# The Plausibility and Feasibility of Remedies for Evaluating Structural Fit

Graham G. Rifenbark[1] & Terrence D. Jorgensen[2]

University of Connecticut[1],
University of Amsterdam[2]

Correspondence:
249 Glenbrook Road, Unit 3064
Storrs CT 06269, USA
graham.rifenbark@uconn.edu

**Abstract.** Various structural fit indices (SFIs) have been proposed to evaluate the structural component of a structural equation model (SEM). Decomposed SFIs treat estimated latent (co)variances from an unrestricted confirmatory factor analysis (CFA) as input data for a path model, from which standard global fit indices are calculated. Conflated SFIs fit a SEM with both measurement and structural components, comparing its fit to orthogonal and unrestricted CFAs. Sensitivity of conflated SFIs to the same structural misspecification depends on standardized factor loadings, but decomposed SFIs have inflated Type-I error rates when compared to rule-of-thumb cutoffs, due to treating estimates as data. We explored whether 2 alternative approaches avoid either shortcoming by separating the measurement and structural model components while accounting for uncertainty of factor-covariance estimates: (a) plausible values and (b) the Structural-After-Measurement (SAM) approach. We conduct population analyses by varying levels of construct reliability and numbers of indicators per factor, under populations with simple and complex measurement models. Results show SAM is as promising as existing decomposed SFIs. Plausible values provide less accurate estimates, but future research should investigate whether its pooled test statistic has nominal Type I error rates.

**Keywords:** structural equation modeling, construct reliability, plausible values, structural-after-measurement, goodness-of-fit

## 1 Evaluating Structural Fit

A structural equation model (SEM) can include both measurement and structural components. The *measurement model* pertains to the relationship between observed and latent variables (i.e., shared variance among indicators of a common factor, vs. error variance unique to each indicator). The *structural model* represents the theorized causal structure among latent variables. Evaluating how well a hypothesized SEM is substantiated by data can be conducted by (a) a

null-hypothesis ($H_0$) test of exact fit, using the likelihood-ratio test (LRT or $\chi^2$) statistic, or (b) quantifying approximate (mis)fit using at least one global fit index (GFI), such as the root-mean-squared error of approximation (RMSEA) or comparative fit index (CFI; see Hu and Bentler, 1998, for an overview).

When the goal is to test/evaluate the hypothesized structural model, its evaluation is complicated by qualities of the measurement model. Specifically, greater construct reliability (determined by the magnitude of factor loadings and the number of indicators per factor in the measurement model) manifests worse apparent data–model fit (e.g., higher $\chi^2$ or RMSEA, lower CFI). That is, the same structural misspecification is easier to detect when using instruments with larger loadings or more indicators than when using fewer or less reliable indicators. Hancock and Mueller (2011) refer to this as the *reliability paradox*: lower reliability yields better apparent data–model fit, inadvertently motivating researchers to use poor-quality measurement instruments. Two existing methods for assessing structural-model fit are conflated and decomposed approaches.

*Conflated* approaches attempt to examine structural model fit by keeping the SEM intact, estimating both components simultaneously. A single SEM's $\chi^2$ statistic conflates misspecification from both components, so Anderson and Gerbing (1988) proposed evaluating structural-model fit with a LRT by comparing a SEM (with hypothesized structural restrictions) to an unrestricted confirmatory factor analysis (CFA), on the assumption[1] that misspecification can only occur in the measurement component. Exact fit is thus tested with a $\Delta\chi^2_{\Delta df}$ statistic: the difference between the hypothesized SEM's $\chi^2_H$ and the structurally saturated CFA's $\chi^2_S$, with $\Delta df = df_H - df_S$. Approximate structural fit can be evaluated using this $\Delta\chi^2$ statistic (and $\Delta df$) in place of a single SEM's $\chi^2$ statistic (and $df$) when calculating common GFIs, for example:

$$\text{RMSEA}_{(\text{D})} \text{ (or RDR)} = \frac{(\Delta)\chi^2 - (\Delta)df}{(\Delta)df \times N}. \tag{1}$$

When using $\Delta\chi^2_{\Delta df}$, Browne and Du Toit (1992) referred to Equation 1 as the root-deterioration per restriction (RDR), which Savalei et al. (2022) more recently called $\text{RMSEA}_\text{D}$. In the specific context of comparing a CFA to a structurally restricted SEM, McDonald and Ho (2002) called it RMSEA-Path, which is the term we use throughout this chapter.

