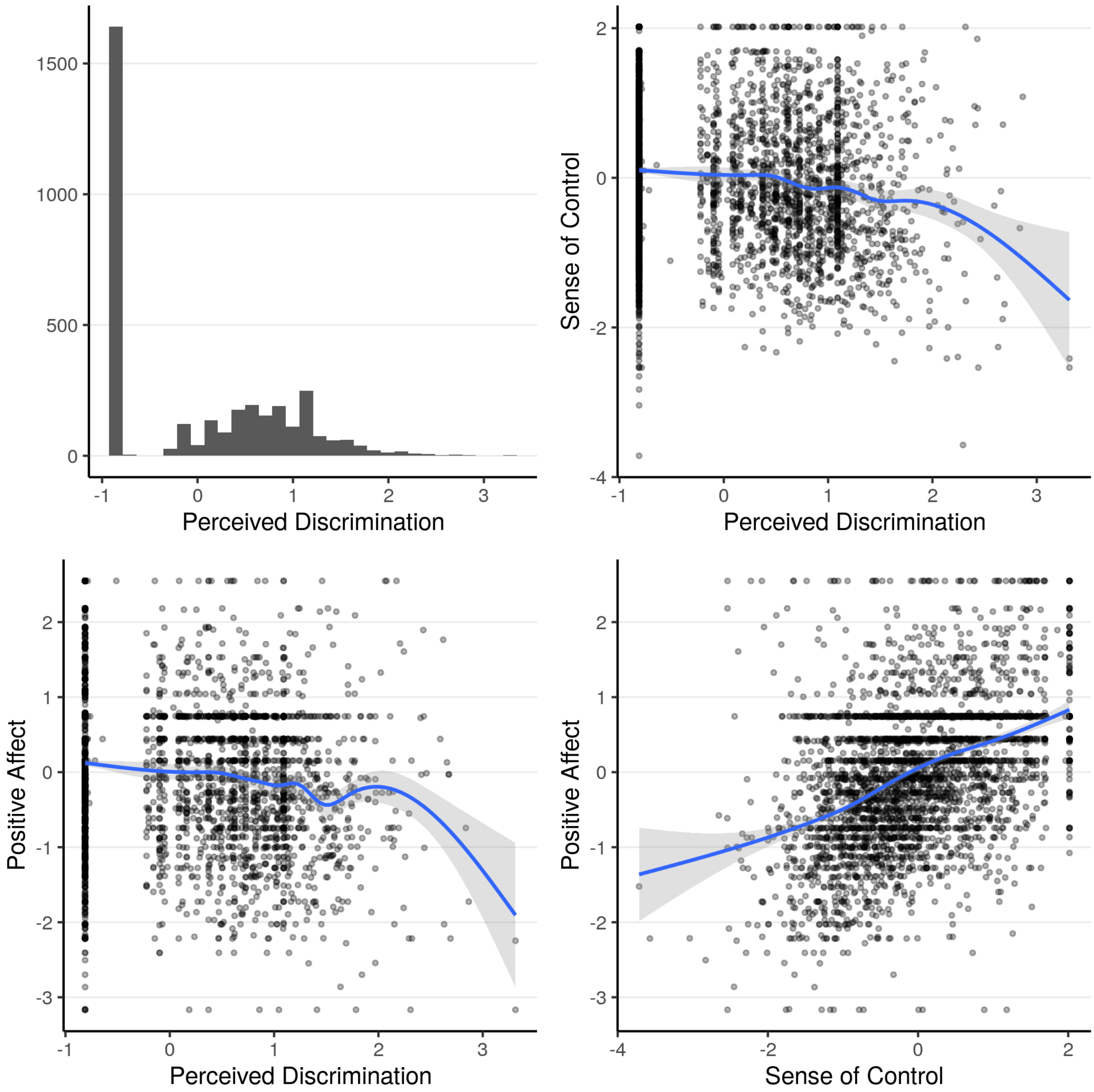
Linear regression with definition variables.



Note. (a) Stage 1: a measurement model for estimating factor scores $\tilde{\eta}_1$ and the corresponding standard errors; (b) Stage 2: path analysis with constraints to fix measurement error variance using definition variables.

