Dynamic Fit Index Cutoffs for Confirmatory Factor Analysis Models

Daniel McNeish[1] & Melissa G. Wolf[2]

[1]Arizona State University, USA
[2]University of California, Santa Barbara, USA

ACCEPTED AT PSYCHOLOGICAL METHODS,
JULY 12, 2021

Contact Information:

Daniel McNeish, PO Box 871104, Arizona State University, Tempe, AZ 85287.
Email: dmcneish@asu.edu

Abstract

Model fit assessment is a central component of evaluating confirmatory factor analysis models and often the validity of psychological assessments. Fit indices remain popular and researchers often judge fit with fixed cutoffs derived by Hu and Bentler (1999). Despite their overwhelming popularity, methodological studies have cautioned against fixed cutoffs, noting that the meaning of fit indices varies based on a complex interaction of model characteristics like factor reliability, number of items, and number of factors. Criticism of fixed cutoffs stems primarily from the fact that they were derived from one specific confirmatory factor analysis model and lack generalizability. To address this, we propose a simulation-based method called *dynamic fit index* cutoffs such that derivation of cutoffs is adaptively tailored to the specific model and data characteristics being evaluated. Unlike previously proposed simulation-based techniques, our method removes existing barriers to implementation by providing an open-source, web-based Shiny software application that automates the entire process so that users neither need to manually write any software code nor be knowledgeable about foundations of Monte Carlo simulation. Additionally, we extend fit index cutoff derivations to include sets of cutoffs for multiple levels of misspecification. In doing so, fit indices can more closely resemble their originally intended purpose as effect sizes quantifying misfit rather than improperly functioning as ad hoc hypothesis tests. We also provide an approach specifically designed for the nuances of one-factor models, which have received surprisingly little attention in the literature despite frequent substantive interests in unidimensionality.

Dynamic Fit Index Cutoffs for Confirmatory Factor Analysis Models

Model fit assessment is an integral part of applying confirmatory factor analysis (CFA) to empirical data. For better or worse, metrics of model fit are often the most influential pieces of information for consumers or producers of research using CFA. Global model fit metrics that summarize the fit of the overall model with a single value generally fall into two broad categorizations: exact and approximate.

The maximum likelihood test statistic (usually referred to as the χ^2 test, though other types of tests are possible; Yuan & Bentler, 1999) is a common test of exact global fit and assesses the hypothesis that the model-implied covariance matrix and (and possibly the model-implied mean vector as well) is equal to the observed covariance matrix (and possibly the observed mean vector as well). Though valued for its clear definition and inferential nature, empirical researchers lament two issues with its use: (a) exact fit is not always a necessary condition for a model to be useful (Browne & Cudeck, 1992; Cudeck & Henly, 1991; MacCaullum, Widaman, Preacher, & Hong, 2001; Mulaik, 2009; Meehl, 1978) and (b) the power of the test rapidly increases with sample size or as data deviate from multivariate normality (Bentler & Yuan, 1999; Hu, Bentler, & Kano, 1992; Tanaka, 1987).

Though these positions are endorsed by many empirical researchers, the rationale sometimes is attributable to the difficulty of achieving exact fit rather than merit of the test itself (e.g., Barrett, 2007; McIntosh, 2012, Ropovik, 2015). Approximate fit indices (e.g., RMSEA, CFI, SRMR) are also popular and originally were intended to function more as effect sizes to supplement exact fit tests by capturing the magnitude of misspecification in the model. However, like effect sizes, approximate fit indices lack inherent null hypotheses, meaning there is ambiguity in which values signal “good” or “bad” fit.

Early in the development of fit indices, values constituting approximate good fit were not easy to discern and were often based on unsubstantiated heuristics or personal experience (Marsh et al., 2004; Maydeu-Olivares, 2017; West et al., 2012). In the late 1990s, more formal inquiry to ascribe qualitative meaning to fit indices emerged (e.g., MacCallum et al., 1996), culminating in the seminal study by Hu and Bentler (1999). Hu and Bentler (1999) conducted an extensive simulation study to investigate which values of various fit indices were able to consistently distinguish between fit index distributions of a CFA model that did and did not contain misspecifications. This work led to the pervasive traditional cutoffs used in empirical studies such as SRMR ≤ .08, RMSEA ≤ .06, and CFI ≥ .96.

At the time of this writing, Hu and Bentler (1999) has received over 80,000 citations on Google Scholar and remains a widespread resource for empirical researchers when presenting evidence for approximate model fit. As a testament to its popularity, a review by Jackson et al. (2009) noted that about 60% of the approximately 350 studies they reviewed from APA journals explicitly mention such fit index cutoffs when evaluating the fit of a CFA model. Though the traditional cutoffs from Hu and Bentler (1999) have reached near canonical status, the authors themselves warned that cutoffs are not rigid and should not be overgeneralized as simulation studies are only applicable to the conditions they investigate (Hu & Bentler, 1998, p. 446).

In the intervening 20 years, cautions against fit index cutoff overgeneralization have intensified and methodological studies have noted that cutoff values can change depending on data and model characteristics such as the number of items or factors (Jackson, 2007; Kenny & McCoach, 2003; Moshagen, 2012; Shi et al., 2019); degrees of freedom (F. Chen et al., 2008; Kenny et al., 2015); magnitude of the standardized loadings and factor reliability (Browne et al., 2002; Hancock & Mueller, 2011; Heene et al., 2011; McNeish et al., 2018); and model type and the nature of the misspecification (Fan & Sivo, 2007; Kang et al., 2016; Sivo et al., 2006). Despite these findings, the review by Jackson et al. (2009) noted, “We also did not find evidence that warnings about strict adherence to Hu and Bentler’s suggestions were being heeded” (p. 18).

Based on Google Scholar citation counts, annual citations of Hu and Bentler (1999) have increased from 2,171 in 2009 when Jackson et al. (2009) was published to 10,333 in 2020, suggesting that this trend has not dissipated in the intervening years but rather has intensified. Although deriving fit index cutoffs is considered a precarious objective to some methodologists, use of traditional fixed cutoffs endures because cutoffs have pragmatic utility in empirical studies, just as effect sizes like Cohen’s *d* have become entrenched in linear regression and analysis of variance models for assessing practical utility of treatment effects.

In this paper, we extend the recent methodological literature on simulation-based approaches to improve generalizability of cutoffs by adaptively updating the model subspace of interest to essentially allow researchers to ask, “What cutoffs would Hu and Bentler (1999) have derived had they used a model like mine in their simulation instead?”. In doing so, researchers no longer need to tenuously generalize fixed cutoffs derived from a single simulation whose conditions only represent a narrow (and possibly disparate) model subspace. Instead, this method allows researchers to derive fit index cutoff values that appropriately quantify the magnitude of misspecification in CFA models for any combination of sample size, number of items, number of factors, factor reliability, etc. Though generalizing fit index cutoffs will not address all concerns from ardent critics of fit index use, empirical studies appear persistent in evaluating fit by comparing fit indices to cutoffs, so our goal is to improve practice by supplying better fit cutoffs within the existing framework and to provide researchers with software that removes barriers to implementation.

To outline the remainder of the manuscript, we first provide some intuition about why it is problematic to generalize the traditional fixed cutoffs. Then, we replicate the simulation from Hu and Bentler (1999) but manipulate factor reliability, a model characteristic that was not manipulated in the original simulation study but has since been shown to affect cutoff values. This replication serves to demonstrate to potentially uninitiated readers how the fit index cutoffs derived by Hu and Bentler (1999) would have varied if the model subspace explored in their simulation were altered even by a single aspect. Next, we review previously proposed simulation-based techniques for deriving custom fit index cutoffs and discuss how we propose to extend this literature with an algorithm that recreates the misspecification from Hu and Bentler (1999) for any CFA model of interest. We apply the approach to a single simulated dataset with detailed descriptions of each step to show how cutoffs are derived and compare the cutoffs to what would be derived using the true model. We follow with a generalization of the algorithm to induce a range of misspecifications so that researchers can obtain a continuum of cutoffs corresponding to different degrees of misspecification to treat fit indices more like the effect sizes that they were intended to be rather than as the de facto hypothesis tests to which they have devolved. We then apply the method to results from a study in the empirical literature to show how conclusions drawn from fixed cutoffs can change when cutoffs are customized for the model being evaluated. Issues specific to one-factor models that are common in scale validation are also discussed because the approach in Hu and Bentler (1999) does not necessarily correspond to such models and a separate approach is proposed instead.

To facilitate implementation, we provide a web-based Shiny application that implements our procedure to generate dynamic cutoffs for the user's empirical model (www.dynamicfit.app). This application relies on two R packages to perform the necessary Monte Carlo simulations to derive custom cutoff value (lavaan and simstandard; Rosseel, 2012; Schneider, 2019) but requires no knowledge of Monte Carlo simulations techniques or any specific software programs and does not require users to manually write any software code. Instead, the application only requires users to upload a .txt file containing the model they wish to evaluate along with their sample size and the application automates the remainder of the process to write code for a Monte Carlo simulation, execute the Monte Carlo study, and collate the results to report cutoffs applicable to the specific model of interest.[1] Appendix A shows a tutorial on how to use the application and Appendix B shows how the method can be implemented (but not automated) within M*plus*. We end with a discussion recounting the difficulties of model fit assessment generally, ways to generalize the method beyond CFA (e.g., growth models, measurement invariance), the limitations we foresee with our proposed method despite potential improvement to the status quo, and implications for evidence of validity for psychological assessments.

The Conceptual Problem

To motivate the issue, consider the entire space of all possible CFA models that one could encounter, which includes any combination of model and data characteristics like sample size, number of items, number of factors, degrees of freedom, or factor reliability (a function of the standardized loadings and the number of items, possibly captured by Coefficient *H*; Hancock & Mueller, 2001). As methodological studies have shown, depending on where the empirical model of interest is located in this space, model misspecification affects fit indices differently and the fit index value that is sensitive to a particular misspecification fluctuates.

[1] The R script used by the application is openly available from the application itself.

To help visualize this process, consider the hypothetical representation of RMSEA cutoffs across the possible model space in Figure 1.[2] The RMSEA value associated with a specific model misspecification is on the z-axis, total number of items is on the y-axis, and factor reliability is on the x-axis. Though other characteristics are relevant, we limit the representation to 3 dimensions so that it can be easily visualized. Throughout this model space, the sensitivity of RMSEA to misspecification fluctuates as a function of model and data characteristics.

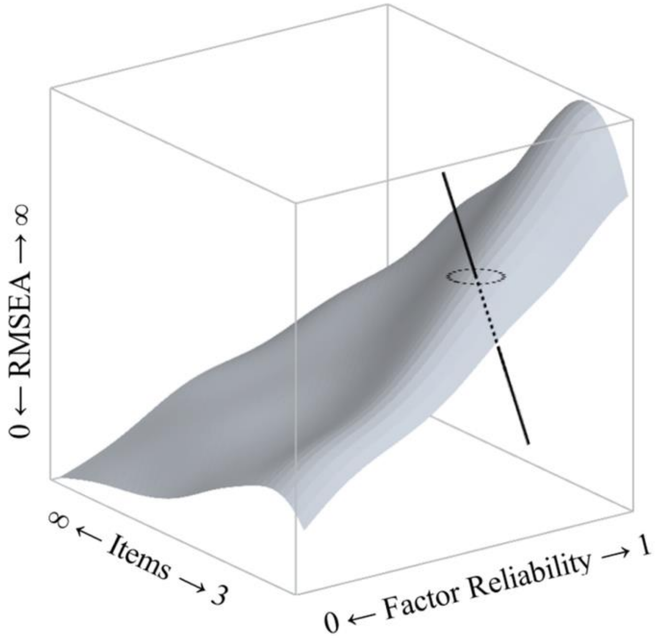


Figure 1. Hypothetical plot showing the variability in sensitivity of RMSEA to a particular misspecification across the model space as a function of model characteristics. The intersecting line represents fixed cutoffs from a single simulation and the elliptical cross-section highlights the subspace generalizable from conditions in a single simulation study.