Incremental fit indices (e.g., CFI) can also be calculated using $\Delta\chi^2_{\Delta df}$ (Savalei et al., 2022), but must also include the $\chi^2_0$ statistic for a structural "null" model—e.g., an independence model with endogenous factors orthogonal to themselves and to exogenous factors—which must be nested in the hypothesized SEM (and CFA). Like Savalei et al. (2022) did with RMSEA, Lance et al. (2016) unified some past definitions by proposing a family of structural fit indices (SFIs) called "C9" that are analogous to incremental GFIs, as well as their complement (C10 = 1− C9) that quantifies badness rather than goodness of fit. For example, a C9 analogous to the normed fit index (NFI; Bentler and Bonett, 1980) is:

[1] The structural component might be misspecified even in a CFA if the number of factors is incorrect (Mulaik and Millsap, 2000).

$$C9 = \frac{\chi_0^2 - \chi_H^2}{\chi_0^2 - \chi_S^2}, \qquad (2) \qquad\qquad C10 = \frac{\chi_S^2 - \chi_H^2}{\chi_0^2 - \chi_S^2}. \qquad (3)$$

One can replace each model's $\chi^2$ in Equation 2 with estimated noncentrality parameter (NCP) $\chi^2 - df$ for a C9 analogous to CFI, or with the ratio $\frac{\chi^2}{df}$ for a C9 analogous to the nonnormed fit index (NNFI; Bentler and Bonett, 1980) or Tucker–Lewis (1973) index (TLI).

Conversely, *decomposed* approaches examine structural model fit by separately estimating the measurement and structural components of a SEM in two steps. First, an unrestricted CFA is fitted and its model-implied latent covariance matrix ($\widehat{\mathbf{\Phi}}$) is extracted. Second, $\widehat{\mathbf{\Phi}}$ is used as input data for subsequent path analysis that models the hypothesized relations among latent variables (i.e., matching the target SEM's structural component). Two-stage estimation attempts to circumvent the reliability paradox by removing the (Stage-1) measurement model's influence on (Stage-2) structural model. Hancock and Mueller (2011) proposed calculating GFIs for the Stage-2 path analysis to serve as SFIs.

### 1.1 Issues with Current Methods

Conflated SFIs have nominal Type-I error rates under correct specification (Lance et al., 2016; Rifenbark, 2019, 2022), but their power to detect structural misspecification is moderated by the magnitude of factor loadings (McNeish and Hancock, 2018). Thus conflated C9/C10 still suffer the reliability paradox: C9 indicates better fit with smaller than larger factor loadings.

Although the decomposed approach appears to disentangle measurement-model misfit from structural misspecifications (Hancock and Mueller, 2011), their SFIs also suffer from inflated Type-I error rates (Rifenbark, 2022; Heene et al., 2021) when rule-of-thumb cutoffs are used (e.g., Hu and Bentler, 1999). Imprecision when estimating $\widehat{\mathbf{\Phi}}$ increases an SFI's sampling variance, which occurs when measuring the factors less reliably (lower factor loadings, fewer indicators). This broadening of an SFI's sampling distribution sends more values past the "critical value" (cutoff), even when a structural model is correctly specified (i.e., due to sampling error alone; Marsh et al., 2004).

Ideally, one would not use fixed cutoffs to judge the quality of a model with SFIs (Groskurth et al., 2021; McNeish and Wolf, 2021); however, while it remains common practice, it is valuable to investigate the practical consequences of doing so. Hancock and Mueller (2011) did not propose a decomposed $H_0$ test of exact fit because treating the Stage-1 $\widehat{\mathbf{\Phi}}$ as observed data would inflate the Type I error rate. Thus, only approximate-fit solutions have been proposed from a decomposed perspective.

