Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

Corrections to Structural Goodness-Of-Fit Test Statistics: A Monte Carlo Investigation of Available Methods

Graham G. Rifenbark [*†] and Terrence D. Jorgensen [‡]

†Wisconsin Center for Education Research, School of Education, University of Wisconsin-Madison, Madison, 53703, WI, USA
‡Methods and Statistics Programme, Faculty of Social and Behavioural Sciences, University of Amsterdam, Amsterdam, The Netherlands
*Corresponding author. Email: rifenbark@wisc.edu

Abstract

Researchers have proposed procedures to construct structural fit indices (SFIs) due to the shortcomings of global fit indices (GFIs) when evaluating the structural component of structural equation models (SEMs). We focus on test statistics used to calculate SFIs. A two-step procedure proposed by Hancock and Mueller (2011) first estimates latent (co)variances (Φ) using confirmatory factor analysis, used as input data to subsequently fit a structural model using path analysis. This procedure results in a *pseudo*-χ^2 test statistic that can be used to construct structural versions of GFIs, but ignoring uncertainty about Φ inflates Type-I error rates, performing worse with less reliable indicators, more indicators per factor, and smaller samples (Cao et al., 2023; Heene et al., 2024; Rifenbark, 2019). Modeling and empirical solutions have been investigated (Rifenbark & Jorgensen, 2023) to improve two-stage SFIs, using the *structural-after-measurement* (SAM) framework (Rosseel & Loh, 2024). Zhang and Wu (2024) proposed post-hoc corrections that consider various computations of information and asymptotic covariance matrices of Φ, some of which are similar to adjustments implemented in `lavaan` for the SAM approach. We use Monte Carlo simulation to compare the performance of these advancements and report results for varying levels of reliability, sample size, and levels of measurement model misspecification. We hypothesized the new methods would result in nominal Type-I error rates for adjusted χ^2 statistics, compared to the uncorrected pseudo-χ^2 statistic.

Keywords: structural equation modeling, structural-after-measurement, two-stage estimation, goodness-of-fit

1. Introduction

In the social sciences it is common for researchers to employ a structural equation model (SEM) to test hypothesized relations among latent variables which are not directly observable. These unobservable phenomena are captured by a collection of observed (or manifest) variables that are thought to measure the intended latent variable(s). A confirmatory factor analysis (CFA) is a special case of SEM, in which each manifest variable is modeled as being composed of variance due to (at least one) common factor and a unique factor. This decomposition is possible when multiple manifest variables $\boldsymbol{\gamma}$ are *indicators* of the same common factor η.

A "full SEM" is made up of two components. First, a measurement component (i.e., the CFA) models the relations between p manifest variables and $k \leq p$ latent variables:

$$\boldsymbol{\gamma} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\nu}$ is a $p \times 1$ vector of indicator intercepts; $\boldsymbol{\Lambda}$ is a $p \times k$ matrix of factor loadings (linear regression slopes) relating indicators to latent variables; and $\boldsymbol{\varepsilon}$ is a $p \times 1$ vector of indicator residuals (the unique factors). Second, a structural component which models the relations among latent variables:

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\zeta}, \tag{2}$$

where $\boldsymbol{\alpha}$ is a $k \times 1$ vector of latent-variable intercepts; $\mathbf{B}$ is a $k \times k$ matrix (with diagonal constrained to zero) of linear regression slopes relating latent variables to each other; and $\boldsymbol{\zeta}$ is a $k \times 1$ vector of latent-variable residuals (disturbances). It is assumed that $\boldsymbol{\varepsilon}$ is multivariate normally distributed: $\boldsymbol{\varepsilon} \sim (0, \boldsymbol{\Theta})$. Likewise, it is assumed that $\boldsymbol{\zeta}$ is multivariate normally distributed: $\boldsymbol{\zeta} \sim (0, \boldsymbol{\psi})$. Further, it is assumed that common and unique factors are orthogonal to (i.e., uncorrelated with) one another.

SEM parameters are estimated (Rosseel & Vidal, 2025) by minimizing discrepancies between observed moments ($\mathbf{S}$ and $\bar{\boldsymbol{\gamma}}$) and model-implied moments ($\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\mu}}$). The model-implied variance-covariance matrix:

$$\mathrm{Var}(\boldsymbol{\gamma}) = \boldsymbol{\Sigma}(\boldsymbol{\vartheta}) = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Psi}\left[(\mathbf{I} - \mathbf{B})^{-1}\right]'\boldsymbol{\Lambda}' + \boldsymbol{\Theta}, \tag{3}$$

and model-implied mean vector:

$$\mathrm{E}(\boldsymbol{\gamma}) = \boldsymbol{\mu}(\boldsymbol{\vartheta}) = \boldsymbol{\nu} + \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\alpha}, \tag{4}$$

are functions of SEM parameters, among which the freely estimated parameters are collected in a vector $\boldsymbol{\vartheta}$ for estimation. Importantly, $\mathbf{I}$ is an identity matrix of the same order of $\boldsymbol{\Psi}$ and $\mathbf{B}$, and it is assumed that $(\mathbf{I} - \mathbf{B})$ is invertible.

1.1 Model Fit

The most common method for estimating SEM parameters is maximum likelihood (ML), which works in an iterative fashion where parameter estimates are optimized with the goal of minimizing this discrepancy function:

$$F_{\mathrm{ML}} = \log|\hat{\boldsymbol{\Sigma}}| - \log|S| + \mathrm{trace}(S\hat{\boldsymbol{\Sigma}}^{-1}) - p + (\bar{\boldsymbol{\gamma}} - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\bar{\boldsymbol{\gamma}} - \hat{\boldsymbol{\mu}}). \tag{5}$$

Once a predetermined criterion has been met, the model is said to have converged on the maximum likelihood estimates (MLEs).