When methodologists use simulation studies to derive fit index cutoffs, they cannot explore all the recesses of this cavernous multidimensional model space simultaneously. Therefore, methodologists necessarily select a subspace upon which to focus. Following from this logic, simulations provide reasonable localized cutoffs that are applicable to detecting a

[2] Large values indicate good fit for some indices (e.g., CFI) while values near 0 indicate good fit for others (e.g., RMSEA, SRMR). We selected an arbitrary index to make the z-axis easier to interpret, but the general idea extends across indices (though not necessarily in equal amounts; Miles & Shevlin, 2007)

misspecification with a particular magnitude in the selected subspace, but cutoffs do not generalize beyond the selected subspace nor to other misspecifications (Yuan, 2005). The conditions used in any single simulation (such as the one used to arrive at the traditional fixed cutoffs) are hypothetically represented by the line intersecting the surfacing in Figure 1 and the dashed cross-section represents the localized area to which such conditions may be generalizable. The derived cutoffs are perfectly reasonable for this narrow subspace, but what happens if one's empirical model is outside of this subspace? Current practice implicitly assumes that the traditional fixed cutoffs from this subspace generalize across the entire model space; however, studies have unambiguously shown that fit index cutoffs fluctuate and are not fixed across the model space (e.g., Hancock & Mueller, 2011; Heene et al., 2011; Shi et al., 2019). In fact, the fluid behavior of model fit measures was proved mathematically by Schönemann (1981), predating fixed cutoffs by nearly 20 years.

As methodologists have repeatedly pointed out, fixed cutoffs are inherently at risk of overgeneralization because there is no single global definition of "good" fit index values. Localized approximations of "good" can be derived, but such values are dependent on a complex multiway interaction of several factors. Essentially, the mapping of misfit magnitude to fit index values is conditionally monotonic but not globally monotonic. For example, an RMSEA of .04 indicates less misfit than an RMSEA of .06 if the model characteristics are constant. However, an RMSEA of .06 might indicate small and possibly forgivable misspecification in some contexts (e.g., the conditions used in Hu & Bentler, 1999) but an RMSEA of .04 might indicate a grossly misspecified model in other contexts (e.g., items that uniformly have weak standardized loadings; Hancock & Mueller, 2011; Heene et al., 2011). Even identical fit index values can be interpreted differently depending on the context such that an RMSEA of .06 in one context does imply the same magnitude of misspecification in another context (Yuan, 2005). Hu and Bentler (1998) themselves have pointed this out by saying "it is difficult to designate a specific cutoff value for each fit index because it does not work equally well with various types of fit indices, sample sizes, estimators, or distributions" (p. 449). Fit indices can be directly compared if model characteristics are constant but cannot necessarily be compared across models with different characteristics because misfit is encoded differently (Schönemann, 1981). This has been a principal grievance with the traditional fixed cutoffs – if the misfit is encoded differently for different model characteristics and the mapping of model misfit to fit index values is not globally monotonic, how can we justify a single cutoff? That is, an RMSEA of .06 indicates reasonable fit in one simulation study but those simulations conditions are not necessarily informative for interpreting an RMSEA of .06 in other conditions.

This is similar to power analysis where the appropriateness of an identical sample size can fluctuate in different contexts. *N* = 50 might be sufficient for a randomized trial with repeated measures to detect a large effect size but *N* = 50 might be absurdly low for a between-subjects comparison to detect a low effect size. Current use of fixed cutoffs is akin to sample size planning decades ago where *N* = 30 per condition was considered sufficient regardless of other pertinent aspects. Just like the traditional fixed cutoffs, *N* = 30 per cell is sufficient in a limited subspace of possible contexts but it can be highly inaccurate in others and knowledge that *N* = 30 is sufficient in one context does not necessarily imply that *N* = 30 is sufficient in a different context. A priori power analysis has evolved to derive better localized solutions (i.e., exactly with analytical approaches for simple models or approximately with Monte Carlo methods for complex models). The ultimate goal of this paper is to derive fit index cutoffs for a localized subspace of interest in order to better tailor fit indices to the characteristics of the data and model being evaluated. Putting this into the context of Figure 1, the intersecting line is dynamically moved around the model space to locate the appropriate localized cutoff for the model being evaluated so that misspecifications can be more accurately quantified.

To show evidence for how fit index cutoffs fluctuate based on variations in model and data characteristics, we replicate the simulation of Hu and Bentler (1999) in the next section but manipulate factor reliability, an aspect that was not manipulated in the original study. This also provides an opportunity to walk through the logic of their seminal simulation study for readers who may have committed the traditional fixed cutoffs to memory but are less familiar with their origin.

Hu and Bentler (1999) Replication Simulation

In this section, we focus on the effect of factor reliability on fit index cutoff values, as this was not systematically manipulated in the original Hu and Bentler (1999) study. Factor reliability – which is a function of standardized factor loadings – has been shown to lead to vacillations in potential cutoff values. Heene et al. (2011) provide a proof of the direct *inverse* relation between factor reliability and fit index values, meaning that high factor reliability (attained via standardized loadings with magnitudes close to 1) leads to seemingly *worse* fit relative to traditional fixed cutoffs. Plainly worded, the more reliable information one has about a latent variable, the more clearly misspecifications can be detected.

This has led to the term *reliability paradox* being applied to fit index interpretation (Hancock & Mueller, 2011) with Miles and Shevlin (2007) facetiously noting “if you wish your model to fit, … ensure that your measures are unreliable” (p. 874). To emphasize, the reliability paradox signifies that models with unreliable factors clear traditional fixed cutoffs more easily and would be more likely to be classified as “good” while models with highly reliable factors will encounter much more difficulty surpassing traditional fixed cutoffs and will be more likely to be classified as “bad”, even with an identical misspecification. This also applies to the χ^2 test where power is a function of communalities as proved by Schönemann (1981) and discussed in depth by Browne et al. (2002), meaning that an identical misspecification increases the χ^2 statistic more in models where factor reliability is higher. This is, of course, problematic because it essentially punishes researchers for having more reliable measures.

Hu and Bentler (1999) Simulation Design

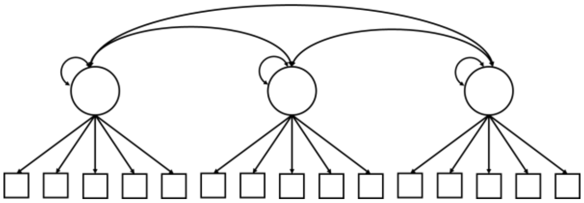
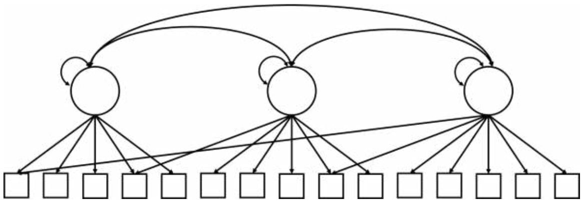
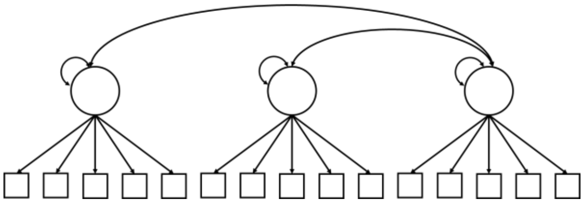
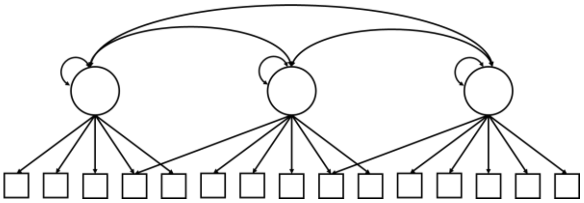
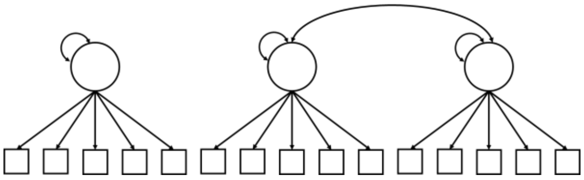
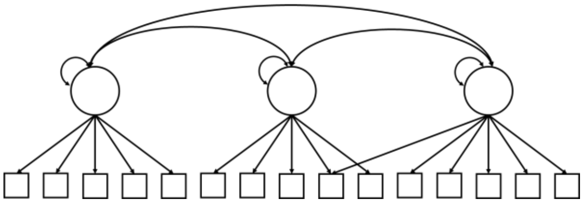
The simulation design exactly replicates the design used in Hu and Bentler (1999). They used two separate models, one deemed the “Simple” model and the other deemed the “Complex” model. The Simple model is a 3-factor CFA model where each factor loads on 5 items and there are no cross-loadings. The Complex model adds 3 cross-loadings to the Simple model: Item 1 loads on Factor 1 and Factor 3, Item 4 loads on Factor 1 and Factor 2, and Item 9 loads on Factor 2 and Factor 3.

For each model, there are three misspecification conditions: True, Minor, and Major. For both models, the empirical model in the True condition exactly matches the data generation model such that there are no misspecifications. For the Simple Model, the Minor condition omits the covariance between Factor 1 and Factor 2 and the Major condition omits all factor covariances involving Factor 1. For the Complex Model, the Minor condition omits the factor loading between Factor 3 and Item 1 and the Major condition omits this same loading and the loading from Factor 2 to Item 4. Table 1 displays conceptual path diagrams for each of these conditions.

There are 7 sample size conditions: 150, 250, 500, 1000, 2500, and 5000. In Hu and Bentler (1999), there are also 7 conditions for normality of the latent factors and/or the error variances. Robustness to non-normality was a greater concern in 1999 than it is now because there were fewer estimators available to accommodate non-normality. Hu and Bentler (1999) alluded to the promise of the Satorra-Bentler scaled test statistic in their discussion, which has since been shown to satisfactorily provide fit criteria that are robust to non-normality (Satorra & Bentler, 2001). Given this advance since the original study, we deviate from the original study's design by only including the multivariate normality condition. All items were continuous and generated to have a mean of 0.

Table 1

Path diagrams of different model and misspecification conditions in Hu and Bentler (1999)

	Simple Model	Complex Model
True		
Minor		
Major		

Note: Error variances for the observed variables are present but not shown in the path diagrams.

In the original study, the factor loadings were not manipulated and were kept constant across all conditions. In the Simple model, the first two items on each factor had loadings of .70 (Items 1, 2, 6, 7, 11, and 12), the middle item on each factor had loadings of .75 (Items 3, 8, and

13), and the last two items on each factor had loadings of .80 (Items 4, 5, 9, 10, 14, and 15). The error variances were chosen such that the items were standardized, meaning that these loadings are standardized. Using Coefficient *H* to estimate factor reliability, these loadings yield a population factor reliability of .87 for each of the three factors. In the Complex Model condition, the loadings were the same as in the Simple Model and the three additional cross-loadings were each set to .70. All items in the Complex Model were standardized, except the items with cross-loadings (Items 1, 4, and 9) whose error variances remained equal to .51, .36, and .36, respectively, meaning that the total variance for these items exceeded 1 and neither the items nor the loadings were standardized.

Manipulating Factor Reliability

As noted above, since the publication of Hu and Bentler (1999), methodological research has shown that factor reliability is directly related to the size of fit indices (Chen et al., 2008; Hancock & Mueller, 2011; Heene et al., 2011; McNeish et al., 2018; Saris et al., 2009). To assess how the cutoffs from the exact same data generation model would vary if the model subspace were broadened slightly, we varied the standardized loadings such that the middle item loadings were 0.35, 0.45, 0.55, 0.65, 0.75, and 0.85. We kept the pattern whereby the first two item loadings on each factor were .05 smaller than the middle item and the last two item loadings on each factor were .05 larger than the middle item. In the Complex Model, the cross loadings were generated to be equal to the value of the lowest item factor loading (i.e., the middle item condition minus .05). These loading conditions resulted in Coefficient *H* values of .42, .57, .69, .79, .87, and .94, respectively.

All the simulation conditions are fully crossed with 252 cells (7 sample sizes × 6 factor reliabilities × 6 model conditions) with 500 replications per cell of the design (increased from

200 replications in the original study). Data were generated and analyzed in SAS using Proc IML and Proc Calis using maximum likelihood estimation. All data generation and analysis files are available on the first author's Open Science Framework page (https://osf.io/wg45r/). SAS was used for consistency as it was the software used in the original study.[3] The results focus on SRMR, RMSEA, and CFI as these are commonly reported fit indices (Jackson et al., 2009).

Outcomes

We tracked fit index values across the replications to locate the value that discriminates between fit index distributions from models that are known to be correctly specified and models that are known to be misspecified. The general guideline used in Hu and Bentler (1999) was to select a cutoff that was able to reject a high percentage of models in the Minor misspecification condition (i.e., having a high true positive rate; similar to power expect that there are no null hypotheses for fit indices) while rejecting very few of models in the True condition (i.e., a low false positive rate; similar to Type-I error rate). The thresholds corresponding to this guideline were that 95% or more of misspecified models be rejected while no more than 5% of true models be rejected (i.e., such that the false positive and false negative rates were both less than or equal to 5%; Hu & Bentler, 1999, p. 16). In some conditions in the original study, these thresholds were not met and the false positive and false negative rate reached as high as 10%. This can be seen in the recommendation of a combination of TLI $\geq .95$ and SRMR $\leq .09$ for $N \leq 500$, where the sum of the false positive and false negative rates was 18.6% for the Complex model in Table 4 of Hu & Bentler (1999, p. 18).