Figure 3

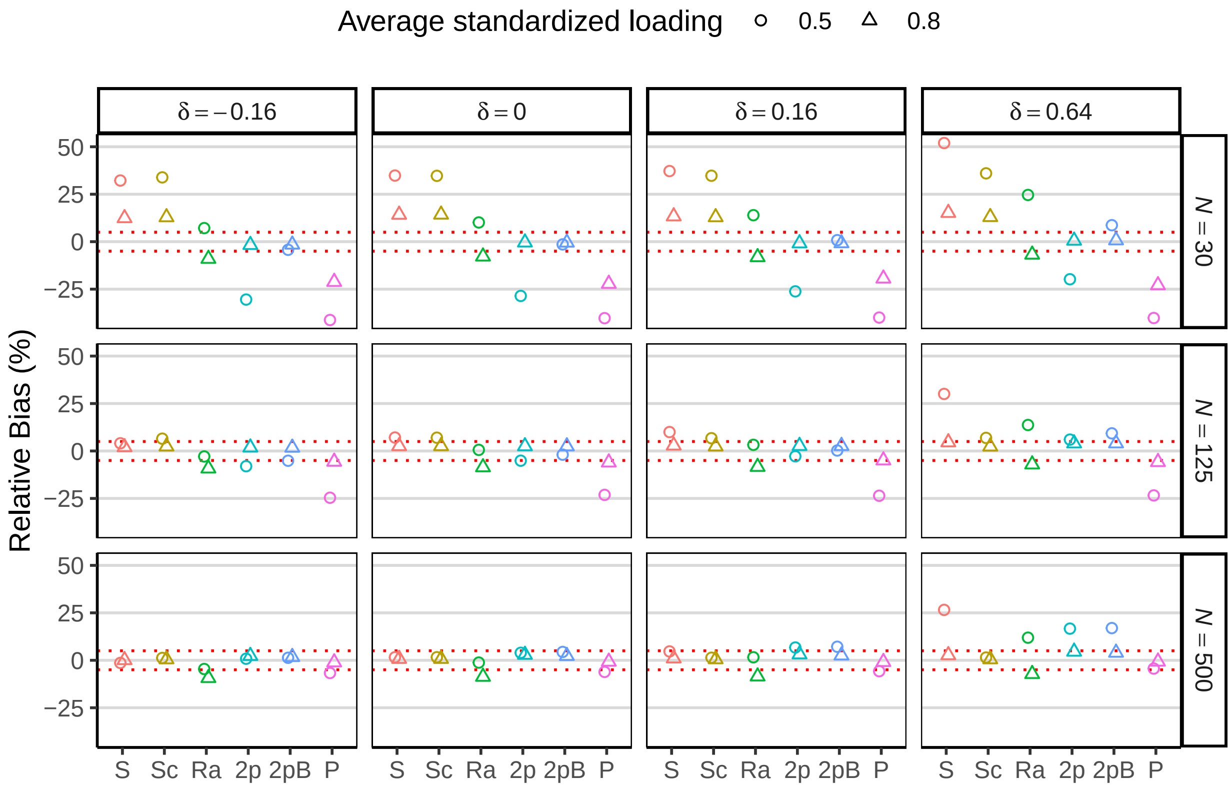
Relations among estimated factor scores for the empirical demonstration.



Note. The distribution of the estimated factor scores for the latent predictor was shown in the top left panel.