A test of exact fit can be constructed $T_{ML} = F_{ML} \times N$. Assuming multivariate normality of $\boldsymbol{\gamma}$, this test statistic is χ^2 distributed with df equal to the number of unique summary statistics used as input data ($p^* = \frac{p(p+1)}{2}$, plus p means if the model includes mean-structure parameters) minus the number of free parameters (i.e., length of the vector $\boldsymbol{\vartheta}$). In addition to an exact test of fit, T_{ML} can be used to construct a host of approximate fit indices (AFIs, e.g., CFI, TLI, or RMSEA), but this study focuses on the test statistic alone because any bias in AFIs can generally be traced to bias of the test statistic used to calculate them.

Hancock and Mueller (2011) illustrated that holding a structural misspecification constant, model misfit appears greater as the magnitude of factor loadings increase (i.e., higher construct reliability). In contrast, as the magnitude of factor loadings decrease, model fit appears to improve when comparing AFI estimates to those proposed by Hu and Bentler (1998). Hancock and Mueller (2011) coined this phenomenon the *reliability paradox*.

1.1.1 Structural Model Fit

To counteract the reliability paradox, Hancock and Mueller (2011) proposed to simply use the model-implied latent covariance matrix from a fitted CFA as input data for a subsequent path analysis to estimate the hypothesized structural parameters. From the fitted path model, one can obtain

a naïve test statistic $\tilde{\chi}^2$ to compute the desired structural analogs of the common AFIs—hereafter referred to as SFIs.

McDonald and Ho (2002) on the other hand, illustrated that because the majority of the global *df* belong to the measurement model, misfit in the measurement model can mask a well fitting structural model or conversely, a well fitting measurement model can evaporate any indication of misfit in the structural model. Therefore, McDonald and Ho (2002) proposed a structural version of the RMSEA (labeled $\text{RMSEA}_{\triangle}$) which is computed using the same information as the traditional method for assessing structural model fit: a nested-model comparison using a likelihood ratio test (LRT, or $\Delta\chi^2$). In a similar vein, Lance et al. (2016) proposed a suite of SFIs based on Anderson and Gerbing (1988, 1992, Conditions 9 & 10), which defined structural fit as being somewhere on a continuum between the best fitting structural model (i.e., a saturated model, represented by a CFA with unrestricted factor covariances) to the worst fitting model (i.e., a null model in which all factor covariances are fixed to 0). Importantly, because $\triangle\chi^2$, $\text{RMSEA}_{\triangle}$ proposed by McDonald and Ho (2002), and the SFIs proposed by Lance et al. (2016) rely on information from models (e.g., CFA [saturated structural model], SEM [hypothesized structural model], and the null structural model [see Lance et al. (2016)]) in which all parameters (measurement and structural) are estimated simultaneously, these approaches are able to maintain nominal Type-I error rates when the structural model is free of misspecifications (Rifenbark, 2019).

McNeish and Hancock (2018) showed that the SFIs proposed by Lance et al. (2016) were still impacted by the reliability paradox and instead recommend the approach proposed by Hancock and Mueller (2011). However, given that this approach requires two stages of estimation (i.e., information from an initial CFA informs the estimation of a subsequent path model), the uncertainty about Stage-1 latent covariances will bias the naïve $\tilde{\chi}^2$ statistic, leading to inflated Type-I error rates. SFIs calculated from it are similarly impacted and depend heavily on the reliability of the measures (Cao et al., 2023; Heene et al., 2024; Rifenbark, 2022; Rifenbark, 2019).

1.2 Possible Solutions

Rifenbark and Jorgensen (2023) conducted a population analysis in which two approaches for improving Type-I error rates were examined: leveraging plausible values (pooled using multiple-imputation methods) and the *structural-after-measurement* framework (SAM; Rosseel and Loh (2024) implemented in the R package `lavaan` (Rosseel, 2012). Results showed that the approach given by Hancock and Mueller (2011) and SAM were nearly identical at the time; however, recently proposed post-hoc corrections (Zhang & Wu, 2024) and new developments in the SAM framework are worthy of further examination (Can & Rosseel, 2025).

1.2.1 Corrections

Zhang and Wu (2024) proposed six corrections that can be applied to the test statistic to improve its performance, naming them V1 through V6. Specifically, two formulations of the asymptotic covariance matrix ($\boldsymbol{\Gamma}$) of estimated SEM parameters were explored: the inverted expected information matrix ($\boldsymbol{W}^{-1}$) [V1, V2, V3] or using a sandwich estimator with the meat being the first-order information matrix ($\boldsymbol{W}^{-1}\boldsymbol{V}\boldsymbol{W}^{-1}$) [V4, V5, V6]. Additionally, Zhang and Wu (2024) investigated different information matrices, as many computational options exist (see Savalei & Rosseel, 2022)—specifically, the observed/expected information matrix from the saturated structural model [V1 and V4], the expected information matrix [V2 and V5], and the observed information matrix [V3 and V6], with the latter two matrices stemming from the hypothesized structural model. We focus on the V2 and V5 corrections, given their better Type-I error control in a simulation study (Zhang & Wu, 2024).

1.2.2 Structural-After-Measurement

The SAM framework proposed by Rosseel and Loh (2024) is implemented in `lavaan` (Rosseel, 2012) with the `sam()` function. SAM is a two-stage approach to SEM in which measurement parameters are estimated in Stage-1 (potentially in separate measurement blocks per factor), to be combined for Stage-2 estimation of structural-model parameters. Two methods are available for Stage-2: *global* and *local*. Global SAM estimates the structural parameters in a full SEM, while holding measurement-model parameters fixed (see Burt, 1976, for an early example). In contrast, local SAM fits a path analysis, using only the summary statistics of common factors—Var(η) and E(η)—as input, which are informed by the estimated measurement parameters and the summary statistics ($\mathbf{S}$ and $\bar{\boldsymbol{y}}$) of indicators (see Wall & Amemiya, 2000).