[3] The EQS software was also used but this software currently has much less exposure than it did at the time of the original study, so we contain all analysis within SAS given that the fit index formulas are the same and also to avoid calling multiple programs

Figure 2 visually demonstrates the process by which cutoffs would be hypothetically derived for SRMR. The distribution of SRMR for misspecified models across simulation replications is shown in light grey and the distribution of SRMR values for correct models is shown in dark grey. The goal of the simulation is to locate the value of SRMR that would be highly likely to belong to the misspecified model fit index distribution but that would be highly unlikely to belong to the true model fit index distribution. The precise definition of "highly likely" is that 95% of misspecified models are at or above the cutoff while no more than 5% of true models are at or above the cutoff (and if that condition cannot be met, the percentages are expanded to 90% and 10%, respectively, instead of 95% and 5%). In Figure 2, an SRMR value of 0.040 meets these criteria – nearly all misspecified models have an SRMR value of 0.040 or higher while nearly all true models have an SRMR below 0.040. So, if a model returns an SRMR of 0.040 or above with these model and data characteristics, one could be confident that the model contains a misspecification at least as large as the "minor" misspecification induced in Hu and Bentler (1999) and is unlikely to be exactly correct.

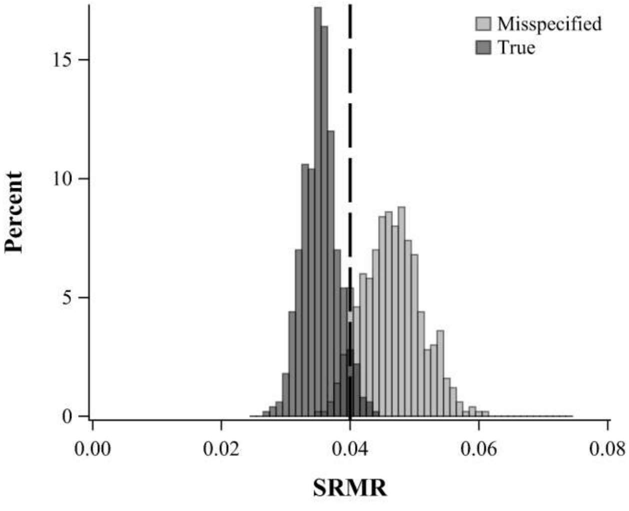


Figure 2. Comparison of SRMR when the true model is fit versus a misspecified model. The black vertical line indicates the cutoff value of .04 that distinguishes between the two distributions

Results

Table 2 presents fit index cutoffs from the replication simulation by sample size and factor reliability for the minor misspecification. Following Hu and Bentler (1999), we reduced the sample size conditions into three groups: ≤ 250, 500, ≥ 1000. The H = .87 row that uses the same loadings from Hu and Bentler's study (1999) is shaded in grey. For consistency, we applied two conditions to derive cutoffs, (a) false positive and false negative rates are both 5% or less and (b) false positive and false negative rates are both 10% or less. Condition (a) is reported whenever it exists; condition (b) is only reported if condition (a) cannot be met. Such cases are marked with an asterisk in Table 2.

Table 2
Fit index cutoffs from replication simulation by sample size and factor reliability

		$N \leq 250$			$N = 500$			$N \geq 1000$		
Coefficient H	Middle Loading	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI
.42	.35	None	None	None	≤ .040*	None	None	≤ .029	None	None
.57	.45	None	None	None	≤ .050	None	None	≤ .046	≤ .018	≥ .982
.69	.55	≤ .068	None	None	≤ .064	≤ .025*	≥ .981*	≤ .066	≤ .031	≥ .973
.79	.65	≤ .087	≤ .038*	≥ .975*	≤ .087	≤ .042	≥ .969	≤ .094	≤ .048	≥ .962
.87	.75	≤ .115	≤ .053	≥ .972	≤ .120	≤ .061	≥ .962	≤ .129	≤ .066	≥ .956
.94	.85	≤ .142	≤ .080	≥ .960	≤ .154	≤ .085	≥ .956	≤ .165	≤ .089	≥ .952

Note: Entries without an "*" indicate that false negative and false positive rates both were below 5%. Entries followed by "*" indicate that at least one rate exceeded 5% but that both rates were less than 10%. Entries listed as "None" convey that there was no cutoff value for which both rates were below 10%. Middle Loading = the standardized loading of the third item on each factor

In the H = .87 row, the RMSEA and CFI cutoffs closely reflect the traditional ≤ .06 and ≥ .96 cutoffs, respectively. The SRMR cutoff is near ≤ .12 and appears different than the traditional ≤ .08 value; however, note that the traditional SRMR cutoff came from combination rules[4] whereas the values in the table are obtained from false positive and false negative rates. When applying rate rules for cutoffs rather than combination rules, Hu and Bentler (1999) similarly arrive at SRMR cutoffs of ≤ .11 or ≤ .12 (e.g., first row of Table 3 on p. 15; text on pp. 16, 22, 26, and 27. This is also discussed on p. 320 of Heene et al., 2011).

As factor reliability increases, the cutoff value needed to detect the same misspecification moves away from exact fit (higher for SRMR and RMSEA, lower for CFI). For instance, with H = .94 and $N \geq 1000$, an RMSEA cutoff of ≤ .089 has equivalent ability to differentiate correct from misspecified models as the traditional fixed cutoff for the conditions in Hu and Bentler (1999). Conversely, clearing traditional fixed cutoffs with lower factor reliability does not necessarily indicate good fit. For $H = .57$ and $N \geq 1000$, an RMSEA cutoff of 0.018 has equivalent ability to differentiate true from misspecified models as traditional fixed cutoffs for the conditions in Hu and Bentler (1999). Additionally, if factor reliability and sample size were both low, there are no cutoff values that can consistently distinguish between true and misspecified models because distributions of fit index values for true and misspecified models overlap too much. Figure 3 shows an example of one such case where excessive overlap in SRMR distributions that would not yield a suitable cutoff because SRMR values for these data and model characteristics could conceivably come from either distribution, meaning that a cutoff would be unable to control false positive and false negative rates.

Substituting the cutoffs in Table 2 as new fixed cutoffs would be slightly more generalizable than traditional fixed cutoffs because they explore more of the model space to account for factor reliability. However, they would still suffer from issues related to degrees of

[4] Hu and Bentler (1998) recommended a two-index strategy whereby SRMR is paired with other complementary indices to increase sensitivity to underparameterized models. The cutoff value of .08 for SRMR reflects optimal classification rates when used in *combination* with other indexes (RMSEA, in particular) rather than in isolation (e.g., Hu and Bentler, 1999 p.6).

freedom, number of items, number of factors, model size, etc. The cutoffs in Table 2 are only applicable to the model space corresponding to a 3-factor CFA model with 87 degrees of freedom and 5 items per factor and are only applicable to the single misspecification represented the minor condition contained in Hu and Bentler (1999). The next section discusses previous work related to generalizing cutoffs with simulation techniques such that the cutoffs are tailored to the exact model and data characteristics being evaluated. We also contextualize the contribution of the current paper in extending this literature in this section.

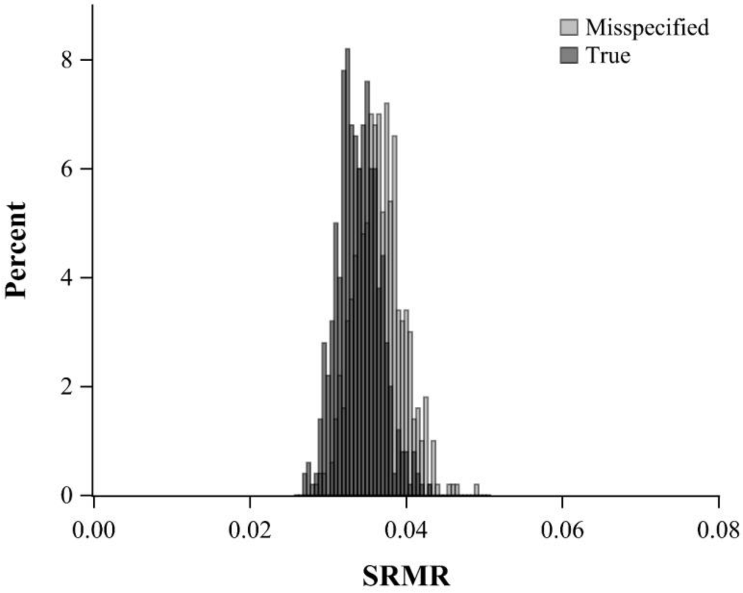


Figure 3. Comparison of SRMR distributions when the overlap is too great to provide a cutoff that could reliably distinguish between the two distributions (i.e., false positive and false negative rates would be excessive). The “None” entries in Table 2 correspond to this type of result.

Simulating Dynamic Cutoffs

Limitations of basing fixed cutoffs on a single set of simulation conditions has been noted (e.g., Marsh et al., 2004) and an alluring proposal has been to perform a custom Monte Carlo simulation based on the specific data and model characteristics of the model under evaluation (Millsap, 2007, 2013; Pornprasertmanit et al., 2013). That is, the Hu and Bentler (1999) cutoffs were derived from a simulation, so by conducting a simulation but altering the conditions, researchers can essentially follow the same logic as Hu and Bentler (1999) and arrive at cutoffs that detect misspecifications in the localized subspace occupied by their model and data.

The core idea is that fit index cutoff derivation is essentially a power analysis. The goal is to uncover fit index values that detect a misspecification of a prespecified magnitude just like power analysis is used to uncover the sample size needed to detect a non-null effect with a particular magnitude with some prespecified probability. Using traditional fit index cutoffs is akin to using a single power analysis across all studies and assuming it generalizes to all situations. Similar to how modern power analyses have evolved to be customized to researchers' unique circumstances, fit index cutoffs are most informative when derived under conditions that closely match the specific data and model characteristics being evaluated (e.g., sample size, number of factors, number items, etc.). Despite the obvious draw of dynamically simulating custom fit index cutoffs, implementation of the method has been limited (e.g., Millsap, 2007 has 113 total Google Scholar citations as of this writing, many of which are citations from methodological articles rather than empirical studies) as researchers continue to rely on traditional fixed cutoffs even though alternatives have existed for many years.

Three barriers to implementation are present when calling for researchers to conduct simulations to custom tailor fit index cutoffs and may be limiting application of simulation-based methods in empirical studies. First, simulation is a methodological skill that many empirical researchers do not possess (e.g., Arend & Schäfer, 2019; Green & MacLeod, 2016). Though there are accessible software options to remove some barriers to conducting a simulation study such as the MONTECARLO utility in M*plus* (Muthén & Muthén, 2002) or the simsem R package (Pornprasertmanit et al., 2020), even with user-friendly software it can be difficult for researchers to engage with these tools if they are unfamiliar with the underlying foundations of simulation techniques. Second, similar to doing a priori power analysis for sample size planning with structural equation models (e.g., Anthoine, Moret, Regnault, Sebille, & Hardouin, 2014; Lai & Kelley, 2011; MacCallum, Browne, & Sugawara, 1996), it can be difficult to articulate the magnitude of the misspecification that is meaningful for complex multivariate models. Third, even when researchers do possess requisite skills to conduct simulation studies, the cost-benefit tradeoff involved with manually coding multiple simulations to derive a custom cutoff for a single model may not be a worthwhile investment, especially if reviewers are content with traditional fixed cutoffs.

Proposed Dynamic Fit Index Cutoff Approach

To address these limitations and increase the ease with which researchers can derive fit index cutoffs that are more attuned to a researcher's specific data characteristics and model, we propose an algorithm to further lower barriers to implementation. We refer to this general approach as *dynamic fit index* (DFI) cutoffs.

First, we provide researchers with a Shiny software application that automates design and execution of a Monte Carlo simulation without any explicit existing knowledge of how to program a simulation. Existing software resources automate some aspects of a simulation study but still require that researchers manually write at least some software code. However, our software application requires no manual code to be written nor does it require users to open or download any specific software program. The program is hosted on a website where it can be freely accessed by users who only need to provide the standardized estimates from their model. From this, any researcher with an internet connection can partake in custom simulations to derive fit index cutoffs, not just those researchers with training in Monte Carlo techniques. This also reduces the time investment for researchers who are knowledgeable about Monte Carlo techniques but would otherwise have to write code from scratch.