Figure 4

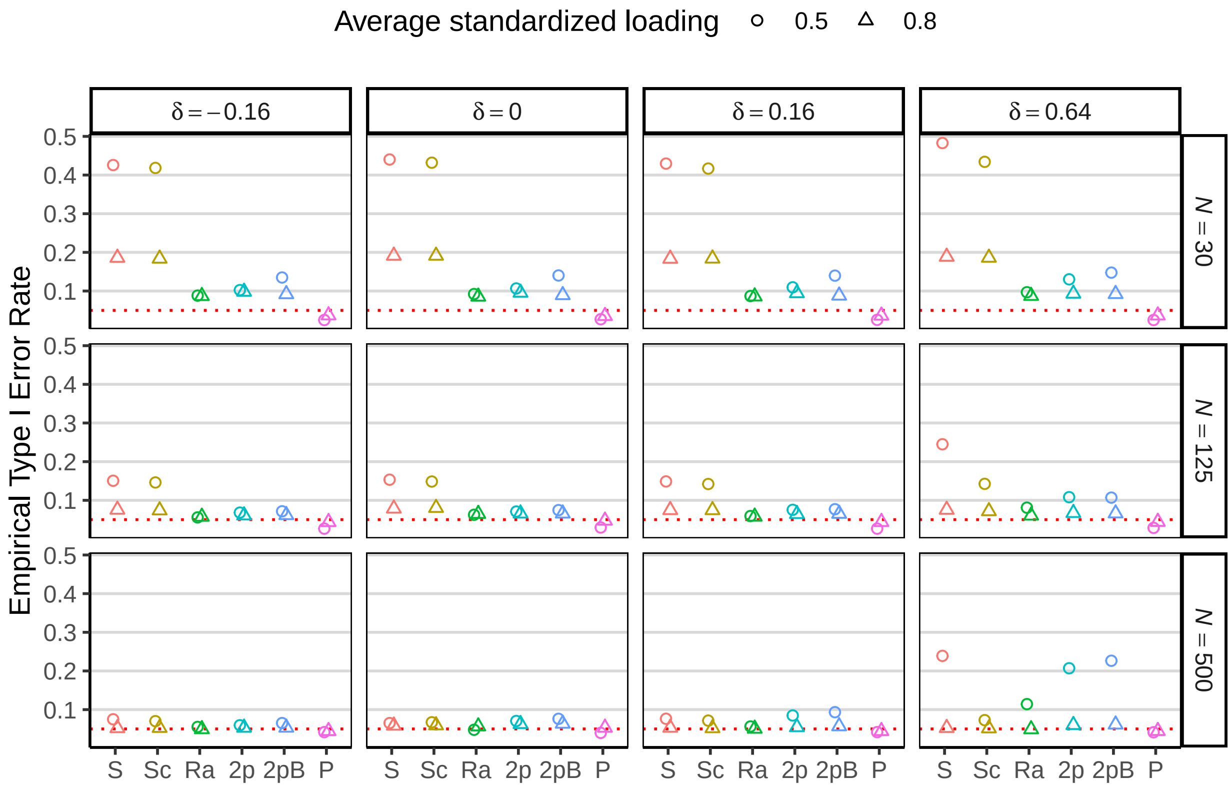
Relative bias of a non-zero path coefficient in Study 2.



Note. δ = unique factor correlation between the third indicators of the latent predictor and the latent outcome. S = structural equation model without unique factor covariance. Sc = SEM with unique factor covariance. 2p = two-stage path analysis with definition variables with maximum likelihood in the first stage. Ra = reliability adjustment with coefficient α. 2pB = 2S-PA with Bayesian estimation in the first stage. P = Polychoric instrumental variable estimator. Values between the two dotted lines (± 5%) were considered to have acceptable bias.

Figure 5

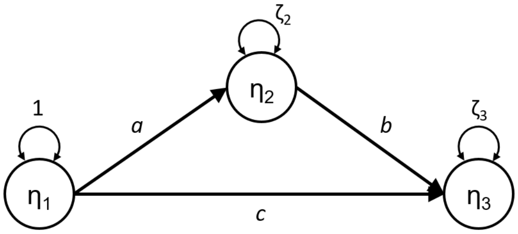
Empirical Type I error rates in Study 2.



Note. δ = unique factor correlation between the third indicators of the latent predictor and the latent outcome. S = structural equation model without unique factor covariance. Ra = Reliability adjustment with coefficient α. Sc = SEM with unique factor covariance. 2p = two-stage path analysis with definition variables with maximum likelihood in the first stage. 2pB = 2S-PA with Bayesian estimation in the first stage. P = Polychoric instrumental variable estimator. The dotted line shows the nominal value of .05.