An important distinction of the SAM approach is that factor covariances do not need to be estimated in Stage-1, as past methods have relied on (Hancock & Mueller, 2011; McNeish & Hancock, 2018; Zhang & Wu, 2024). Instead, SAM calculates factor (co)variances as a function of observed-data covariances ($\boldsymbol{S}$) and Stage-1 measurement parameters. Regardless of local or global SAM, a *mapping matrix* $\boldsymbol{M}$ is defined by rearranging Equation 1 to define common factors as a function of observed data and measurement-model parameters:

$$\begin{aligned} \boldsymbol{\Lambda}\boldsymbol{\eta} &= \boldsymbol{y} - \boldsymbol{\nu} - \boldsymbol{\varepsilon} \\ \boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\eta} &= \boldsymbol{M}\left(\boldsymbol{y} - \boldsymbol{\nu} - \boldsymbol{\varepsilon}\right) \\ \boldsymbol{\eta} &= \boldsymbol{M}\left(\boldsymbol{y} - \boldsymbol{\nu} - \boldsymbol{\varepsilon}\right), \end{aligned} \tag{6}$$

where $\boldsymbol{M}\boldsymbol{\Lambda} = \boldsymbol{I}$ (Rosseel & Loh, 2024, Eq. 11). Then $\boldsymbol{M}$ is used to calculate both latent means:

$$\mathrm{E}(\boldsymbol{\eta}) = \boldsymbol{M}\left(\bar{\boldsymbol{y}} - \boldsymbol{\nu}\right), \tag{7}$$

and latent (co)variances:

$$\mathrm{Var}(\boldsymbol{\eta}) = \boldsymbol{M}\left(\widehat{\boldsymbol{S}} - \boldsymbol{\Theta}\right)\boldsymbol{M}'. \tag{8}$$

Numerous choices are available to calculate $\boldsymbol{M}$, but we use the ML estimator (Rosseel & Loh, 2024, Eq. 16):

$$\boldsymbol{M} = (\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}. \tag{9}$$

As Rosseel and Loh (2024) noted, this mapping matrix corresponds to the Bartlett factor-score matrix, but their SAM approach does not necessitate calculating factor scores.

We focus on local SAM, as it is possible to evaluate the fit of the structural model. A decision to make when employing SAM is the number of measurement blocks (the `sam(mm=)` argument) to be used. There can be as many blocks as there are latent variables, one measurement block that contains all latent variables, or somewhere in between. Using more measurement blocks can prevent some misfit from propagating to other parts of the model (Rosseel & Loh, 2024). Recently, Can and Rosseel (2025) developed new *SE*s for local SAM, along with adjusted $\tilde{\chi}^2$ statistics. It is worth noting that SAM obtains $\boldsymbol{\Gamma}$ in a different manner, making it possible to accommodate multiple measurement blocks.

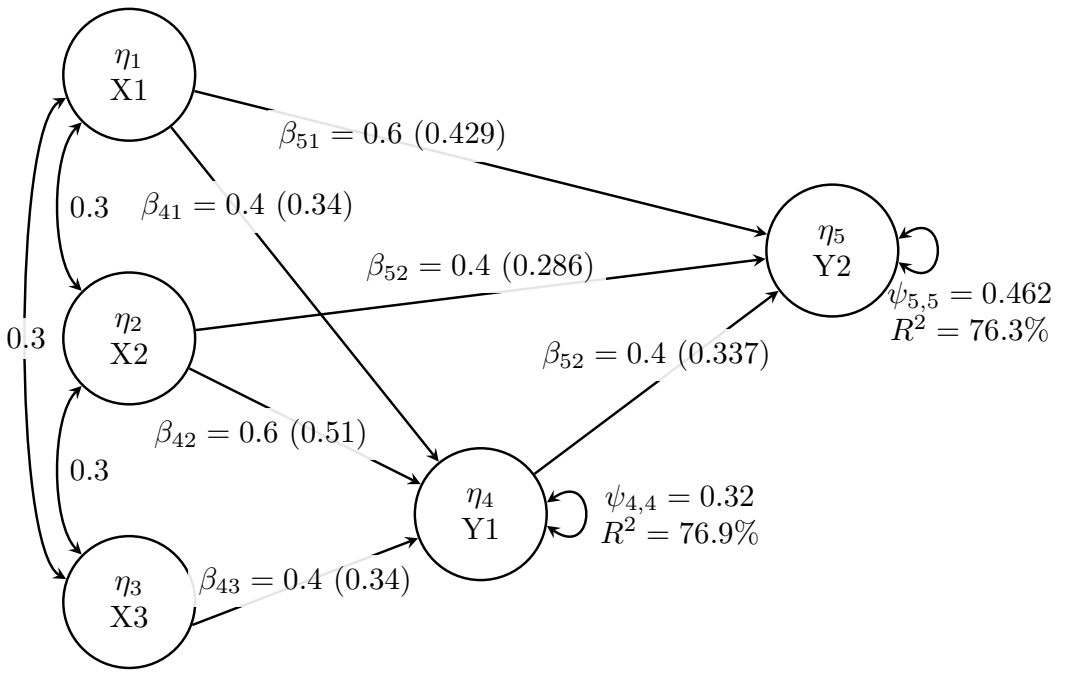
Therefore, our study was designed to examine the performance of the adjusted test statistics from SAM and to compare them to the corrections advanced by Zhang and Wu (2024). The latter fits the Stage-2 path model to estimated factor (co)variances from Stage-1, whereas SAM calculates the Stage-2 input instead from Stage-1 measurement parameters and $\boldsymbol{S}$. We investigated eight tests:

- Single-stage $\Delta\chi^2$
- Two-stage:

- Naïve $\tilde{\chi}^2$ from Stage 2
- Unadjusted $\tilde{\chi}^2$ from SAM (using `mm=1` and `mm=5`)
- Adjusted tests:
  - V2: $\Gamma = W^{-1}$
  - V5: $\Gamma = W^{-1} \times V \times W^{-1}$
  - SAM (using `mm=1` and `mm=5`)

2. Method

We designed a Monte Carlo simulation to examine the impact of three design factors (see Section 2.1) on three study outcomes (see Section 2.4). The population structural model used in the study contained five latent variables and is pictured in Figure 1 which also illustrates the structural relations. Of the five latent variables, three are exogenous (**X1**, **X2**, **X3**) and two are endogenous (**Y1** and **Y2**) with **Y1** serving as a mediator. This structural model has been used in previous research (Lance et al., 2016; MacCallum, 1986; Rifenbark & Jorgensen, 2023).