Second, because selecting the level of misspecification is difficult even if one is familiar with Monte Carlo techniques, our algorithm generalizes the misspecifications used by Hu and Bentler (1999) to other multifactor models (special considerations are necessary for one-factor models, which are discussed in detail later in this paper). Though other methods for capturing misspecifications in covariance matrices have been suggested (e.g., Maydeu-Olivares, 2017; Wu & Browne, 2015), researchers overwhelmingly continue to reference the traditional fixed cutoffs, which would seem to imply at least tacit interest in which fit index values can identify the type of misspecifications encoded in these traditional cutoffs (granted, this assumes that researchers know the origin of the traditional fixed cutoffs, which may not be the case). This removes potential guesswork involved in deciding what type of misspecification researchers wish to encode in their simulation while also making the DFI cutoffs consistent with guidelines used for deriving the traditional cutoffs.

Of course, prespecifying the misspecification eliminates users' ability to customize the misspecification used in the simulation. Our software partially addresses this by featuring multiple types of misspecifications to give researchers a range of potential options (specific details are discussed shortly). Such a practice would be analogous to limiting the possible effect sizes in a power analysis to a select number of prespecified values. However, power analyses for sample size planning rarely stray from preselected cutoffs such as 0.20/0.50/0.80 for Cohen's *d* without a reduction in the utility of the approach.

Nonetheless, our hope is that giving all researchers (regardless of quantitative savviness) an approachable tool by which to simulate custom cutoffs – even if only in a constrained manner

– will (a) raise awareness of the precariousness of fit index cutoffs across different data and model characteristics, (b) foster a greater appreciation of the origins of the traditional cutoffs, and (c) encourage researchers to be more attentive to how they use fit indices. We also hope that it moves the methodological literature away from bemoaning generalizability issues with the traditional fixed cutoffs (which, as we demonstrate, can be addressed with the current open-source computing environment) and towards thinking about the types of misspecifications to which fit indices should be sensitive and creating ways to better quantify model misspecification so that fixed cutoffs become obsolete and reliance upon them is discontinued.

The next section outlines the algorithm we implement to incorporate the misspecification used in Hu and Bentler (1999) across more model and data conditions to generalize cutoffs. After going through these details, we discuss how to expand the algorithm to misspecifications with varying levels of severity to produce a broader set of cutoffs rather than a single binary cutoff that often leads to fit indices to be treated as ad hoc hypothesis tests. Afterwards, we discuss special considerations for one-factor models common in scale validation.

Algorithm to Derive DFI Cutoffs

DFI cutoffs retain the idea of using simulation to derive suitable fit index cutoffs but differ in that the model upon which the simulation is not fixed but rather is updated to match the empirical model being evaluated. Specifically, the steps of the DFI algorithm for generalizing the Minor misspecification from Hu and Bentler (1999) to other multifactor models is:

1. Fit the empirical model and obtain the standardized parameter estimates.
2. The standardized estimates from the empirical model are used to create a data generation model for a subsequent Monte Carlo simulation. Rather than use the empirical model as the data generation model, an additional path of a magnitude

that corresponds to Hu and Bentler's Minor misspecification condition is added to the data generation model. In other words, it is assumed that the empirical model now represents the "Minor" misspecification row in Table 1 such that the "true" model has an extra path not present in the empirical model. The goal is to reverse engineer a plausible model for the "True" row of Table 1, as if the researcher's empirical model were misspecified.

3. The model created in Step 2 is used to generate 500 datasets. The empirical model is fit to each generated dataset, producing a fit index that captures the value of a model containing a misspecification of the magnitude used in the Minor condition of Hu and Bentler (1999) for the characteristics of the researcher's empirical model.
4. The empirical distribution of fit indices from analyses of all generated datasets is then formed. The 5th percentile of the fit index distribution for lower-is-better indices (or the 95% percentile for higher-is-better indices) is the value of the fit index that consistently detects a misspecification of similar magnitude as used in Hu and Bentler (1999) for the researcher's empirical model characteristics.
5. Repeat Step 2 through Step 4 but change the data generation model in Step 2 to be exactly equal to the empirical model to inspect behavior of fit indices if the empirical model were congruent with the data generation model (i.e., treat the empirical model as the True row in Table 1). The cutoff value derived in Step 4 should be further from exact fit (i.e., further from 0 for SRMR and RMSEA, further from 1 for CFI) than when the data generation model is congruent with

the empirical model to ensure that fit values are not ambiguous and that the cutoff can distinguish between correct and misspecified models.

5b. If the cutoff value derived in Step 4 is closer to exact fit, then use the more lenient criteria using the 10th percentile in Step 4 and 90th percentile in Step 5 for lower-is-better indices (and vice versa for higher-is-better indices). If no such value exists with the more lenient criteria, then the characteristics of the model are not amenable to a cutoff that can unambiguously distinguish between congruent and misspecified models.

The logic of the strategy is that it mimics the strategy used to derive the traditional fixed cutoffs but adapts the simulation design to the model subspace that corresponds to the empirical model. A bottom-up approach is taken whereby the empirical model is used to create a plausible data generation model from the same model subspace for purpose of comparison. That is, whether the data generation model in Step 2 is the "true" model is irrelevant because the goal is merely to identify a model whose model-implied covariance structure differs by an amount consistent with the Minor misspecification conditions from Hu and Bentler (1999). That is, the goal is not to identify misspecifications in the empirical model but rather to determine the scaling of fit indices for the model being evaluated.

This approach follows previous work in the causal inference literature on omitted confounders (Rosenbaum, 2002, 2010; Rosenbaum & Rubin, 1983) where the interest is not in identifying the omitted variables themselves but rather to quantify possible effects of hypothetical unseen or uncollected variables. For instance, the sensitivity test from Rosenbaum (2002) calculates how large the effect of an omitted cofounder would need to be in order to alter a conclusion that treatment effect is significant. The Rosenbaum's method is not interested in identifying the omitted variable. Instead, it gives researchers an idea of the magnitude of effect an omitted variable would need to have in order to reverse the conclusion so that researchers can consider whether variables they did not collect might have an effect with such a magnitude. Harring et al. (2017) adapted this approach into structural equation modeling to test whether omitted variables could change conclusions and similar ideas have been suggested for sensitivity analyses in mediation models (MacKinnon & Pirlott, 2015).

The DFI approach follows the same thought process – there is no presupposition that the data generation model in Step 2 is actually the true model nor is the goal to isolate and identify the true model. The point of Step 2 is merely to reverse engineer a hypothetical model that would render the empirical model misspecified to a similar degree as in Hu and Bentler (1999) in order to determine what the distribution of fit index values would look like for a misspecified model in the subspace occupied by the empirical model. In other words, the scaling of the fit indices is dependent on data and model characteristics and using a consistent, representative misspecification in the data generation model (such as the one used by Hu and Bentler) helps to uncover the scaling of the fit indices. The end-goal is to determine how a particular misspecification is quantified by fit indices under particular data and model characteristics to derive a custom cutoff value with similar properties as the traditional fixed cutoffs anywhere in the subspace of CFA models regardless of the number of items, number of factors, factor reliability, degrees of freedom, etc.

Note that the actual misspecification in the empirical model could be different in magnitude and pattern than the misspecification featured in the data generation model. The data generation model is selecting but one possible misspecification to help better understand the scaling of fit indices for a specific set of data and model characteristics. This helps to anchor the scale of the fit indices in a particular subspace, but how the misfit is distributed throughout the empirical model is not necessarily the same as the representative misspecification used in the DFI simulations. The DFI simulations are a tool to guide interpretation of fit indices across models but do not imply anything about the nature of specific misspecifications in the empirical model.

Though challenges in generalizing the Hu and Bentler (1999) misspecification is one presumed reason for lack of implementation in previous simulation-based techniques for deriving custom cutoffs, lack of familiarity with simulation techniques in general is another presumed reason for lack of application of such methods in empirical studies. Therefore, we provide a web-based Shiny application that only requires users to enter their sample size and upload a .txt file with their model's standardized estimates to a web browser that then automates Steps 2 through 5 and reports the DFI cutoffs corresponding to their model. We walkthrough an example analysis in the next section to elucidate the details of the method.

Simulated Data Example

In this section, we demonstrate the DFI approach for deriving RMSEA, SRMR, and CFI cutoffs for multifactor models using the data from Replication 1 of the Complex model, N = 500, H = 0.69 conditions from the simulation whose results were presented in Table 2. We begin with generated data because – unlike empirical data – we know the true model so we can compare the results of our procedure to the results using the true model in order to better gauge accuracy. We go through each of the 5 steps in detail to clarify the process of obtaining DFI cutoffs.

Step 1: Fit the Empirical Model

The empirical model is a 3-factor model where all factors covary with each other. Items 1 through 5 load on Factor 1, Items 6 through 10 load on Factor 2, Items 11 through 15 load on

Factor 3, Item 4 cross-loads on Factor 2 and Item 9 cross-loads on Factor 3. The path diagram for the empirical model is shown in Figure 4. We fit the model in M*plus* Version 8.3 with maximum likelihood estimation. Model fit metrics are reported in Table 3 along with both unstandardized and standardized parameter estimates. The RMSEA was .033 (90% CI = [.021, .044]), SRMR was .041, and CFI was .970. If referencing traditional fixed cutoffs, the model would appear to fit well. However, the factor reliability values are in the low .60s/high .70s and are far below the .87 factor reliability used to derive the traditional fixed cutoffs. So even though the model characteristics are identical to those from Hu and Bentler (1999) in all respects but one, the reduced factor reliability could potentially place this model in a subspace where fit indices quantify misfit differently.

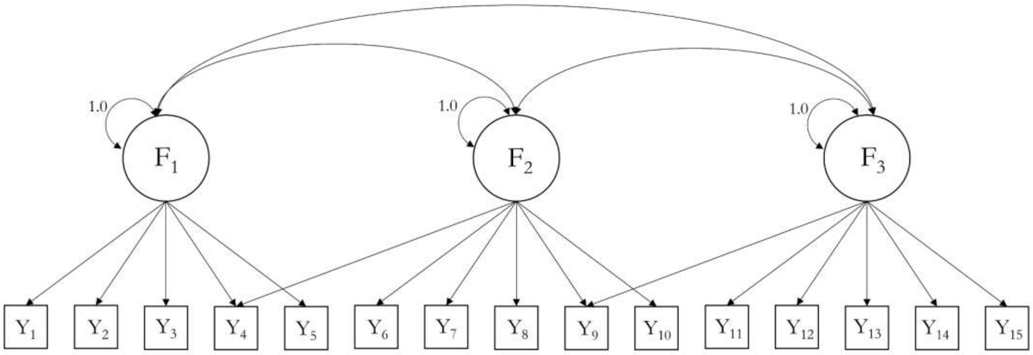


Figure 4. Path Diagram for empirical model

Table 3

Empirical model estimates for replication 1 of the N =500, H *=0.69, complex condition*

Loadings			
	Item	Unstandardized	Standardized
Factor 1			
	1	.861	.705
	2	.444	.445
	3	.478	.515
	4	.453	.373
	5	.480	.497
Factor 2			
	4	.593	.489
	6	.601	.595

Factor Correlations		
Factor 1	Factor 2	.485
Factor 1	Factor 3	.657
Factor 2	Factor 3	.196
Error Variances	Unstandardized	Standardized
Item 1	.752	.503
Item 2	.798	.802
Item 3	.632	.735
Item 4	.653	.444
Item 5	.700	.753
Item 6	.658	.645

	7	.540	.507	Item 7	.845	.743
	8	.565	.559	Item 8	.703	.688
	9	.634	.532	Item 9	.693	.488
	10	.627	.638	Item 10	.573	.593
Factor 3				Item 11	.759	.702
	9	.460	.386	Item 12	.679	.706
	11	.567	.546	Item 13	.670	.771
	12	.532	.542	Item 14	.627	.676
	13	.447	.479	Item 15	.595	.605
	14	.549	.570	Model Fit		
	15	.623	.628	RMSEA		.033
Factor Reliability (Coefficient *H*)				RMSEA 90% CI		[.021, .044]
Factor 1		.676		SRMR		.041
Factor 2		.733		CFI		.970
Factor 3		.710		χ^2 (85)		131.39, p <.01

Note: Factors are given scale by constraining factors variances to 1 for all factors; factor correlations are equal to factor covariances

Step 2: Create the Data Generation Model

We use the standardized estimates from Table 3 to create a data generation model for a Monte Carlo simulation with one addition: we add a cross-loading to mimic the Minor misspecification condition in Hu and Bentler's Complex model. The logic of Hu and Bentler's Complex model misspecification was to determine the fit index value that could consistently identify an omitted cross-loading whose magnitude is equal to the weakest loading present in the model. Item 2 has the weakest standardized loading (.445), so the data generation model will be the same as the empirical model *with an additional cross-loading* to Item 2 with the value equal to .445 (the standardized path of this weakest item).[5] This cross-loading is added to the factor on which the item did not originally load that has the highest factor reliability because this will provide the largest misspecification that could be associated with such a cross-loading (it also

[5] Item 4 and Item 9 have lower standardized loadings for one factor, but higher standardized loadings for another factor. If items load on multiple factors, we consider the item with the highest standardized residual variance. The magnitude of the path that can be added to the data generation model is limited by the size of the standardized residual variance, so the item with a larger standardized residual variance provides the most flexibility. This is discussed in detail in the next section.

facilitates replicating the same data generation model if the method is applied repeatedly). In this example, Item 2 originally loaded on Factor 1, so the cross-loading would originate from Factor 2 because its Coefficient *H* is slightly higher than Factor 3 (0.733 vs. 0.710). In this way, we treat the empirical model as if it has a minor misspecification in attempt to determine fit index values that detect a misspecification consistent with an omitted cross-loading with the same magnitude as the weakest loading in the empirical model.[6] The path diagram of the data generation model is shown in Figure 5.