Figure 6

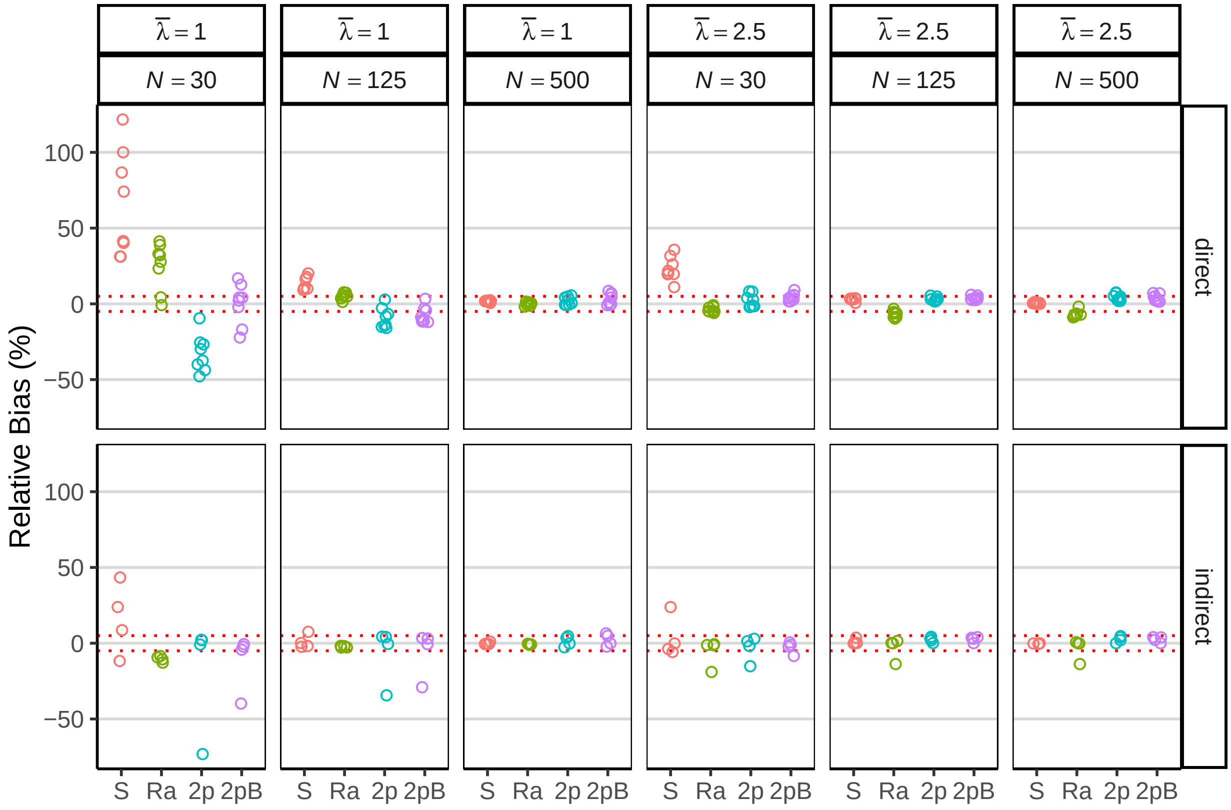
Mediation model for Study 3.



Note. Each latent variable was measured by five categorical indicators (which were not presented in the graph).

Figure 7

Percentage relative bias of non-zero direct (a = .39, b = .39, and c = .15) and indirect effects (ab = .39^2) in Study 3.



Note. $\bar{\lambda}$ = average factor loading. S = structural equation model without unique factor covariance. Ra = reliability adjustment with coefficient α. 2p = two-stage path analysis with definition variables with maximum likelihood (Bayesian) estimation in the first stage. 2pB = 2S-PA with Bayesian estimation in the first stage. Values between the two dotted lines (± 5%) were considered to have acceptable bias.