### 1.2 Potential Remedies for Evaluating Structural Fit

An ideal method would allow structural misspecifications to be identified independent from measurement-model misfit, but without ignoring the measurement model's imprecision when using $\widehat{\boldsymbol{\Phi}}$ as input data. A true test of exact fit with nominal Type I error rate would also be welcome.

We explore two potential solutions based on factor score regression (FSR; Thurstone, 1935; Thomson, 1934), which uses factor-score estimates (derived from Stage-1 measurement models) as input data for a path analysis. FSR suffers from the same limitation as decomposed SFIs: the input data are estimated (not known) factor scores, whose imprecision is not accounted for in Stage-2 estimation. One solution is numerical, the other is analytical.

**Numerical Solution: Sample Plausible Values** Rather than obtain a single point estimate of subject $i$'s vector of factor scores, we can draw a sample of *plausible values* from their sampling distribution, whose variance reflects their imprecision. It was first proposed for Item Response Theory (IRT; Mislevy et al., 1992; von Davier et al., 2009) and has since been applied in SEM (Asparouhov and Muthén, 2010; Jorgensen et al., 2022). The motivation is similar to sampling multiple imputations of missing values (Rubin, 1987), where the (100%-)missing values are the factor scores. Drawing $m$ samples of plausible values provides $m$ imputed data sets, where M should be large enough to minimize additional Monte Carlo sampling error.

To use plausible values to evaluate a structural model's fit, we first estimate an unrestricted CFA, draw $m$ samples of plausible values, fit the hypothesized structural model (as a path analysis) to each of the $m$ data sets, then use Rubin's (1987) rules to pool parameter estimates across $m$ results. The LRT statistic can also be pooled (Meng and Rubin, 1992) and the pooled statistic can be used to calculate SFIs in Equations 1 and 2. Variability of results across $m$ imputations (i.e., between-imputation variance) captures the uncertainty around $\widehat{\boldsymbol{\Phi}}$ and factor scores estimated from it. Imprecision should therefore be accounted for, resulting in decomposed SFIs that yield more robust inferences about structural fit, including a test of exact fit with approximately nominal Type I error rate.

**Analytical Solution: Use Bias-Correcting Formulas** Croon (2002) developed a bias-correcting method for FSR, which Devlieger et al. (2016) showed outperforms other FSR methods in terms of bias, mean-squared error, and Type I error rates. Devlieger et al. (2019) extended Croon's (2002) correction to construct fit indices (RMSEA, CFI, SRMR) and approximate $\chi^2$ for nested-model tests, validating their method with simulation results. These analytical solutions even outperform SEM when there are fewer observations than indicators.

More recently, Rosseel and Loh (2021) developed *structural-after-measurement* (SAM) which generalizes Croon's correction further to be applicable when analyzing summary statistics ($\widehat{\boldsymbol{\Phi}}$) rather than raw data. Thus, factor-score estimates are no longer required. SAM is implemented in the R package `lavaan` (Rosseel,

2012) via the `sam()` function. As the name implies, measurement parameters are estimated first, potentially in separate independent measurement blocks to prevent misfit from propagating across factors (e.g., cross-loadings, residual correlations between indicators of different factors). There can be as many measurement blocks as there are latent variables or as few as one, and there are equivalent "local" and "global" SAM procedures (Rosseel and Loh, 2021). Only local SAM provides a "pseudo-$\chi^2$ statistic" (and fit indices calculated with it) to evaluate the fit of the structural model, so we focus only on local SAM.

## 2 Asymptotic Investigation

We compared how well SFIs from SAM or plausible values could evaluate structural fit, relative to the flawed decomposed SFIs (Hancock and Mueller, 2011) and to the conflated test (Anderson and Gerbing, 1988) and SFIs (Lance et al., 2016). We analyze population moments at the factor level ($\boldsymbol{\Phi}$) and item level ($\boldsymbol{\Sigma}$) to obtain asymptotic results free from sampling error. Factor-level results enable us to determine "true" values (benchmarks for SFIs) of an overly restricted structural model. Item-level results enable evaluating how much each method's SFIs are affected by different measurement-model conditions.