Importantly, note that the true model does not need to be known when forming the data generation model. In this example using simulated data, we are privy to the fact that the true model has no cross-loading from Factor 2 to Item 2 but rather that the actual omitted path in the empirical model is a loading from Factor 3 to Item 1. This is irrelevant for the DFI approach because the goal is not to uncover the true model. Rather, the goal is to generate data from a plausible model within the same approximate subspace as the empirical model that contains a misspecification with a similar magnitude used by Hu and Bentler (1999).

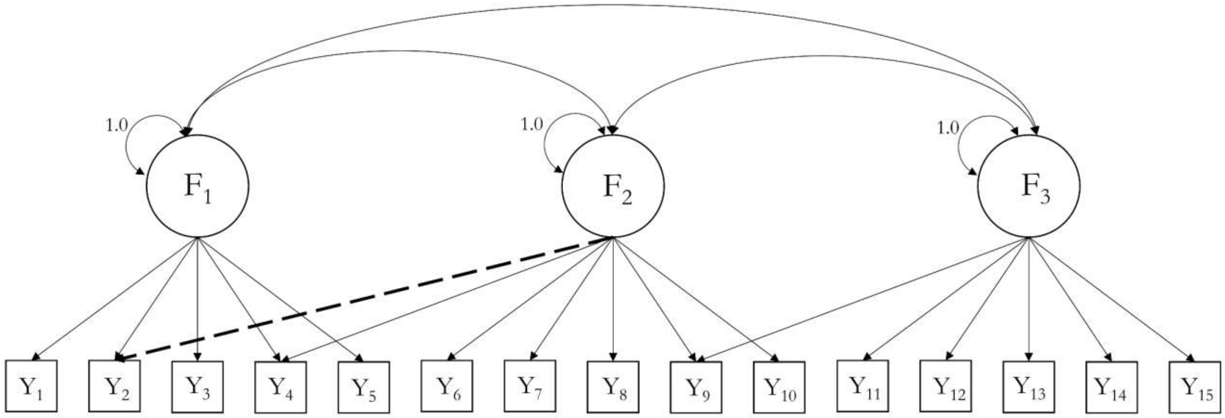


Figure 5. Path diagram for the data generation model. The addition path not included in the empirical model is represented by a bold dashed line.

[6] In the accompanying Shiny application, we select the lowest loading from factors with 3 or more items when possible. During experimentation with models with irregular characteristics, results were less stable when cross-loadings were added to an item from a two-item factor, presumably due to local under-identification. The application also will not run for models with very small degrees of freedom whereby additional paths would make the model just-identified or under-identified.

Ensuring the Cross-Loading is Admissible. Note that for models with high standardized loadings across all items, it is not possible to add a cross-loading with a high standardized loading. For example, if the lowest standardized loading present in the model was 0.80, adding a cross-loading of the same magnitude to the same item would lead to a model-implied variance of at least $0.80^2 + 0.80^2 = 1.28$, which would require a negative population residual variance to keep items on a standardized metric. Hu and Bentler (1999) encountered this same problem and addressed it by unstandardizing items with cross loadings in their data generation model (p. 7). To streamline software automatization, we take a slightly different approach.

Instead, we compare the magnitude of the lowest standardized loading in the empirical model to the maximum allowable cross-loading that would maintain a nonnegative population residual variance. For items that do not already have cross-loadings in the empirical model, the maximum allowable standardized cross-loading is

$$\lambda_{CL} < \sqrt{\lambda_O^2 \psi_{(F_O,F_{CL})}^2 + \theta_O} - \lambda_O \psi_{(F_O,F_{CL})} \tag{1}$$

where λ_{CL} is the standardized cross-loading to be added to the data generation model, λ_O is the original standardized loading in the empirical model, $\psi_{(F_O,F_{CL})}$ is the factor correlation between the factor the item originally loads on (F_O) and the factor associated with the additional cross-loading in the data generation model (F_{CL}), and θ_O is the original standardized error variance of the item in the empirical model. The standardized cross-loading is then equal to $\min\left[0.95\left(\sqrt{\lambda_O^2 \psi_{(F_O,F_{CL})}^2 + \theta_O} - \lambda_O \psi_{(F_O,F_{CL})}\right), \lambda_O\right]$. We multiply by 0.95 simply to prevent a nonpositive definite population covariance matrix that may be attributable to rounding error (this possibility occurred during experimentation using the exact maximum).

Using Table 3, the empirical model standardized loading of Item 2 (λ_O) was 0.445, the correlation between Item 2's original factor (Factor 1) and the factor associated with the cross-loading (Factor 2) is 0.485, and the standardized error variance of Item 2 in the empirical model (θ_O) is 0.802. Evaluating Equation 1 with this example would yield

$$\begin{aligned} &\lambda_{CL} < \sqrt{0.445^2 \times 0.485^2 + 0.802} - 0.445 \times 0.485 \\ &\lambda_{CL} < 0.705 \end{aligned} \tag{2}$$

The maximum standardized cross-loading Item 2 can accommodate is 0.705 because if a cross-loading of this size were added, the explained variance of Item 2 would be

$$0.445^2 + 0.705^2 + 2\left(0.445 \times 0.705 \times 0.485\right) = 1.00 \tag{3}$$

The lowest standardized loading of 0.445 is below 95% of the maximum allowable cross-loading in Equation 2 ($0.95 \times 0.705 = 0.670$), so we set the cross-loading equal to .445 in this example.

Factor Covariance Misspecification. In Hu and Bentler (1999), two misspecifications are present: one for misspecified cross-loadings (the Complex condition) and one for misspecified factor covariances (the Simple condition). In the original study, the Simple misspecification omitted a factor covariance such that two factors were made orthogonal. Though this involves removing only a single parameter, Fan and Sivo (2005) note that this has widespread consequences in the model-implied covariance matrix because it restricts the covariance between any items on the two orthogonal factors. Fan and Sivo (2005) note that reported sensitivity of SRMR to covariance misspecifications is primarily an artifact of this specific misspecification and is not a general rule.

Empirical models typically feature factor covariances, so the Hu and Bentler (1999) approach of making factors orthogonal is not implementable when trying to reverse engineer a data generation model from an empirical model. The Hu and Bentler (1999) misspecified models had orthogonal factors, but if the empirical model of interest posits non-orthogonal factors, this type of misspecification cannot be recreated because reverse engineering involves adding more parameters to the data generation model, which cannot be accomplished if those paths were already estimated as part of the empirical model. Put another way, because we effectively treat the empirical model of interest as the minor misspecification condition in the simulation, if the empirical model includes all possible factor covariances (as is the default in many software programs), the factor covariance matrix is saturated and cannot be misspecified.

Given this issue and based on findings from Fan and Sivo (2005), we do not differentiate types of misspecifications to derive cutoffs for different fit indices. That is, we rely on the procedure described in the Complex condition of Hu and Bentler (1999) to create the data generation model from the empirical model for all indices. As a result, we expect that our RMSEA and CFI values will closely match what is presented in Table 2, but the SRMR values are expected to be different (specifically, they will be smaller).

Step 3: Generate Data and Fit the Empirical Model

Multivariate normal data consistent with the model-implied covariance matrix from the data generation model are generated in a Monte Carlo simulation. Generated data contain the same number of variables and the same sample size as the empirical data (15 variables and 500 people, in this example). This is repeated such that there are 500 unique datasets, each containing 500 observations and 15 variables. The empirical model from Figure 4 is then fit to each simulated dataset such that the resulting fit index values reflect typical values from a misspecified model because the empirical model has fewer estimated paths than the data generation model. Again, the accompanying Shiny application executes this automatically so unfamiliarity with Monte Carlo simulation or an undesirable time investment related to writing simulations scripts from scratch are not barriers to implementation.

Step 4: Locate the 5th Percentile of the Fit Index Distribution

The Monte Carlo simulation generates 500 different datasets and each one is analyzed with the empirical model, meaning that there are 500 different sets of output. The distribution of fit indices from these 500 analyses are then summarized. To coincide with Hu and Bentler's method for deriving cutoffs, we are interested in the 5th percentile of the distribution for lower-is-better indices (SRMR and RMSEA) which identifies the fit index value to which 95% of misspecified model fit index values are equal or greater. That is, in the simulation we know that the empirical model that is fitted to the generated data is misspecified because the data generation model and the empirical model do not match, so the 5th percentile tells us the value of the fit index that would detect this misfit 95% of the time. CFI is a higher-is-better index, so this same information is captured by the 95th percentile of the distribution instead of the 5th percentile. Table 4 shows the 5th percentile of SRMR and RMSEA and the 95th percentile of CFI when the model is known to have a misspecification. Using the cutoffs shown in the "Value" column of Table 4 would accurately reject a model with this misspecification 95% of the time (i.e., a false negative rate of 5%) and therefore shows how misfit caused by an omitted cross-loading scales the fit indices for these data and model characteristics.

Table 4
5th percentile of SRMR and RMSEA and 95th percentile of CFI fit index distributions when the model is misspecified

Index	Distribution	Percentile	Value
RMSEA	Misspecified	5	.023
SRMR	Misspecified	5	.035
CFI	Misspecified	95	.987

Step 5: Repeat Using the Empirical Model as the Data Generation Model

The values from Step 4 are not necessarily the final cutoffs because we also need to ensure that these values do not excessively reject true models (i.e., that they do not yield high false positive rates). We repeat Step 2 through Step 4 but make the data generation model equal to the empirical model (i.e., the data generation model is the model shown in Figure 4 and there are no additional cross-loadings included). In doing so, we inspect the typical fit index values we would encounter in this subspace were the empirical model correct. Ideally, 5% or fewer of correct model replications will be rejected when using the cutoff value obtained in Step 4. That is, the 95th percentile of the correct model fit index distribution should be at or below the 5th percentile of the misspecified model fit index distribution for lower-is-better indices (and vice versa for higher is better indices). Otherwise, the overlap in fit index values for misspecified and correct models would be ambiguous since the same fit index value could conceivably come from either the correct or misspecified model distribution. Table 5 shows the 95th percentile of SRMR and RMSEA and the 5th percentile of CFI when the empirical model is correct and matches the data generation model.

Table 5
95th percentile of SRMR and RMSEA and 5th percentile of CFI fit index distributions when the model is correct

Index	Distribution	Percentile	Value	Step 4 Cutoff
RMSEA	Correct	95	.024	.023
SRMR	Correct	95	.033	.035
CFI	Correct	5	.984	.987

For SRMR, the 95th percentile for the correct model distribution (0.033) is below the 5th percentile of misspecified models (0.035), so the 0.035 value from Step 4 can reasonably differentiate between misspecified and correct models because the false positive and false negative rates are both below 5%. However, for RMSEA, the 95th percentile of the correct model distribution (0.024) exceeds the 5th percentile of misspecified models (0.023), meaning that the value from Step 4 is ambiguous in that it is observed for both correct and misspecified models. The same is true for CFI such that 5th percentile for the correct model distribution (0.984) is below the 95th percentile of the misspecified model distribution (0.987), so the CFI value obtained in Step 4 is similarly ambiguous.

Figure 6 demonstrates this possible ambiguity using CFI as an example. Based on the logic of deriving fit index cutoffs, values at or worse than the cutoff indicate unacceptable fit. In Figure 6, this would mean that CFI values at or to the left of the 95th percentile of the misspecified model distribution would be considered unacceptable. However, in this case, the 5th percentile of the correct model distribution falls in this area, meaning that an unacceptable proportion of correct models would be rejected (i.e., the false positive rate is too high). Therefore, the SRMR value in Step 4 is an acceptable cutoff for detecting misfit for this model, but the CFI and RMSEA values in Step 4 are unacceptable using a 5% threshold, so we test the 10% threshold in Step 5b.