Appendix A

Measurement Error of Factor Scores With Categorical Indicators

This Appendix provides a simple demonstration that the error variance of the factor score is heterogeneous under the factor model for categorical data defined in equation (4), even though the error variance for the underlying latent response variates were assumed constant such that $\mathrm{Var}(\epsilon_i) = \theta_\epsilon$ for all is. For simplicity, we assume $\lambda = 1$, which was one of the values used in our simulation conditions, and that the test has only one binary item without loss of generality. It is sufficient to show that the error variance of factor score depends on the observed item response. We also assume that the expected a posteriori (EAP) score is used as a factor score, but the heterogeneity applies to essentially all types of factor scores.

Based on the above model, the EAP score can be obtained as the posterior mean of η given the observed data $Y = y$. By Bayes's theorem, the posterior distribution of $\eta|y$ is

$$P(\eta|y) = \frac{\pi(\eta)P(Y = y|\eta)}{\int_{-\infty}^{\infty} \pi(h)P(Y = y|h)\,\mathrm{d}h},$$

and the EAP score is the expected value of $\eta|y$. Often, $\pi(\eta)$ is chosen to be $N(0, 1)$ to match the scaling of the latent variable.

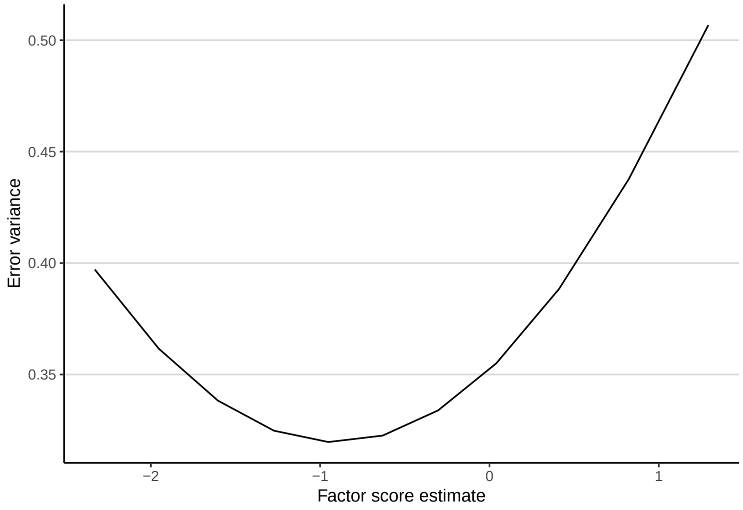
The error variance of the EAP score is the posterior variance of $\eta|y$:

$$\mathrm{Var}(\eta|y) = \mathrm{E}(\eta^2|y) - [\mathrm{E}(\eta|y)]^2,$$

where $\mathrm{E}(\eta^m|y) = \int_{-\infty}^{\infty} \eta^m P(\eta|y)\,\mathrm{d}\eta$. In general, the above expression depends on y such that $\mathrm{Var}(\eta|y)$ is different for different response patterns, except in some special cases such as when $\tau = 0$ or when $P(Y|\eta)$ is normal. To illustrate, if $\tau = 2.20$ (one of the values used in our simulation conditions), which corresponds to $P(Y = 1|\eta = 0) = 0.9$, using numerical integration to evaluate $\mathrm{Var}(\eta|y)$, the error variance for the EAP score is 0.91 when $y = 0$, and 0.87 when $y = 1$.

The graph below shows the association between the factor score estimates and the corresponding error variance where there are 10 items, assuming that the measurement parameters are known, $\lambda = 1$, and other parameters as specified in Study 1.

Figure A1



Appendix B

More Details of 2S-PA with Bayes

To reduce the small-sample bias found in 2S-PA in Study 1, we tested a Bayesian variant that used Bayesian estimations in the first stage for obtaining factor scores. Specifically, we incorporated Bayesian priors by assigning a normal prior with mean of 0 and *SD* of $\sqrt{5}$ to the loadings (which was the default in Mplus) to stabilize the parameter estimates. Note that the probit link was used in Bayesian estimation, which is the default in Mplus, as opposed to the logit link in maximum likelihood estimation. Therefore, the priors on the loadings were considered weakly informative priors. For other parameters, we used the default priors in Mplus, which were uniform on the real line for thresholds and means, and uniform on the positive real line for variance parameters.