### 2.1 Hypotheses

We know from past research (McNeish and Hancock, 2018) that for a given structural misspecification, SFIs of Lance et al. (2016) indicate better (or worse) fit with lower (or higher) factor loadings and fewer (or more) indicators; conversely, Hancock and Mueller (2011) SFIs are not affected (on average) by measurement quality. However, measurement-model misspecifications (e.g., omitted cross-loadings) should bias estimates of factor (co)variances, thus biasing even Hancock and Mueller (2011) SFIs.

Regardless of whether a measurement model is correctly specified, we expect plausible values to yield asymptotically identical SFIs as the decomposed SFIs of Hancock and Mueller (2011) regardless of measurement quality. Plausible values and decomposed SFIs both estimate $\widehat{\boldsymbol{\Phi}}$ from a CFA, which will not be biased by poor measurement quality, but can be biased by measurement misspecifications (e.g., omitted cross-loadings). The advantage of plausible values is that beyond SFIs, a pooled $\chi^2$ statistic can be calculated, which should be similar to the $\chi^2$ obtained by fitting the same model to the population $\boldsymbol{\Phi}$.

Likewise, we expect SAM to yield asymptotically identical SFIs as the decomposed SFIs of Hancock and Mueller (2011) regardless of measurement quality, but only when a measurement model is correctly specified. Given measurement misspecifications (e.g., omitted cross-loadings), SAM's independent measurement blocks provide a layer of protection from propagated errors, which should make SAM's SFIs more robust than plausible values or Hancock and Mueller (2011) SFIs.

### 2.2 Factor-Level Population Model

First, we specified population parameters to derive $\mathbf{\Phi}$, which enabled us to determine population-level SFI values for more-restricted models. We refer to these true-value results to evaluate the accuracy of SFI estimates under four different methods in the indicator level analysis. We selected a frequently used structural model for our population (Lance et al., 2016; McNeish and Hancock, 2018; Rifenbark, 2019, 2022), depicted in Figure 1. These population parameters imply population covariance matrix $\mathbf{\Phi} = (\mathbf{I} - \mathbf{B})^{-1} \times \mathbf{\Psi} \times \left[(\mathbf{I} - \mathbf{B})^{-1}\right]'$, to which we fit four models:

- saturated Model $S$: all variables freely covary
- null Model 0: only X1, X2, and X3 freely covary
- true partial-mediation Model $T$: all paths in Figure 1 estimated
- misspecified full-mediation Model $M$: Model $T$ with fixed $\beta_{51} = \beta_{52} = 0$

Models were estimated with maximum likelihood (ML) in `lavaan()`, and the `fitMeasures()` function was used to obtain $\chi^2$ (with $N = 500$) and GFIs for Model $M$. We used Models $M$, $S$, and 0 to calculate C9 (Eq. 2), and we verified that Model $T$ estimates matched population parameters in Figure 1.

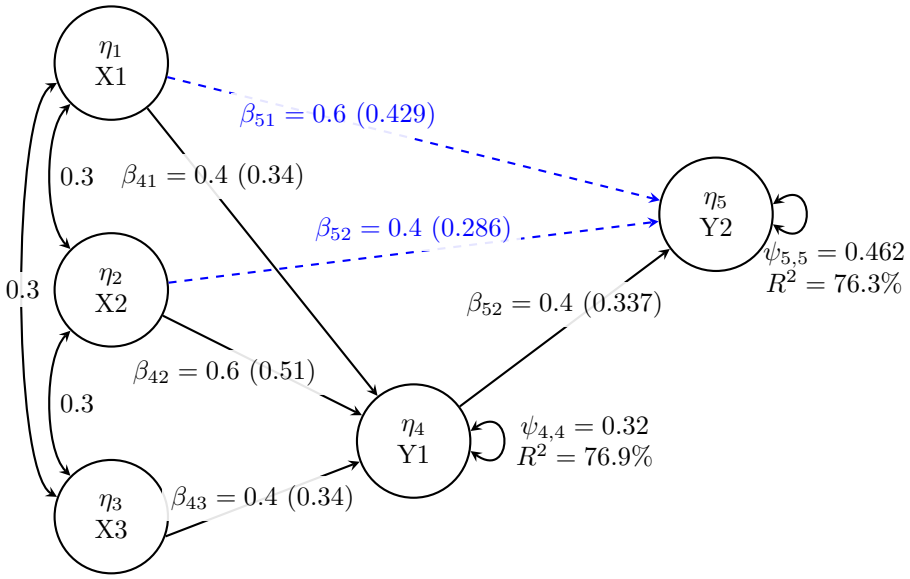


**Fig. 1.** Population structural parameters. Each exogenous-factor variance $\psi_{X,X} = 1$, so exogenous covariances are correlations. Standardized slopes in parentheses.