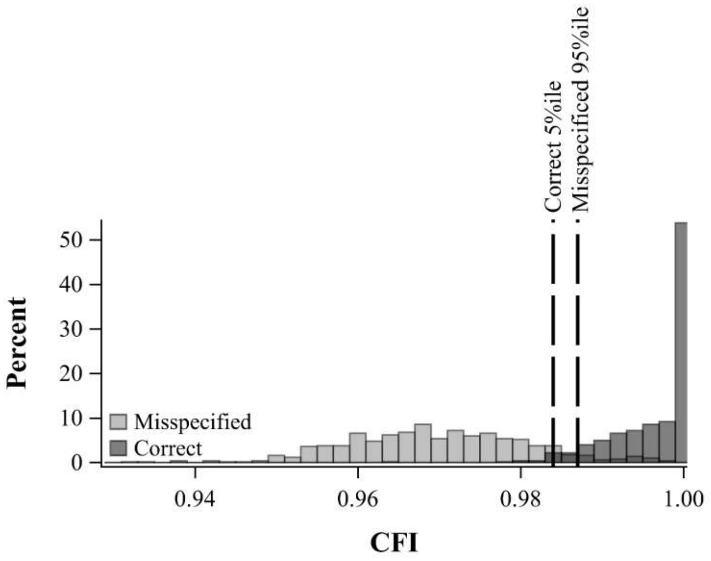


Figure 6. Comparison of CFI distributions when the model is misspecified and when the model is correct. The 95th percentile of the correct model distribution is further from perfect fit than the 5th percentile of the misspecified model distribution, indicating that the false positive rate would be too high and that this cutoff is not suitable to distinguish between distributions.

Step 5b. If the false positive and false negative rates cannot both be kept under 5%, then we move on to Step 5b and determine if there is a fit index value that keeps both rates at or below 10%. Table 6 shows the same information as Step 4 and Step 5 for RMSEA and CFI but expands the false positive and false negative rate thresholds to 10%. For RMSEA, the 10th percentile of the misspecified model distribution (0.026) is now further from exact fit than the 90th percentile of the correct model distribution (0.021). So, RMSEA ≤ 0.026 meets the 10% threshold and would be acceptable as a value to distinguish between true and misspecified models in this context.

Similarly, the 10th percentile of the CFI misspecified model distribution of 0.983 is further from exact fit than the 90th percentile of the CFI correct model distribution of 0.987. So, CFI ≥ 0.983 appears to be a reasonable CFI value for distinguishing between correct and misspecified models for these data and model characteristics. Readers can replicate these results in the Shiny application by uploading a .txt file with the model standardized model estimates to https://dynamicfit.app and setting the sample size to 500 (see Appendix A for more details; .txt files for all the models in this paper are provided on the OSF page associated with this paper).

Table 6
Comparison of RMSEA and CFI fit index distributions when the model is correct and misspecified using a 10% false positive and false positive threshold

Index	Distribution	Percentile	Value
RMSEA	Correct	90	.021
	Misspecified	10	.026
CFI	Correct	10	.987
	Misspecified	90	.983

Conventionally, values satisfying the primary 5% threshold are preferable because the false positive and false negative rates are much smaller, so the Shiny application does not report the 10% threshold cutoffs when the 5% threshold cutoffs are available. Alternatively, if no value can be obtained that satisfies the 5% or 10% threshold, then the fit index distributions may not be sufficiently precise to differentiate between correct and misspecified models in the subspace occupied by the empirical model (this would correspond to the “None” cells of Table 2 and the example shown in Figure 3). This outcome will be more common at smaller sample sizes where the sampling variability of the fit index distributions will tend to be larger.

Summary

Although the fit index values for the empirical model appear to fit well compared to the traditional fixed cutoffs, none of the indices meet the DFI cutoffs that replicate the Hu and Bentler (1999) misspecification applied to the data and model characteristics present in this analysis because misfit is quantified differently in the current model subspace than it is in the model used in Hu and Bentler (1999). Figure 7 compares the fit index distributions for SRMR, RMSEA, and CFI and highlights the differences between the DFI and traditional cutoffs in this example. Note that the traditional cutoffs are completely insensitive to a misspecification consisting of a single omitted cross-loading for this model when the factor reliability is low (*H* = .69) and the distributions of both models fall completely (or nearly-completely) to one side of the traditional cutoff. The DFI approach relocates the cutoff so that it more closely corresponds to the point that demarcates correct and misspecified models to reflect misfit quantification more accurately in this context.

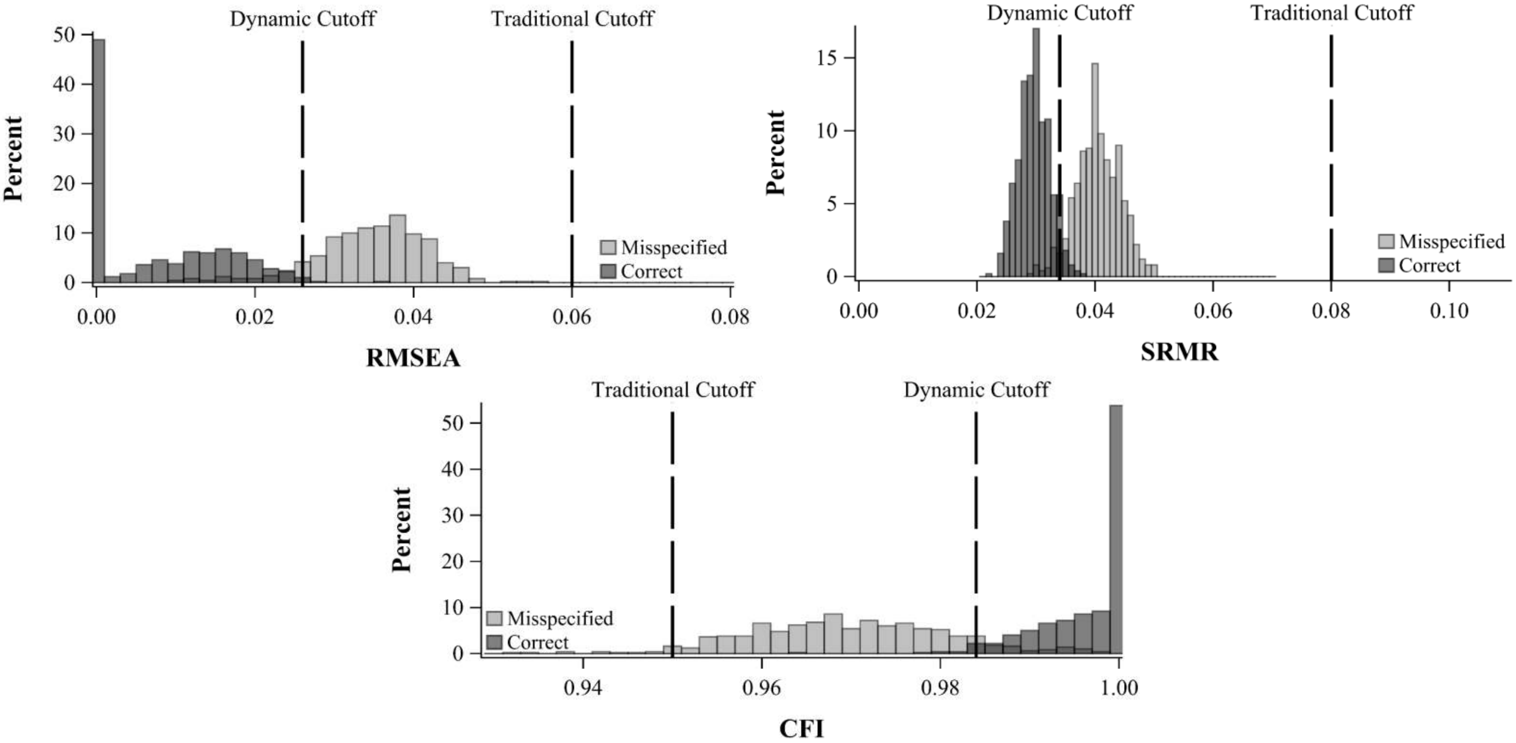


Figure 7. SRMR, RMSEA, and CFI distributions for true and misspecified models with comparison of dynamic cutoffs and traditional fixed cutoffs

Further, note the proximity of the cutoffs from Table 2 for *H* = .69 and the DFI cutoffs here, which we summarize in Table 7 (the SRMR value in Table 4 is from the Complex condition rather than the Simple Condition presented in Table 2). Table 2 compared the empirical model to the actual true model (which was known because the data were simulated) whereas the DFI approach reverse engineered a plausible data generation model featuring a misspecification of the same magnitude as used in Hu and Bentler (1999). The advantage of the DFI approach is that we did not need any knowledge of the true model to arrive at essentially the same cutoffs: we only analyzed a single dataset and used the estimates to reverse engineer a plausible data generation model with the same level of misspecification used in Hu and Bentler (1999) to generalize their approach to these data and model characteristics. Furthermore, both methods also came to the same conclusion that SRMR passed the primary 5% threshold but that both RMSEA and CFI did not and the secondary 10% threshold was required.

Table 7
Comparison of DFI cutoffs from the first replication data

	Cutoff		Threshold	
	Table 2	DFI	Table 2	DFI
SRMR	.036	.035	5%	5%
RMSEA	.025	.026	10%	10%
CFI	.981	.983	10%	10%

Note: the SRMR value reported for Table 2 comes from the Complex condition used by the DFI approach, not the Simple condition used to derive SRMR cutoffs in Hu and Bentler (1999) so the values differs from what was reported in Table 2.

Reproducing Table 2 with Dynamic Fit Cutoffs

To provide further evidence for how our procedure is effective without requiring knowledge of the true model, we reproduce each cell for the $N = 500$ condition in Table 2 using the DFI procedure using the first generated data set for each condition. In Table 2, the true model was known and was fit to the data. In this section, we used the DFI approach to reverse engineer a plausible model with a similar magnitude of misspecification to use as the data generation model. A comparison of the DFI cutoffs and the cutoffs from the full factorial simulation Table 2 are presented in Table 8. The DFI cutoffs closely reproduce the Table 2 cutoffs across conditions where nearly all the cutoffs differ only in the third decimal place. The SRMR cutoff is a little different in the $H = .94$ condition because the DFI approach relies on the maximum allowable cross-loading is this condition rather than unstandardizing any items with a cross-loading as done in the original Hu and Bentler (1999) study. This leads to a slightly different misspecification and affects SRMR more than RMSEA of CFI because SRMR is an absolute index that has no parsimony adjustment or comparison to a baseline model.

Table 8
Comparison of cutoffs from full simulation using true model in Table 2 and dynamic fit cutoffs using data generated from the first replication in the N =500 condition

		SRMR		RMSEA		CFI	
Coefficient *H*	Middle Loading	Table 2 Cutoff	DFI Cutoff	Table 2 Cutoff	DFI Cutoff	Table 2 Cutoff	DFI Cutoff
.42	.35	None	None	None	None	None	None
.57	.45	None	None	None	None	None	None
.69	.55	.036	.035	.025*	.026*	.981*	.983*
.79	.65	.044	.047	.042	.052	.969	.963
.87	.75	.051	.051	.061	.067	.962	.959
.94	.85	.068	.049	.085	.088	.956	.953

Note: SRMR in Table 2 is reported for the Simple misspecification. The dynamic cutoff approach uses the Complex misspecification from Table 2, so the SRMR reported here corresponds to the Complex condition not reported in Table 2. Cells with a "*" indicate the secondary 10% threshold was necessary to determine a cutoff value.

Fit Indices as a Continuum, not an Ad Hoc Hypothesis Test

Fit indices were originally intended to provide continuous information to supplement the χ^2 test to help quantify the degree of misfit (Hu & Bentler, 1999, p.2; Tucker & Lewis, 1973). As noted by Ropovik (2015), "a significant chi-square does not necessarily imply a useless model" and fit indices can assist in differentiating between a model that retains some merit despite misspecifications and a model that is grossly incorrect and should be discarded (Millsap, 2007). However, over time, fit indices have migrated into being used as binary arbiters of fit and function as ad hoc hypothesis tests – much like the χ^2 test – rather than a supplement to quantify the magnitude of misfit (Gomer, Jiang, & Yuan, 2019, p. 372).

To return fit indices more closely to their intended purpose, the simulation design for deriving fit index cutoffs can include multiple conditions for the severity of misspecification.

Rather than having a single cutoff to differentiate “good” versus “bad” models as with traditional fixed cutoffs, varying misspecification severity would provide a semi-continuous set of cutoffs. The result would be similar to bins created by Cohen’s *d* type effect sizes where researchers can better articulate the potential magnitude of misspecification in their model. Unlike Cohen’s *d* effect size bins that are static (e.g., between d = .20 and d =.50 is a typically considered a “medium” effect), DFI cutoff bins adaptively change based on the data and model characteristics. The idea is to allow researchers to be more forthright with their evaluation of fit by not limiting the possible outcome of fit assessment to only “good” or “bad” but rather to allow for more nuance and clarity in model evaluation.