Note. Values presented for latent covariances correspond to correlations since the latent variances for X1, X2, and X3 are 1.0. Latent regressions present both unstandardized and standardized regression weights, with the latter in parentheses. Finally, for the dependent variables (Y1 and Y2) we provide the population values for the latent disturbances and their respective variance explained by the system of equations.

Figure 1. Population Structural Model

2.1 Manipulated Factors

Next we detail the simulation design factors and their respective levels, while holding other factors constant across conditions.

2.1.1 Population Measurement Model

We generated data from two populations, with the same structural component (Figure 1) but different measurement components. The first model exhibited simple structure: the measurement model contained six indicators-per-factor where $\boldsymbol{\Lambda}$ was block-diagonal and off-diagonal elements of $\boldsymbol{\Theta}$ were equal to zero. The second model had a more complex measurement structure: a single cross-loading ($\lambda_{18,2}$) that was half the magnitude of the primary loading ($\lambda_{18,3}$), and a correlation (ρ = .20) between residuals of two manifest variables (ε_{19} and ε_{25}) that indicate different latent variables (**Y1** and **Y2**).

2.1.2 Construct Reliability

We manipulated measurement quality by varying the magnitude of factor loadings (holding the number of indicators per factor = 6, in contrast to Rifenbark & Jorgensen, 2023). Primary loadings of all indicators per factor are displayed in Table 1, which were chosen to yield low or high CR (specifically, ω = .60 or .90, respectively) in the simple-structure population.

Table 1. Population Factor Loadings

	Latent variable		
Construct Reliability	**X1, X2, X3**	**Y1**	**Y2**
0.60	.448	.380	.320
0.90	.775	.658	.554

2.1.3 Sample Size

We examined three levels of sample size: limited (N = 75), small (N = 150), and large (N = 1000).

2.2 Data Generation

To generate the data for this study we took the following steps.

1. We derived the population latent covariance matrix via $\boldsymbol{\Phi} = (\mathbf{I} - \mathbf{B})^{-1} \times \boldsymbol{\Psi} \times \left[(\mathbf{I} - \mathbf{B})^{-1}\right]'$ using the population values found in Figure 1 above.
2. Next, we solved for population observed covariance matrix via $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$ where elements of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$ differed across simulation conditions:
   - Simple measurement model:
     - The population values for $\boldsymbol{\Lambda}$ across simulation conditions are given in Table 1
     - The population values for the manifest residuals $[diag(\boldsymbol{\Theta})]$ were set to $1 - diag(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}')$
   - Complex measurement model:
     - The population values for $\boldsymbol{\Lambda}$ found in Table 1 remained the same; however, when CR = .60 the primary ($\boldsymbol{\Lambda}_{18,3}$) and secondary ($\boldsymbol{\Lambda}_{18,2}$) loadings were set to .359 (X3) and .179 (X2), respectively, whereas when CR = .90, the primary ($\boldsymbol{\Lambda}_{18,3}$) and secondary loading ($\boldsymbol{\Lambda}_{18,2}$) were set to .622 (X3) and .311 (X2) respectively.
     - The population value for the off-diagonal estimate of $\boldsymbol{\Theta}$ was scaled via $.20 \times \sqrt{\theta_{19,19}\theta_{25,25}}$
3. Finally, we generated multivariate normal data using the `rockchalk` package (Johnson, 2022) in R.

A total of 5000 data sets were generated for each of the 2 (population measurement models: simple vs. complex) × 2 (CR) × 3 (N) = 12 simulation conditions. Next, we detail the procedures undertaken in the simulation.

2.3 Procedure

Below, we provide information related to the fitted models, steps taken to perform the V2/V5 corrections proposed by Zhang and Wu (2024), as well as the correction available in `lavaan::sam()`.

2.3.1 Fitted Models

Prior to describing the procedure for executing the various corrections to the structural goodness-of-fit test statistic, we detail the specific models needed to be fit and their input (i.e., model syntax).

Table 2. Fitted Models Detailed

Model String	$\triangle\chi^2$	$\tilde{\chi}^2_{PATH}$/V2/V5	SAM
CFA	✓	✓	
SEM	✓		✓
PATH		✓	

Note. CFA = confirmatory factor analysis, SEM = full structural equation model (including measurement and structural components), PATH = path analysis (Fig. 1), and SAM = structural-after-measurement.

Specification. Regardless of whether the measurement model was simple or complex in the population, the fitted measurement model always conformed to simple structure in the relevant model strings (e.g., CFA and SEM). Thus, it was a correctly specified model when fitted to data generated from the simple population, but it was incorrectly specified when fitted to data generated from the complex population. The structural model was always correctly specified (i.e., freely estimated all nonzero parameters in Figure 1) in the relevant model strings (e.g., SEM and PATH), but Stage-2 tests could yield inflated Type I errors when the structural model was fitted to a biased estimate of the factor covariance matrix. Table 2 displays the required models to be fit across all methods of interest.