Model $M$'s $\chi^2_{df=3} = 216.99$, so population RMSEA = 0.378 indicated very poor fit. Population CFI = .863 and analogous C9 = .852 were also unacceptable

by most standards (Bentler and Bonett, 1980; Hu and Bentler, 1999). These "true values" are the benchmarks we will use to compare the four methods for evaluating structural fit using indicator-level data.

### 2.3 Indicator-Level Population Model

Holding the structural model constant, we specified different measurement models to investigate the impact of different measurement-model attributes on structural model evaluation. We manipulated three factors:

- We used 3 or 6 indicators per factor (pF). Therefore, the full SEM (Lance et al., 2016) or CFA (plausible values Hancock and Mueller, 2011) was fitted to 15 or 30 indicators. Local SAM's fitted 5 single-indicator CFAs to each factor's 3 or 6 indicators before fitting the structural component (Model $M$).
- Whereas McNeish and Hancock (2018) manipulated factor loadings directly, we selected loadings that would yield low or high construct reliability (CR = 0.6 or 0.9), which also depends on pF (Gagne and Hancock, 2006). As such, for a given construct reliability, factor loadings were lower when pF = 6 than when pF = 3. Table 1 shows the population $\mathbf{\Lambda}$ values (of all pF indicators) for each factor under various conditions. They are standardized loadings, such that residual variances were set to $\text{diag}(\mathbf{\Theta}) = 1 - \text{diag}(\mathbf{\Lambda\Phi\Lambda}')$.
- In the population, the measurement model had either simple or complex structure. Simple structure implies each observed indicator loads onto only one latent variable, and residuals are uncorrelated. Our complex measurement model contained both a cross-loading and a correlated residual. In the complex population, the covariance between the first indicators of Y1 and Y2 was $r = .20$ (scaled to a covariance by multiplying residual $SD$s: $.2\sqrt{\theta_{y_1}\theta_{y_7}}$), and the last indicator of X3 cross-loaded onto X2. Table 1 shows that across pF and CR conditions, the cross-loading (in parentheses) was half as large as the primary loading, while maintaining indicator variances $\theta_{x,x} = 1$.

**Table 1.** Population values for $\mathbf{\Lambda}$.

| | pF = 3 | | pF = 6 | |
|---|---|---|---|---|
| | CR = .90 | CR = .60 | CR = .90 | CR = .60 |
| X1–X3 | .866 | .578 | .775 | .448 |
| PL (CL) | .696 (.348) | .464 (.232) | .622 (.311) | .359 (.179) |
| Y1 | .736 | .491 | .658 | .380 |
| Y2 | .620 | .413 | .554 | .320 |

*Note.* Simple-structure parameters given in the top row. Second row shows PL = primary loading and CL = cross-loading of indicators of X1–X3 in complex-structure conditions. Bottom rows show loadings for Y1 and Y2 under either simple or complex structure.

In all six conditions, we computed the population indicator-level covariance matrix implied by our SEM parameters in Figure 1 and Table 1: $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$.

### 2.4 Procedure

The same four structural models that we fitted to $\boldsymbol{\Phi}$ were augmented with a simple-structure model. Thus, augmented Model $S$ was an unrestricted CFA, augmented Model 0 was an orthogonal CFA, and augmented Models $T$ and $M$ were "full" SEMs representing partial and full mediation, respectively. In simple-structure conditions, the measurement model was correctly specified, but it was misspecified in complex-structure conditions because it omitted the cross-loading and residual covariance. Misspecifying the measurement model (which biases $\widehat{\boldsymbol{\Phi}}$) allowed us to compare how SFIs are influenced across the four methods.