This could also moderate some of the heated debates between proponents of exact fit and approximate fit. Currently, fit indices are used like hypothesis tests rather than as effect sizes, which justifiably elicits consternation among exact fit adherents because treating fit indices this way circumvents principles of null hypothesis significance testing. Treating fit indices more like the effect sizes they were intended to be protects the sanctity of exact fit tests while allowing researchers who are willing to accept some degree of misspecification in their models a way to quantify misfit more accurately. Separating these approaches to model evaluation by giving fit indices the vocabulary and framework it needs to operate as intended gives each perspective space to operate without encroaching on the mechanisms of the other perspective.

More candid evaluations of fit that do not restrict the outcome to only “good” or “bad” will hopefully encourage researchers to embrace a rejected hypothesis that the model fits exactly and use it an opportunity to inspect the nature of the misspecification in their model, such as by examining modification indices (e.g., Saris et al., 2009) or the standardized residual matrix containing the difference between the model-implied and observed covariance elements

(McDonald, 2010; though others do note that such follow-up investigations have caveats and may encourage atheoretic model tinkering, Markland, 2007; MacCallum, Roznowski, & Necowitz, 1992). Rejection of exact fit tests indicates that the model does not exactly reproduce the observed covariance matrix and examination of the standardized residual matrix – for example – can aid in differentiating whether misfit is attributable to a single correctable misspecification or to a scattering of small discrepancies throughout the model, which may suggest the model is a reasonable approximation to reality (McDonald & Ho, 2002, p. 73; Millsap, 2007, p. 879).

For instance, models with high factor reliability (and therefore with high communalities) can often have larger fit indices but no obvious areas of local strain in the standardized residual matrix because power to detect deviations from exact fit are a direct function of the communalities (Browne et al., 2002). Though examination of the standardized residual matrix is often suggested in methodological sources (e.g., Hancock & Mueller, 2011; West et al., 2012), this practice is rarely reported in empirical studies with Ropovik (2015) finding it in just 3% of reviewed articles. Presumably, this is due to fit indices that meet cutoffs being interpreted similar to exact fit rather than being interpreted as lack of exact fit but possibly containing only minor misspecifications.

Varying Severity of Misspecification

Marsh et al., (2004) made strong arguments that the misspecifications used in Hu and Bentler (1999) may not be of interest across different model types and that different levels of misspecification may be of interest in different contexts. For instance, Marsh, Hau, and Grayson (2005) note that it becomes more difficult to meet traditional fixed cutoffs with larger models and that this has encouraged models with few items per factor, similar to how researchers prior to popularization of fit indices might prefer smaller samples to more easily achieve nonsignificant χ^2 tests.

The DFI approach can be generalized to provide varying levels of misspecification. Rather than solely mimicking the single omitted cross-loading to produce the equivalent traditional fixed cutoffs for different data and model characteristics, the simulation can include additional cross-loadings to provide DFI cutoffs for more severe levels of misspecification. In essence, the general procedure outlined previously remains intact but additional misspecification severity conditions are sequentially added with DFI cutoffs being derived for each level. The result is that a model evaluated with fit indices is not merely “good” or “bad” but rather that the level of goodness or badness can be quantified with more nuance.

In the corresponding Shiny application, the default number of misspecification levels is determined by the number of the factors in the model such that the maximum degrees of misspecification is equal to the number of factors minus one. This follows from Hu and Bentler (1999), who simulated two levels of misspecification for a three-factor model (minor and major; see Table 1). The rationale for this choice is that adding multiple cross-loadings to multiple items that load on a single factor in the empirical model does not necessarily worsen misspecification because multiple misspecifications of this type can be absorbed into a single factor covariance, which increases the local strain on this part of the model but does not necessarily affect global strain.

That is, if Item 1 and Item 2 both load on Factor 1 in the empirical model and a data generation is created by adding two additional cross-loadings from Factor 2 to Item 1 and Factor 2 to Item 2, parameter estimates can more easily account for such a localized misspecification and the discrepancy function is little affected by the presence of one versus two omitted cross-loadings (we encountered this situation exactly when testing different ways to allot multiple additional paths in the data generation model in the algorithm). When adding multiple paths to a data generation model, it is useful to spread them throughout the model to avoid this hyper-localized misspecification, which is why the number of levels of misspecification is based on the number of factors to ensure that the discrepancy function will actually increase as more paths are added to the data generation model. The algorithm to create the data generation model continues to add paths equal to each successive loading in ascending order of magnitude but does so conditionally to not repeat adding cross-loadings that affect the same pair of factors. This is demonstrated in the next section with the same simulated example used previously.

Simulated Data Example

We demonstrate using the same data from Replication 1 from the Complex model, N = 500, H = 0.69 condition reported in Table 3. As discussed previously, the data generation model for one additional cross-loading that mimics the minor misspecification in Hu and Bentler (1999) would consist of an additional standardized cross-loading of .445 from Factor 2 to Item 1. We can extend beyond the misspecification in Hu and Bentler (1999) and test additional levels of misspecification to get a broader sense of the sensitivity of fit indices to different types of misspecifications.

To add a second additional path to the data generation model, Step 2 of the DFI algorithm is repeated with Item 2 removed from consideration because it has already been used. The next item selected is Item 13 whose standardized loading in the empirical model was .479. This loading (or the maximum allowable loading described in Equation 1) is also added to the data generation model so that there are two-cross-loadings in the data generation model that did not appear in the empirical model (similar to Hu and Bentler's "major" misspecification condition).

Choosing the factor associated with this second cross-loading is more nuanced with multiple additional cross-loadings present in the data generation model. As before, the factor on which the item did not originally load with the highest factor reliability is preferred. If selecting this factor would repeat the same factor pairing as a previously added path, then the factor with the next highest factor reliability is selected.

In this example, Item 13 loads on Factor 3, Factor 2 has the highest factor reliability, and the previously added cross-loading was between Factor 2 and an item that loaded on Factor 1. The DFI algorithm will note that there is also an additional misspecification already in the data generation model between Factor 1 and Factor 2 and will not allow another misspecification to be added between these factors. However, misspecifications can be added between either Factor 1 and Factor 3 or between Factor 2 and Factor 3. Factor 2 has the highest factor reliability, so the .479 loading is therefore added from Factor 2 to Item 13 because this factor pairing is still available to receive a misspecification in the data generation model.

If the item selected for the second additional cross-loading also belonged to Factor 1, the cross-loading would have been assigned to Factor 3 to avoid including multiple misspecifications between Factor 2 to an item belonging to Factor 1 in the data generation model. For a 3-factor model, two misspecifications are tested by default in the Shiny application (similar to Hu and Bentler’s minor and major misspecification conditions) and DFI cutoffs corresponding to each level of misspecification are provided. For this example dataset, these cutoffs are presented in Table 9.

Table 9
DFI cutoffs for two levels of misspecification and the magnitude of the standardized cross-loading added to create at each level of misspecification.

Additional Paths In Data Generation Model	Magnitude of Omitted Loading	SRMR	RMSEA	CFI
1	.445	.035	.026	.983
2	.479	.051	.049	.948

In Table 9, the first row corresponds to the DFI cutoffs for a misspecification of one omitted cross-loading with a standardized magnitude of .445 (these were also presented in the previous section where we outlined the details of the DFI algorithm). The second row corresponds to the DFI cutoffs for a misspecification of two omitted cross-loadings with standardized magnitudes of .445 and .479. Rather than having a single set of cutoffs that obliges researchers into a “good” versus “bad” determination, there are now multiple bins in which the model evaluation can fit:

1. The model reproduces the observed covariance matrix exactly (non-significant χ^2 test).
2. The model does not fit exactly, but the amount of misfit is consistent with or less than an omitted standardized cross-loading equal to .445 (fit index values better than the DFI cutoffs in the first row).
3. The model does not fit exactly and the amount of misfit is consistent with somewhere between one omitted standardized cross-loadings equal to .445 and two omitted standardized loadings equal to .445 and.479 (fit index values better than the DFI cutoffs in the second row but worse than the DFI cutoffs in the first row).

4. The model does not fit exactly and the amount of misspecification exceeds two omitted standardized cross-loadings equal to .445 and .479 (fit index values worse than the DFI cutoffs in the second row).

For this particular model, the fit index values were SRMR = .041, RMSEA = .033, and CFI = .970, so the interpretation would fall into Category 3 – the fit index values are worse than the DFI cutoffs in the first row but better than the DFI cutoffs in the second row. This would indicate that the model does have some misspecifications, but the magnitude of those misspecifications appears to be moderate and the model may retain some merit (as we discuss in the next section, just like effect sizes used in other modeling frameworks, it is up to the researcher to justify the level of misspecification that is acceptable). For models with more factors, more levels of misspecification are output by default in the application to address the points in Marsh et al. (2004) and Marsh et al. (2005) that the definition of 'minor misspecification' is not fixed, can differ across model types and scales with model size.

How Misspecified is "Too Misspecified"?

Where to draw the line for which level of misspecification is "too much" is up to interpretation. This is no different from effect sizes in other analyses – a treatment effect may be statistically significant, but the real question is whether the impact is sufficiently high to warrant a policy change or intervention. To some researchers, "sufficiently high" might correspond to a Cohen's *d* of .25 but to others it might correspond to a Cohen's *d* of .50. Further, the traditional effect size cutoffs provided by Cohen (1988) have recently been noted to be more fluid than originally thought and vacillate across contexts (e.g., Correll et al., 2020; Kraft, 2020), much like fit indices. To be clear, such arguments are not about whether the effect is different from 0 as tested by null hypotheses, it is about the point at which difference becomes practically meaningful.

The same principle holds for model evaluation with CFA – there is no arguing about whether the model fits exactly or not from an inferential standpoint. If the χ^2 test is significant, then the model does not reproduce the observed covariance matrix and there is a misspecification either in the structural relations in the model or in the distributional assumptions of the model. Fit indices should not be used as an ad hoc hypothesis test to replace a null hypothesis significance test whose outcome is unfavorable. This is analogous to the fact that a small Cohen's *d* does not invalidate a significant treatment effect because the two metrics are testing different things. Conversely, a significant χ^2 test does not imply that a model is useless just as a significant treatment effect does not mean a policy change is necessarily warranted if the practical difference between groups is negligible.

As in other statistical models where there is a complementary interplay of effect sizes and significance tests, exact fit and approximate fit should not be competing approaches in CFA but rather are complementary approaches that together can provide more holistic evaluations of models. Instead, fit indices should be used to quantify the degree of misspecification and to present evidence for whether the impact of the misspecification is sufficiently large to invalidate the model and its conclusions. As with policy decisions using effect sizes, there will be differing opinions about what magnitude misspecification is too big in a particular context, but this is an inherent quality of effect size measures (and fit indices) that should be embraced rather than overlooked by those who use fit indices to evaluate their models.

Furthermore, the guidelines presented by AERA, NCME, and APA (2014) describe five equally important sources of validity evidence for the use of psychological measurements.

Researchers heavily emphasize quantitative evidence of validity based on internal structure – often via CFA and model fit measures – to the detriment of the others. Validity arguments via internal structure can be important, but statistical models themselves are atheoretical and psychometric interpretations are only valid insofar as there is a strong theory behind a model. Internal structure is only one piece of the puzzle and can (and should) be supplemented with other sources of validity information to present a more holistic case for validity that can extend behind model fit measures. In other words, validity is always an argument that can be supported by – but not exclusively based upon – quantitative information like fit indices.

Empirical Example

This section provides an empirical example to show how fixed cutoffs can impact conclusions from empirical studies that do not necessarily share model characteristics with the data generation model in Hu and Bentler (1999). *Psychological Assessment* focuses on scale validation where evaluating fit of CFA models is often a primary goal and was a natural location from which to draw an example. We discuss a study from Kearns et al. (2018) which involved scale validation using standard CFA models along with traditional fixed cutoffs to inform decisions about adequacy of model fit.

To be clear, this study followed currently accepted protocols and we are not faulting their methodology in any way. In fact, the study was selected because it was so thorough in reporting its results that all relevant information needed to recreate and evaluate the analysis was present and we commend the authors' transparency. Accordingly, we highlight this study as an example of limitations of existing recommended practices, even when followed exactly. We give a short description of the study and its reported results prior to applying the DFI approach to the reported model and discussing discrepancies between traditional and DFI cutoffs.