Estimation. All models were fitted in R using `lavaan`[1] (Rosseel, 2012) using *maximum likelihood* (ML) estimation. For each replication, we first fitted the single-stage models (CFA and SEM) which require the simultaneous estimation of all measurement and latent parameters. We elected to employ bounded ML in which lower limits were imposed on manifest residuals and latent variances/disturbances. This was accomplished by specifying: `bounds = "pos.var"` within `lavaan::sem()`. Rifenbark and Rosseel (2024) found that the use of bounded estimation bolstered model convergence rates while having no impact on parameter recovery, even with small sample sizes. For more information on bounded estimation, see De Jonckere and Rosseel (2022). We compared these initial models with a LRT to obtain the standard $\Delta\chi^2$.

2.3.2 Uncorrected Structural Test Statistic

To construct the naïve structural test statistic ($\tilde{\chi}^2$)—referred to as a *pseudo test statistic* (Hancock & Mueller, 2011; Rosseel & Loh, 2024)—it was necessary to extract the model-implied factor covariance matrix from the fitted CFA ($\widehat{\boldsymbol{\Phi}}_{\text{CFA}}$). This was accomplished via `lavaan::lavInspect(cfa.object, what = "cov.lv")`. Using $\widehat{\boldsymbol{\Phi}}_{\text{CFA}}$ as input in a subsequent path analysis, the hypothesized structural model (depicted in Figure 1) was specified and fitted to obtain $\tilde{\chi}^2$.

2.3.3 Zhang and Wu (2024) Corrections

To construct the V2 and V5 adjusted test statistics advanced by Zhang and Wu (2024) the following steps were taken:

1. Extract the W^{-1}:
   `lavaan::lavInspect(cfa.object, what = "inverted.information.expected")`
2. Extract the V:
   `lavaan::lavInspect(cfa.object, what = "information.first.order")`
3. Compute $\boldsymbol{\Gamma}$:
   - V2: W^{-1}
   - V5: $W^{-1} \times V \times W^{-1}$
4. Fit path analysis using $\widehat{\boldsymbol{\Phi}}_{\text{CFA}}$ as input
5. Extract W_2:
   `lavaan::lavInspect(path.object, what = "h1.information.expected")`
6. Extract the Jacobian matrix ($\boldsymbol{\Delta}$):
   `lavaan::lavInspect(path.object, what = "delta")`
7. Solve for the residual weight matrix (U): $W_2 - W_2 \times \Delta \times (\Delta' W_2 \Delta)^{-1} \times \Delta' W_2$
8. Compute the constant:
   - V2: $c = \frac{1}{df} trace(U \times \Gamma_{V2})$
   - V5: $c = \frac{1}{df} trace(U \times \Gamma_{V5})$
9. Scale the unadjusted test statistic via $\frac{1}{c}\chi^2$

2.3.4 Structural-After-Measurement

To carry out the SAM method it is necessary to supply `lavaan::sam()` the SEM model syntax which includes the specification of the measurement model and the structural relations among latent variables. Depending on the number of measurement blocks used, a list object is supplied which specifies which latent variables are to be estimated in a given measurement block. Therefore, when using a single measurement block, a list object with a single element was supplied: `mm.list = list(XY = c("X1",X2","X3","Y1","Y2")`; in contrast, when five measurement blocks were requested, we specified five elements: `mm.list = list(X1="X1",X2="X2",X3 ="X3",Y1="Y1",Y2="Y2")` – see Figure 1.

Importantly, it is necessary to specify the argument `sam.method="local"` to use the SAM[2] method. In order to obtain the adjusted test statistic from `sam()`, the argument `se="local"` must be specified and by default (in the experimental version we used), the uncorrected (naïve) structural-fit test statistic is also available.

2.4 Outcomes

The study contained three outcomes of interest:

1. Convergence rate
2. Expected values of the test statistics, estimated as the average (per condition) across converged replications
3. Type-I error rates, estimated as the proportion of converged replications (per condition) that rejected H_0

The performance of the various test statistics related to their expected values and Type-I error rates depend heavily on convergence rates. Simply, if a required model fails to terminate without error then this will make the construction of the test statistic impossible. Therefore, convergence rates are a key attribute when assessing the performance of available test statistics. Across all simulation conditions, the structural model was fitted with $df = 1$, which is the expected value of the χ^2 fit statistics under H_0, under which condition we expect to observe Type-I error rates near nominal levels (i.e., rejecting H_0 in 5% of converged replications when using $\alpha = .05$).

3. Results

Upon completion of the simulation, we compiled the outcomes of interest across all simulation conditions. Interested readers can access these results and R code used to execute the simulation from our OSF project page.

3.1 Convergence Rates

Table 3. Convergence rates per condition when construct reliability is low

Measurement Model	Simple			Complex		
Sample Size	75	150	1000	75	150	1000
Models Estimated						
CFA	98.9	99.9	100	99.0	100	100
SEM	98.6	99.9	100	98.3	99.9	100
PATH	38.2	69	99.9	36.3	66.2	99.9
SAM (mm=1,mm=5)	100	100	100	100	100	100

Note. CFA = confirmatory factor analysis, SEM = structural equation model, PATH = path analysis, and SAM = structural-after-measurement. To construct the single-stage test ($\Delta\chi^2$), information from CFA and SEM models are required. To construct $\tilde{\chi}^2$ (unadjusted test statistic) and the V2/V5 adjustments per Zhang and Wu (2024), information from CFA and PATH models are required. SAM (unadjusted and adjusted) statistics are solely dependent on the `lavaan::sam()`.

We observed 100% convergence rates for models when CR was high (CR = .90), regardless of whether the measurement model was misspecified or not; therefore, all test statistics and adjustments were available. When CR was low (CR = .60), convergence rates differed across models but remained constant across measurement model conditions for the two-stage approaches. Table 3 shows convergence rates for all models when CR was low.