The four full SEMs were fitted to the indicator-level population $\boldsymbol{\Sigma}$, and resulting $\chi^2$ values were used to calculate conflated SFIs for augmented Model $M$: RMSEA-Path (Eq. 1; McDonald and Ho, 2002) and C9 with NCP (Eq. 2, analogous to CFI; Lance et al., 2016). To calculate decomposed versions of these SFIs (Hancock and Mueller, 2011), we saved the model-implied $\widehat{\boldsymbol{\Phi}}$ and fitted the (nonaugmented) Model $M$ to it, just as we did to obtain "true" population SFIs by fitting Model $M$ to the population $\boldsymbol{\Phi}$. However, $\widehat{\boldsymbol{\Phi}}$ could vary across the 2 (simple vs. complex) $\times$ 2 (pf = 3 or 6) $\times$ 2 (CR = .60 or .90) = 8 conditions.

To obtain SFIs using plausible values and SAM, raw data were necessary for analysis. We used the `rockchalk::mvrnorm()` function to generate a single data set with the argument `empirical=TRUE` to guarantee our sample's covariance matrix was identical to the population $\boldsymbol{\Sigma}$. This minimized sampling error, although some Monte Carlo error was still expected because different raw data (even with identical covariance matrices) yield different factor-score estimates.

**Plausible values.** We fitted an unrestricted CFA (augmented Model $S$) to the raw data, then used the `semTools::plausibleValues()` function (Jorgensen et al., 2022) to sample $m = 100$ sets of plausible values. We used the `semTools::sem.mi()` function to fit Model $M$ to each sample of plausible values. The `fitMeasures()` function provided SFIs using the pooled $\chi^2$ statistic (the "D3" method; Meng and Rubin, 1992).

**SAM.** We used the `lavaan::sam()` function to fit augmented Model $M$ to the raw data, which internally fitted five single-factor CFAs (i.e., 5 measurement blocks using the argument `mm=5`), followed by fitting Model $M$ to the $\widehat{\boldsymbol{\Phi}}$ estimate obtained via the local-SAM method (Rosseel and Loh, 2021). SFIs are printed by the `summary()` function.

### 2.5 Results & Discussion

We verified that all GFIs, SFIs, and $\chi^2$ showed perfect data–model fit when both the measurement and structural (Model $T$) components were correctly

specified. Table 2 presents estimated SFIs (RMSEA and CFI) for Model $M$ across conditions, with their true values from Section 2.2 in the column headers.

**Table 2.** Asymptotic Estimates of SFIs across conditions.

| | | | RMSEA (= 0.378) | | | | CFI (= .863) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Measurement | CR | pF | $\widehat{\boldsymbol{\Sigma}}$ | $\widehat{\boldsymbol{\Phi}}$ | PV | SAM | $\widehat{\boldsymbol{\Sigma}}$ | $\widehat{\boldsymbol{\Phi}}$ | PV | SAM |
| Simple | low | 3 | 0.088 | 0.378 | 0.285 | 0.378 | .971 | .863 | .848 | .863 |
| (correctly | | 6 | 0.090 | 0.378 | 0.291 | 0.378 | .970 | .863 | .842 | .863 |
| specified) | high | 3 | 0.255 | 0.378 | 0.320 | 0.378 | .906 | .863 | .850 | .863 |
| | | 6 | 0.257 | 0.378 | 0.323 | 0.378 | .905 | .863 | .848 | .863 |
| Complex | low | 3 | 0.080 | 0.356 | 0.254 | 0.358 | .977 | .890 | .888 | .887 |
| (misspecified) | | 6 | 0.085 | 0.364 | 0.261 | 0.366 | .973 | .878 | .877 | .877 |
| | high | 3 | 0.251 | 0.373 | 0.309 | 0.373 | .910 | .872 | .865 | .871 |
| | | 6 | 0.254 | 0.375 | 0.310 | 0.375 | .907 | .867 | .862 | .867 |

*Note.* True RMSEA and CFI provided in column headers as benchmarks. CR = high (.9) or low (.6) construct reliability. pF = number of indicators per factor. $\widehat{\boldsymbol{\Sigma}}$ = conflated SFIs (i.e., RMSEA-Path or C9). $\widehat{\boldsymbol{\Phi}}$ = decomposed SFIs of Hancock and Mueller (2011). PV = decomposed SFIs pooled from plausible values. SAM = decomposed SFIs from pseudo-$\chi^2$ of SAM approach.