For each measurement model, we used Markov Chain Monte Carlo (with Gibbs sampling) with two chains to perform fully Bayesian estimations. Gibbs sampling stopped when the potential scale reduction factor dropped below 1.01, or when it reached 500,000 iterations. For each observation, we obtained the factor scores and the corresponding *SE*s as the means and *SD*s of 200 draws from the posterior predictive distributions of the latent variable, with a thinning interval of 10.

For simulated data in Study 1, the priors drastically reduced the bias to -3.59% for the worst condition, and also improved convergence rate for conditions with small sample sizes.

These regularizing estimates can similarly be obtained using the mirt package in R, which treated the input priors as penalty terms to obtain penalized maximum likelihood estimates for measurement parameters and factor scores. See the sample Mplus syntax and R code in the supplemental materials for carrying out 2S-PA with Bayes for the empirical example.

Appendix C

Polychoric Instrumental Variable (PIV) Estimator With Model-Implied Instrumental Variables

We used the R package *MIIVsem* (Version 0.5.5, Fisher et al., 2020) to perform PIV estimations and obtained estimates for the standardized latent regression coefficient. Based on the theory of instrumental variable estimation and the simulation results from Nestler (2013) and Jin et al. (2016), for each equation, the PIV estimator is consistent under certain model misspecifications such as the omitted unique covariances in Study 2. However, unlike other methods in the study, PIV requires a scaling indicator (i.e., with loading set to 1) for each latent factor, and in this case the first indicator was used for that purpose. The software automatically identified model-implied instrumental variables (IVs) for each estimating equation: for estimating loadings, the IVs are all other indicators that are not scaling indicators; for the latent regression coefficient, the IVs are the non-scaling indicators for η_1. Because the scaling of the latent variables in PIV is different from other methods, we also obtained the standardized latent regression coefficient estimate as

$$\hat{\beta} = \hat{b} \times \frac{\sqrt{\hat{\mathrm{Var}}(\eta_1)}}{\sqrt{\hat{\mathrm{Var}}(\eta_1)\hat{b}^2 + \hat{\zeta}}},$$

where $\hat{b}$, $\hat{\mathrm{Var}}(\eta_1)$, and $\hat{\zeta}$ are the estimates of unstandardized path coefficient, variance of the latent predictor, and disturbance of the latent outcome from MIIVsem. At the time of writing, however, MIIVsem does not provide the estimates of variance parameters by default. Using the `var.cov = TRUE` option would provide the point estimates of the variance parameters based on the diagonally weighted least square estimations, but it does not provide the asymptotic covariance matrix of the variance parameter estimates, which are needed to apply the delta method to compute the *SE* of $\hat{\beta}$. Therefore, we followed equations (26) to (31) in Bollen and Maydeu-Olivares (2007, p. 315) to obtain the *unweighted least squares* estimates of $\mathrm{Var}(\eta_1)$ and ζ, and the corresponding asymptotic covariance matrix. The formulas in Bollen and Maydeu-Olivares (2007) did not cover the covariances between $\hat{b}$ and $(\hat{\mathrm{Var}}[\eta_1], \hat{\zeta})$, which are also needed to apply the delta method, so we compute them as, following equation (31) of Bollen and Maydeu-Olivares (2007) on p. 315,

$$\widehat{\mathrm{Acov}}(\hat{\theta}_1, \hat{\theta}_2) = \frac{1}{N}\hat{\mathbf{K}}^{*}\hat{\boldsymbol{\Sigma}}_{\rho\rho}\hat{\mathbf{K}}^{*},$$

where $\hat{\mathbf{K}}^{*} = [\mathbf{K}^\top | \hat{\mathbf{H}}_2^\top (\mathbf{I} - \hat{\boldsymbol{\Delta}}_1 \hat{\mathbf{K}})^\top]^\top$, and all other matrices were defined in Bollen and Maydeu-Olivares (2007). The R code for carrying out the delta method estimation of the standardized path coefficient can be found in the supplemental materials.