The test statistics advanced by Zhang and Wu (2024) were found to be unavailable due to low convergence rates for the path analysis model. Specifically, when N = 75 we observed convergence rates between 36% (1817 replications) and 38% (1911 replications) for the path analysis model. When N = 150, convergence rates for the path analysis model nearly doubled to 66% (3310 replications) and 69% (3450 replications). In terms of the SAM test statistics, both adjusted and unadjusted test statistics were available regardless of simulation conditions or the number of measurement blocks used, as 100% convergence rates were observed. Finally, due to the high convergence rates for the CFA and SEM models (98.0% across all simulation conditions) the test statistics stemming from theses models (e.g., $\Delta\chi^2$) were largely available.

3.2 Average Test Statistics

For each approach we only retained replications that converged without error. Within each simulation condition we computed the mean for each test statistic over all available replication to examine the degree to which the average approximated the expected value (df) for the structural model. Table 4 contains averages per condition and method.

3.2.1 Simple Measurement Model

When CR was high, the single-stage and adjusted two-stage test statistics closely approximated df = 1, with the SAM test statistics performing slightly better than $\Delta\chi^2$, V2, and V5 test statistics. The V5 test statistic had slightly higher expected values than the V2 test statistic, and adjusted tests from SAM were slightly larger with one measurement block than with five blocks. Overall, the two-stage adjusted test statistics outperformed $\Delta\chi^2$, except V5 when $N \leq 150$. As expected, average naïve test statistics from two-stage approaches were inflated (2.27–3.15 times > df).

Table 4. Ratio of average test statistic over degrees of freedom across converged replications per condition

	CR = .60			CR = .90		
	N = 75	N = 150	N = 1000	N = 75	N = 150	N = 1000
Measurement Model	S; C	S; C	S; C	S; C	S; C	S; C
Single-stage						
$\Delta\chi^2$	1.60; 1.63	1.07; 1.14	1.05; 1.30	1.10; 1.14	1.08; 1.02	1.01; 1.03
Two-stage, Naïve						
$\tilde{\chi}^2$	25.21; 24.34	32.14; 35.76	20.93; 32.85	3.15; 3.36	2.63; 2.55	2.28; 2.37
SAM, mm=1	53.30; 53.88	74.81; 83.86	21.33; 34.19	3.15; 3.32	2.63; 2.55	2.28; 2.37
SAM, mm=5	19.98; 21.18	31.51; 35.66	18.03; 25.37	2.67; 2.79	2.49; 2.38	2.27; 2.34
Two-Stage, Adjusted						
V2	0.81; 0.78	0.89; 0.93	1.04; 1.32	1.07; 1.11	1.05; 1.00	1.01; 1.03
V5	0.86; 0.85	0.93; 0.96	1.05; 1.33	1.15; 1.19	1.10; 1.05	1.01; 1.04
SAM, mm=1	1.02; 1.03	1.19; 1.27	1.01; 1.26	1.05; 1.08	1.05; 1.00	1.01; 1.03
SAM, mm=5	0.56; 0.58	0.80; 0.81	0.99; 1.14	0.98; 1.02	1.03; 0.97	1.00; 1.02

Note. CR = construct reliability, S = simple measurement population (i.e., correct fitted measurement model), C = complex measurement population (i.e., misspecified measurement model), $\Delta\chi^2$ = LRT comparing nested CFA and SEM, $\tilde{\chi}^2$ = unadjusted test statistic from path analysis fitted to estimated latent (co)variances, SAM = the structural-after-measurement approach (with and without adjustments to the naïve statistic) with mm = number of measurement blocks, V2 & V5 = adjustments proposed by Zhang and Wu (2024).

Low CR exacerbated the inflation, yielding average naïve test statistics 18.03–53.30 times greater than expected. The $\Delta\chi^2$ statistic deviated much less from expectation: 1.60 when N = 75, 1.05 when N = 150. Two-stage adjusted tests had mixed results with smaller sample sizes. When N = 75, SAM with one measurement block performed the best (1.02) followed by the V2 (0.81) and V5 (0.86) statistics, whereas SAM with five measurement blocks performed the worst (0.56). When N = 150, average adjusted tests from SAM were 1.19 (mm = 1) and 0.80 (mm=5), which were greater deviations than V2 (0.89), V5 (0.93), or $\Delta\chi^2$ (1.07). When N = 1000, SAM seemed to approach expected values (0.99 and 1.01 for mm = 1 and 5 blocks, respectively) slightly better than V2 (1.04) or V5 (1.05).

3.2.2 Complex Measurement Model

Similar patterns were found in complex-population conditions. As with simple structure (when the fitted measurement model was correctly specified), naïve two-stage tests remained higher than expected, even when CR was high (2.34–3.36), but especially when CR was low (21.18–83.86). Generally (with 4 small exceptions), inflation was worse when the measurement model was misspecified (i.e., simple structure assumed for data generated from a complex measurement model) than when the simple structure assumption was met. Again, naïve tests from SAM were less inflated with mm = 5 measurement blocks than when a single measurement block was used. Fitting the structural model to the Stage-1 $\widehat{\mathbf{\Phi}}_{\text{CFA}}$ inflated $\tilde{\chi}^2$, but interestingly, inflation in smaller samples with low CR was more similar to SAM with mm = 5 than mm = 1 measurement blocks.

When CR was low, the V2 and V5 adjusted test statistics seemed to be inflated in large samples (N = 1000) but deflated in smaller samples, as was SAM's adjusted test with mm = 5 measurement blocks. It is difficult to identify a discernible pattern of results when CR was low (we defer to their consequences for Type-I error inflation in the next section), but under high CR the adjusted tests from SAM remained closer to expected values than V2 or V5, especially in smaller samples.