**Conflated SFIs.** As expected (McNeish and Hancock, 2018; Rifenbark, 2019, 2022), RMSEA-Path (McDonald and Ho, 2002) and C9 (Lance et al., 2016) in the $\widehat{\boldsymbol{\Sigma}}$ column of Table 2 were affected by measurement quality (CR), with lower CR inducing better apparent fit. One might not even reject the model using SFIs when construct reliability was low. Even the additional misfit from the measurement model (complex populations) did not yield SFIs that indicated fit being as poor as the true values did, although the impact of measurement misspecification was small. Holding CR constant, number of indicators (pF) also did not substantially affect expected values of RMSEA-Path or C9.

**Decomposed SFIs.** When the measurement model was correctly specified, Hancock and Mueller (2011) SFIs (in the $\widehat{\boldsymbol{\Phi}}$ column of Table 2) nearly matched SAM's results across all CR and pF conditions, indicating their SFIs have asymptotically equivalent expected values. Both methods estimated true SFIs accurately for simple-structure populations. But their equivalence did not hold for misspecified measurement models. Failing to model the cross-loading and residual correlation induced small differences between SAM and Hancock and Mueller (2011) SFIs, with SAM estimates being slightly closer to true values. Although the impact of pF was small (somewhat better fit with fewer indicators), its effect was greater when CR was low.

Using plausible values also showed some promise, although its pooled SFIs were less accurate estimates of true values than SAM or Hancock and Mueller (2011). Pooled RMSEA showed better fit than the true values (particularly with low CR), and pooled CFI estimates were somewhat more accurate than RMSEA. However, pooled CFI showed better fit than true values (like RMSEA) only when the measurement model was misspecified; with correct specification, pooled CFI always showed worse fit than true values across conditions. Pooled SFIs always showed slightly worse fit with more indicators, but again this was negligible.

## 3 Conclusion

Population analyses show that SAM and the decomposed SFIs of Hancock and Mueller (2011) are identical in the case of the simple measurement model. However, slight differences were observed when the complex measurement model was misspecified. This was expected because SAM isolates local misfit in each measurement block, which may enable SAM to outperform Hancock and Mueller (2011) in cases of greater measurement misspecification.

In the current investigation, Hancock and Mueller (2011) SFIs appear asymptotically equivalent to SAM's SFIs. Although their sampling distributions may have the same expected values, their sampling variances may yet differ. Caution is warranted until Monte Carlo studies reveal whether increasing either's sampling variability inflates Type I errors (i.e., in smaller samples and lower CR). Holding CR constant, pF had negligible impact on SFIs, which warrants ignoring it in future Monte Carlo study, varying only CR via the magnitude of factor loadings.

The oddly inconsistent plausible-value results are likely due to the relative misfit of Model 0 and Model $M$, but could also be due to Monte Carlo sampling error (we drew a finite sample of plausible values, so these results were not entirely asymptotic). Results could also depend on the method for pooling the $\chi^2$ statistic; alternatives include the "D2" method (Li et al., 1991) and "D4" (Chan and Meng, 2017). Grund et al. (2021) found that D2 can be too liberal, while D3 and D4 can be too conservative. Given how these patterns could be exacerbated in the extremely poor-fitting null Model 0, further investigation is warranted. The greatest promise of plausible values may not be for SFIs themselves, but in its ability to provide an actual (pooled) *test* of the $H_0$ of exact fit.

# Bibliography


Anderson, J. C. and Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103(3):411–423.

Asparouhov, T. and Muthén, B. (2010). Plausible values for latent variables using M*plus*. Available from http://www.statmodel.com/download/Plausible.pdf.

Bentler, P. M. and Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88(3):588–606.

Browne, M. W. and Du Toit, S. H. (1992). Automated fitting of nonstandard models. *Multivariate Behavioral Research*, 27(2):269–300.

Chan, K. W. and Meng, X.-L. (2017). Multiple improvements of multiple imputation likelihood ratio tests. *Statistica Sinica*, 32:1489–1514.