3.3 Type-I Error Rates

Reviewing the inflated averages in Table 4 does not make it immediately apparent how large differences were between conditions or methods. Their practical impact can be seen by comparing estimated Type-I error rates, which are shown in Table 5. As with Table 4, each cell contains two estimates (calculated only from converged models), to compare Type-I errors between whether the fitted measurement model was correctly specified (simple population) or misspecified (complex population).

Table 5. Type-I error rates (using α = .05) across converged replications per condition

	CR = .60			CR = .90		
	N = 75	*N* = 150	*N* = 1000	*N* = 75	*N* = 150	*N* = 1000
Measurement Model	S; C	S; C	S; C	S; C	S; C	S; C
Single-stage						
$\Delta\chi^2$	0.11; 0.12	0.05; 0.06	0.06; 0.08	0.06; 0.07	0.06; 0.05	0.05; 0.06
Two-stage, Naïve						
$\tilde{\chi}^2$	0.64; 0.64	0.67; 0.69	0.63; 0.71	0.24; 0.27	0.22; 0.21	0.20; 0.20
SAM, mm=1	0.80; 0.80	0.75; 0.78	0.63; 0.71	0.24; 0.27	0.22; 0.21	0.20; 0.20
SAM, mm=5	0.57; 0.56	0.61; 0.62	0.62; 0.68	0.22; 0.23	0.21; 0.20	0.20; 0.20
Two-Stage, Adjusted						
V2	0.03; 0.03	0.04; 0.04	0.06; 0.08	0.06; 0.06	0.05; 0.05	0.05; 0.06
V5	0.03; 0.04	0.04; 0.05	0.06; 0.08	0.07; 0.07	0.06; 0.05	0.05; 0.06
SAM, mm=1	0.06; 0.07	0.08; 0.08	0.05; 0.07	0.05; 0.06	0.06; 0.05	0.05; 0.06
SAM, mm=5	0.02; 0.02	0.03; 0.03	0.05; 0.06	0.05; 0.05	0.05; 0.05	0.05; 0.05

Note. CR = construct reliability, S = simple measurement population (i.e., correct fitted measurement model), C = complex measurement population (i.e., misspecified measurement model), $\Delta\chi^2$ = LRT comparing nested CFA and SEM, $\tilde{\chi}^2$ = unadjusted test statistic from path analysis fitted to estimated latent (co)variances, SAM = the structural-after-measurement approach (with and without adjustments to the naïve statistic) with mm = number of measurement blocks, V2 & V5 = adjustments proposed by Zhang and Wu (2024).

The overall impact of our measurement model misspecification on Type-I error rates was relatively small, with the exception of two-stage naïve tests when N = 1000. This is especially apparent considering the negligible differences for single-stage $\Delta\chi^2$, for which the misspecification was present in the CFA and full SEM compared to obtain the LRT.

In general, inflation (or even deflation of some two-stage adjusted tests) was more extreme under low than high CR. As with expected values, some counter-intuitive patterns among two-stage adjusted tests were observed as N increases when CR was low. Robitzsch (2022, Table 1) similarly showed counterintuitive patterns for local and global SAM's (lack of) robustness to measurement-model misspecification. Although this seems to run counter to the claim that separate measurement blocks can somewhat protect against propagation of error (Rosseel & Loh, 2024), SAM's ability to do so depends on having high-quality measures. Similar to our low-CR condition, Robitzsch (2022, Eq. 81) only observed his counter-intuitive patterns by simulating data from populations with very low factor loadings.

For the most part, Type-I error rates were closer to nominal levels in larger samples and with higher CR. Type-I error rates appeared to differ only negligibly from the nominal α = .05 (i.e., within 2 percentage points) for two-stage adjusted statistics across almost all conditions—a few in Table 5 were only as high as 8%. Likewise, single-stage estimation of the full SEM yielded nearly nominal Type I error rates (when compared to a CFA) in most conditions, except under the lowest levels of N and CR.

4. Discussion

With respect to convergence rates, we found the test statistics advanced by Zhang and Wu (2024) suffered due to low convergence rates of the fitted path analysis model—see Table 3. Recall these test statistics require models to be fitted in a serial fashion where a CFA model is fitted first, followed by fitting a subsequent path analysis using $\hat{\boldsymbol{\phi}}$ as its input. Even though bounded ML estimation was used for the CFA, the resulting model-implied $\widehat{\boldsymbol{\Phi}}$ possessed eigenvalues that were close to zero or negative, causing the estimation of the path model to fail as a result of a non-positive definite matrix. On the other hand, the unadjusted and adjusted test statistics were able to be computed in 100% of the replications across all simulation conditions when using the SAM method (regardless of the number of measurement blocks used (e.g., one or five). The difference in convergence rates between the path analysis and SAM is due to the default settings in `lavaan::sam()`. Specifically, in the `local.options` argument it is possible to toggle whether to use the lambda correction or not (`lambda.correction = TRUE/FALSE`) which by default is `TRUE`. When employed, the covariance matrix which is used to estimate the structural parameters is forced to be positive definite. This explains the striking difference in convergence rates between SAM and the path analysis models.

In an ad hoc manner, we examined replications where the path analysis model failed to converge, (a) to examine whether using bounded ML estimation of the path analysis model would improve convergence rates, (b) to confirm whether $\hat{\boldsymbol{\phi}}$ from the CFA contained near-zero or negative eigenvalues, and (c) to test whether setting `lambda.correction` to `FALSE` would also result in model non-convergence. From this exercise, we concluded that (a) using bounded ML estimation of the path analysis model did not result in model convergence, (b) the model-implied $\hat{\boldsymbol{\phi}}$ matrix was ill-conditioned as evidenced by near zero and negative eigenvalues, and (c) we were able to mimic model non-convergence for SAM when the lambda correction was not employed.

Aside from the difference in convergence rates between SAM and the path analysis models, it is worth highlighting another major difference between the SAM method and the method of Zhang and Wu (2024). It is important to note that the general steps outlined in Section 2.3.3 are the same for the SAM method; however, `lavaan::sam()` removes many opportunities for human error to occur due to the execution of a single R function. Besides having to manually fit multiple models (CFA and path analysis models), extreme care must be taken to ensure that the elements from the required matricies stemming from the fitted CFA and path analysis models match. Specifically, the row and column names of the inverted expected information matrix from the CFA (W^{-1}, or $\boldsymbol{\Gamma}$ for the V2 adjustment) and the first-order information matrix ($\boldsymbol{V}$, required to compute $\boldsymbol{\Gamma}$ for the V5 adjustment) must match the row and column names of the expected information matrix (W_2) and the row names of the Jacobian matrix ($\boldsymbol{\Delta}$) from the path analysis model (Stage 2). If the matrix elements are not in agreement, the resulting adjustment will be incorrect.

The adjusted test statistics outperformed the naïve test statistics with respect to recovering expected values and Type-I error rates. The impact of our measurement model misspecification was found to have little overall impact on the adjusted test statistics with respect to expected values and Type-I error rates. We also observed the adjusted test statistics from SAM were often close to the V2 and V5 corrections and at times performed better.

As with any simulation, we caution against generalizing results beyond the conditions we investigated. More modeling conditions should be investigated in future research. New directions include investigating how robust SFIs (and their confidence intervals) are when calculated from the adjusted tests we investigated.

Funding Statement: This research was supported by a grant from the Institute of Educational Sciences (#R324A190170).

Competing Interests: None.

Notes

1 We used the development version of `lavaan` available on `GitHub` (version 0.6-20.2325), which was necessary at the time of this work to access this functionality via the `lavaan::sam()` function. By the time of publication, this functionality is available in the version on CRAN (0.6-21).

2 Structural-after-measurement is a framework; therefore, `lavaan::sam()` can be used to estimate a wide variety of two-stage methods. This is done by supplying various values for the `sam.method` argument. For example, "fsr" will conduct a naïve factor score regression.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological bulletin*, *103*(3), 411.

Anderson, J. C., & Gerbing, D. W. (1992). Assumptions and comparative strengths of the two-step approach: Comment on fornell and yi. *Sociological Methods & Research*, *20*(3), 321–333.

Burt, R. S. (1976). Interpretational confounding of unobserved variables in structural equation models. *Sociological methods & research*, *5*(1), 3–52.

Can, S., & Rosseel, Y. (2025). Evaluating the standard error estimation of the local structural-after-measurement (lsam) approach in structural equation modeling [author accepted manuscript]. *Methodology*.

Cao, C., Man, K., & Ge, Y. (2023). Revisiting the impact of measurement quality on targeted structural model fit indexes in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(3), 458–466.

De Jonckere, J., & Rosseel, Y. (2022). Using bounded estimation to avoid nonconvergence in small sample structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(3), 412–427.

Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement*, *71*(2), 306–324.

Heene, M., Maraun, M., Glushko, N. J., & Pornprasertmanit, S. (2024). The devil is mainly in the nuisance parameters: Performance of structural fit indices under misspecified structural models in sem. *Meta-Psychology*, *8*.

Hu, L.-t., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological methods*, *3*(4), 424.

Johnson, P. E. (2022). Package 'rockchalk'.

Lance, C. E., Beck, S. S., Fan, Y., & Carter, N. T. (2016). A taxonomy of path-related goodness-of-fit indices and recommended criterion values. *Psychological methods*, *21*(3), 388.

MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological bulletin*, *100*(1), 107.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological methods*, 7(1), 64.

McNeish, D., & Hancock, G. R. (2018). The effect of measurement quality on targeted structural model fit indices: A comment on lance, beck, fan, and carter (2016).

Rifenbark, G. G. (2022). Impact of construct reliability on proposed measures of structural fit when detecting group differences: A monte carlo examination. *The Annual Meeting of the Psychometric Society*, 313–328.

Rifenbark, G. G., & Jorgensen, T. D. (2023). The plausibility and feasibility of remedies for evaluating structural fit. In M. Wiberg, D. Molenaar, J. González, J.-S. Kim, & H. Hwang (Eds.), *Quantitative psychology: The 87th annual meeting of the psychometric society, bologna, italy, 2022* (pp. 147–159).

Rifenbark, G. G., & Rosseel, Y. (2024). Small sample estimation of structural equation models: A comparison of alternative estimation approaches. *Paper Presented at The International Meeting of the Psychometric Society, Prague, Czech Republic*.

Rifenbark, G. G. (2019). *Misfit at the intersection of measurement quality and model size: A monte carlo examination of methods for detecting structural model misspecification*. University of Connecticut.

Robitzsch, A. (2022). Comparing the robustness of the structural after measurement (sam) approach to structural equation modeling (sem) against local model misspecifications with alternative estimation approaches. *Stats*, *5*(3), 631–672. https://doi.org/10.3390/stats5030039

Rosseel, Y. (2012). Lavaan: An r package for structural equation modeling. *Journal of statistical software*, *48*, 1–36.

Rosseel, Y., & Loh, W. W. (2024). A structural after measurement approach to structural equation modeling. *Psychological Methods*, *29*(3), 561–588. https://doi.org/10.1037/met0000503

Rosseel, Y., & Vidal, M. (2025). A tutorial for understanding sem using r: Where do all the numbers come from? *British Journal of Mathematical and Statistical Psychology*.

Savalei, V., & Rosseel, Y. (2022). Computational options for standard errors and test statistics with incomplete normal and nonnormal data in sem. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(2), 163–181.

Wall, M. M., & Amemiya, Y. (2000). Estimation for polynomial structural equation models. *Journal of the American statistical association*, *95*(451), 929–940.

Zhang, X., & Wu, H. (2024). Investigating structural model fit evaluation. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*(5), 863–881.