Croon, M. (2002). Using predicted latent scores in general latent structure models. In Marcoulides, G. and Moustaki, I., editors, *Latent variable and latent structure models*, pages 195–223. Erlbaum, Mahwah, NJ.

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

Devlieger, I., Talloen, W., and Rosseel, Y. (2019). New developments in factor score regression: Fit indices and a model comparison test. *Educational and Psychological Measurement*, 79(6):1017–1037.

Gagne, P. and Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, 41(1):65–83.

Groskurth, K., Bluemke, M., and Lechner, C. (2021). Why we need to abandon fixed cutoffs for goodness-of-fit indices: A comprehensive simulation and possible solutions. Available from PsyArXiv: https://doi.org/10.31234/osf.io/5qag3.

Grund, S., Lüdtke, O., and Robitzsch, A. (2021). Pooling methods for likelihood ratio tests in multiply imputed data sets. Available at PsyArXiv: https://doi.org/10.31234/osf.io/d459g.

Hancock, G. R. and 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. D., Glushko, N. J., and Pornprasertmanit, S. (2021). The devil is mainly in the nuisance parameters: Performance of structural fit indices under misspecified structural models in SEM. Available on PsyArXiv: https://doi.org/10.31234/osf.io/d8tuy.

Hu, L.-t. and Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3(4):424–453.

Hu, L.-t. and 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.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., and Rosseel, Y. (2022). *semTools: Useful tools for structural equation modeling*. R package version 0.5-6.

Lance, C. E., Beck, S. S., Fan, Y., and Carter, N. T. (2016). A taxonomy of path-related goodness-of-fit indices and recommended criterion values. *Psychological Methods*, 21(3):388–404.

Li, K.-H., Meng, X.-L., Raghunathan, T. E., and Rubin, D. B. (1991). Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica*, 1(1):65–92. Retrieved from https://www.jstor.org/stable/24303994.

Marsh, H. W., Hau, K.-T., and Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, 11(3):320–341.

McDonald, R. P. and Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, 7(1):64–82.

McNeish, D. and Hancock, G. R. (2018). The effect of measurement quality on targeted structural model fit indices: A comment on lance, beck, fan, and carter (2016). *Psychological Methods*, 23(1):184–190.

McNeish, D. and Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*. Advanced online publication.

Meng, X.-L. and Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, 79(1):103–111.

Mislevy, R. J., Johnson, E. G., and Muraki, E. (1992). Chapter 3: Scaling procedures in NAEP. *Journal of Educational Statistics*, 17(2):131–154.

Mulaik, S. A. and Millsap, R. E. (2000). Doing the four-step right. *Structural Equation Modeling*, 7(1):36–73.

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*. PhD thesis, University of Connecticut.

Rifenbark, G. G. (2022). Impact of construct reliability on proposed measures of structural fit when detecting group differences: A Monte Carlo examination. In Wiberg, M., Molenaar, D., González, J., Kim, J.-S., and Hwang, H., editors, *Quantitative Psychology: The 86th Annual Meeting of the Psychometric Society, Virtual, 2021*, pages 313–328. Springer.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.

Rosseel, Y. and Loh, W. W. (2021). A structural after measurement (SAM) approach to SEM. *Psychological Methods*. Preprint retrieved from https://osf.io/pekbm.

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

Savalei, V., Brace, J., and Fouladi, R. T. (2022). We need to change how we compute RMSEA for nested model comparisons in structural equation modeling. *Psychological Methods*. Preprint available from PsyArXiv: https://doi.org/10.31234/osf.io/wprg8.

Thomson, G. H. (1934). The meaning of "i" in the estimate of "g". *British Journal of Psychology. General Section*, 25(1):92–99.

Thurstone, L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of primary traits*. University of Chicago Press.

Tucker, L. R. and Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1):1–10.

von Davier, M., Gonzalez, E., and Mislevy, R. (2009). What are plausible values and why are they useful. In von Davier, M. and Hastedt, D., editors, *IERI monograph series: Issues and methodologies in large-scale assessments*, pages 9–36. IEA-ETS Research Institute, Hamburg, Germany.