Kearns et al. (2018)

Kearns et al. (2018) was interested in validating the Brief-Caffeine Expectancy Questionnaire (B-CaffEQ) and collected responses on 47 candidate items from 975 people regarding their caffeine consumption. Half of the participants (*N* = 488) were used in an exploratory analysis, which narrowed the scale down to 21 items loading on 7 factors. The final CFA featured 20 items on 7 factors for the remaining 487 people (one item was removed between the exploratory and confirmatory analysis); the first 6 factors each had 3 items and the last factor had 2 items. All factors were allowed to covary with all other factors. The standardized factor loadings were quite high, ranging from 0.70 to 0.92, with an average of 0.84. Figure 8 shows the path diagram with standardized loadings for the CFA model (factor correlations are not shown to improve readability of the figure). The fit of the CFA was reported to be reasonable relative to traditional cutoffs where RMSEA = 0.066 [90% CI = (0.060, 0.072)], SRMR = 0.040, and CFI = 0.953. Reporting of model fit was done thoughtfully and the authors were forthright in noting that the RMSEA was slightly above the traditional cutoff.

The factor reliability is high in this model, ranging from 0.76 to 0.92 across the seven factors, which falls close to the values used to derive the traditional cutoffs in Hu and Bentler (1999). The number of factors in the model was higher than the number used to derive traditional cutoffs (7 versus 3) while the number of items per factor was lower (2 or 3 versus 5). Given these differences, the model subspace used when deriving the traditional cutoffs is not adjacent to this empirical model and the cutoffs would be unlikely to generalize.

The DFI cutoffs corresponding to a misspecification of single omitted cross-loading for these model and data characteristics that mimic the procedure used to derive the traditional cutoffs are:

- RMSEA ≤ 0.038
- SRMR ≤ 0.037
- CFI ≥ 0.986

None of the fit indices from this empirical model are better than these cutoffs, indicating that the cumulative magnitude of misspecification in this model is greater than that contained by a single omitted cross-loading.

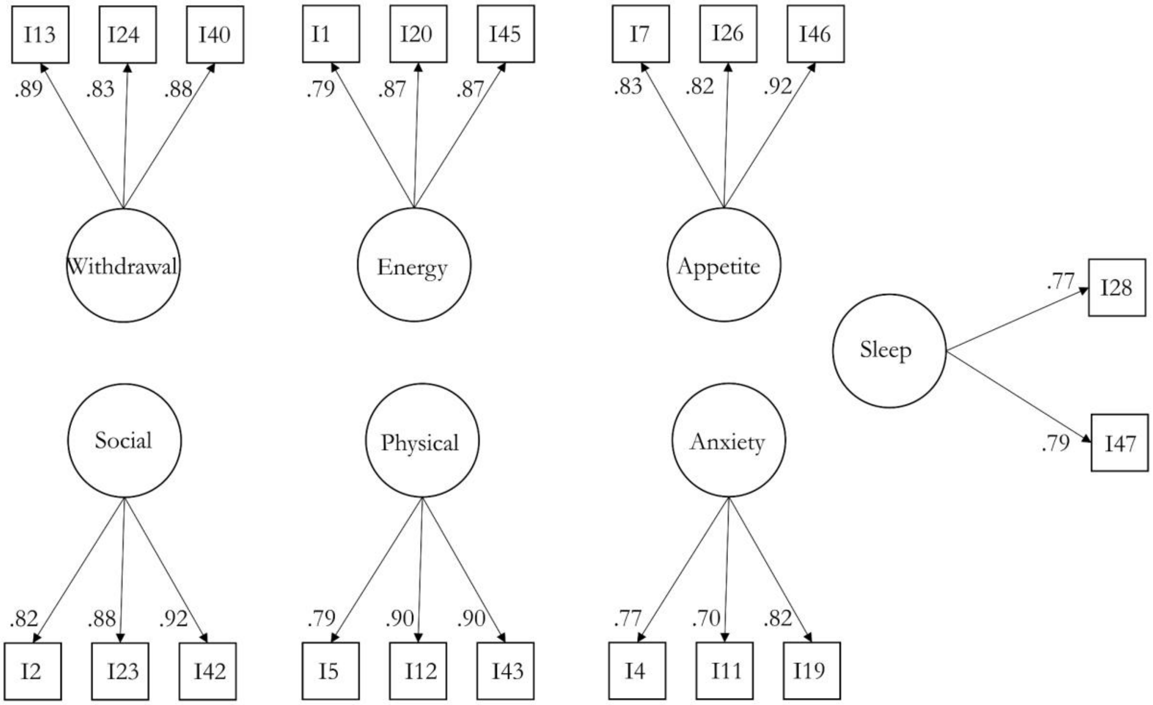


Figure 8. Path diagram for CFA model in Kearns et al. (2018) with standardized loadings. All factors have a variance of 1 and all factor covary with each other, which are not shown

Because commonly used fit indices track global fit, the more unique elements present in the model-implied covariance matrix, the easier it is for the impairment caused by a single misspecification to be washed out after being combined with or averaging over the other elements (see e.g., Table 3 of Shi, Lee, & Maydeu-Olivares, 2019).[7] Additionally, with many factors and few items per factor, one omitted cross-loading will impact few other elements of the

[7] For example, we reran the $H = .94$ condition for the simulation presented in Table 2 but added 5 more items with loadings between .80 and .90 and the Level-1 cutoffs corresponding to a single-omitted cross-loading became stricter (RMSEA ≤ 0.064, CFI ≥ 0.970, SRMR ≤ 0.037).

model-implied covariance matrix (i.e., the contagion of the misspecification is more isolated with few factors and many items per factor). As model size expands, perhaps a model with only a single misspecification – upon which the traditional fixed cutoffs are based – can still make a contribution to the literature (e.g., Marsh et al., 2005). Therefore, it would be useful to obtain cutoffs based on additional misspecification(s) that are less isolated to better quantify what level of misspecification is actually present in the model rather than concluding that, if the model does not meet the cutoffs for a single omitted cross-loading, then the model fit is necessarily poor.

As described above, additional levels of misspecification can be tested to better quantify the magnitude of misspecification that may be present to better evaluate fit and extend beyond simple conclusions of "good" or "bad" fit. Using the generalization of the DFI algorithm described previously, our Shiny application by default will test 6 levels of misspecification for a model with 7 factors. The misspecification and the cutoffs corresponding to these different levels are provided in Table 10.

The first row of Table 10 is the one omitted cross-loading misspecification that corresponds to the approach used to derive the traditional cutoffs. Each successive row adds another cross-loading to the data generation model such that the omitted cross-loadings are cumulative. That is, the second row of Table 10 provides DFI cutoffs for a data generation model that omits a standardized cross-loading of .446 and a standardized cross-loading of .234 in the same model. The magnitudes of the omitted cross-loadings are small for this model because the standardized loadings in the model are high and items cannot accommodate larger standardized cross-loadings while retaining a nonnegative standardized residual variances because there is little unexplained variance remaining in most items. The magnitude of the omitted cross-loadings is also not strictly monotonic because the overall misspecification is also affected by the magnitude of factor correlations in addition to cross-loading magnitude.

Table 10
DFIs cutoffs for six levels of misspecification and the magnitude of the new cross-loading added to create each subsequent misspecification.

Additional Paths In Data Generation Model	Magnitude of Omitted Loading	SRMR	RMSEA	CFI
1	0.446	.038	.038	.986
2	0.234	.039	.042	.984
3	0.272	**.044**	.053	.977
4	0.255	.049	.062	.970
5	0.239	.049	**.072**	.964
6	0.279	.057	.088	**.949**

Note: Bold entries indicate the level for each index at which the fit index from the Kearns et al. (2018) model is better than the DFI cutoff

From this information, we can see that the SRMR in this model (.040) is consistent with an omission of three standardized cross-loadings with magnitudes of .446, .234, and .272. Similarly, the RMSEA (.066) is consistent with an omission of five standardized cross-loadings with magnitudes of .446, .234,.272, .255, and .239 and CFI (.953) is consistent an omission of six standardized cross-loadings with magnitudes of 446, .234, .272, .255, .239, and .279. As a reminder, this does not mean that the empirical model necessarily has omitted cross-loadings. The omitted cross-loadings included in the data generation model are merely one representative misspecification to help better understand the scaling of fit indices for these data and model characteristics. This would be interpreted as the cumulative misspecification in the empirical model being on par with misfit that would be caused by a particular number of cross-loadings of a particular magnitude. The actual pattern or distribution of misfit throughout the empirical model could look very different what is generated in the DFI simulations. In other words, the

DFI simulations are a tool to guide interpretation of fit indices but do not imply anything about the nature of misspecifications in the empirical model.

The ultimate question is, does this model fit? Strictly speaking, it does not because the χ^2 test is significant ($\chi^2(149) = 469.29, p < .01$, as reported in the original paper) which indicates a misspecification is present. However, as noted by Ropovik (2015), a significant χ^2 test does not imply a model is useless and Millsap (2007) notes that fit indices can be used to differentiate models with merit versus models that are grossly incorrect. Here, the model is not consistent with a misspecification of the size used to derive the traditional cutoffs (or smaller) when updated to the current model characteristics. However, the model may retain some merit depending on how this information in contextualized. In our opinion, even though the misfit is consistent with several omitted cross-loadings, the magnitudes associated with these hypothetically omitted standardized loadings is mostly in the low .20s and misspecification of this magnitude would likely not be severe enough to warrant dismissing the model.

Nonetheless, model evaluation using fit indices is not wholly objective by design and involves some argumentation to support whether the model misspecification is sufficiently small such that its use could still be warranted despite deviation from exact fit. The DFI approach provides a framework where validity arguments can be built around the types of misspecifications that are consistent with fit index values (possibly including other sources of validity beyond internal structure) and what types of misspecifications researchers deem acceptable rather than imprudently overgeneralizing fixed cutoffs. Incidentally, this fits neatly within the mainstream arguments-based approach to instrument validation in education and psychology whereby researchers are encouraged to present any evidence needed to justify the intended uses and interpretations of assessment scores (Kane, 2013).

One-Factor Models

Despite the wealth of interest in unidimensionality and one-factor models, the misspecifications in Hu and Bentler (1999) do not translate to one-factor contexts. The covariance misspecification in their Simple misspecification condition featured an omitted factor covariance and their Complex misspecification condition featured omitted cross-loadings on other factors. However, in the one-factor context, there are no additional factors with which the sole factor can covary nor are there additional factors upon which items can cross-load. Despite the frequency with which traditional fixed cutoffs are applied with one-factor models, the misspecifications to which that traditional fixed cutoffs are designed to be sensitive do not exist in one-factor models. Though some research has delved into issues pertaining to fit indices and associated cutoffs with one-factor models (e.g., Shi et al., 2019; Shi & Maydeu-Olivares, 2020), this area remains understudied with limited guidance for practice compared to the sizeable literature on multifactor models.

Unidimensionality is a common question in scale development and validation and there has been recent renewed interest in one-factor models for assessing psychometric properties of or as alternatives to sums or averages of item scores (e.g., Edwards & Wirth, 2009; Fried et al., 2016; Fried & Nesse, 2015; Kuhfeld & Soland, 2020; McNeish & Wolf, 2020; Slof-Op't Landt et al., 2009). To meet the needs of researchers seeking to evaluate evidence of fit of their one-factor models, we have created a separate DFI algorithm that deviates from the multifactor approach of Hu and Bentler (1999) to better address the issues present in one-factor models. Due to differences between one-factor and multifactor models, we felt it was appropriate to separate Shiny applications for these different types of models.

One-Factor DFI Algorithm

A common interest with one-factor models is in evaluating whether a set of items are reasonably unidimensional or whether there may be multiple factors present. One proxy for inducing multidimensionality with one-factor models is to add residual correlations between individual items, which has been used in previous methodological studies in this area (Shi et al., 2019; Shi & Maydeu-Olivares, 2020). Previous simulation results have suggested that the magnitude of factor loadings was less salient for models with omitted residual correlations but that the number of items was far more influential in the context of one-factor models (e.g., Table 3 of Shi & Maydeu-Olivares, 2020). Given these findings and the omission of one-factor models for the simulation used to derive the traditional cutoffs, we retain the simulation-based framework presented in previous section but break from following Hu and Bentler's approach for computing DFI cutoffs for one-factor models.

Just as we featured multiple levels of misspecification for multifactor models, we also feature multiple levels of misspecification for one-factor models. Residual correlations are a more localized misspecification, so cutoffs are affected by the number of items in a one-factor model. For example, the presence of one omitted residual correlation in a one-factor model with 30 items will be much less impactful than one omitted residual correlation in a one-factor model with 6 items. Rather than provide the DFI cutoff for a single omitted residual correlation for all models, we base the DFI cutoffs on the *proportion* of items in the data generation model with residual correlations. Whereas a single omitted residual correlation may be worrisome in a model with 6 items, the threshold for concern in a 30-item model would likely be much larger. Consequently, the number of residual correlations added will vary depending on model size and will not necessarily correspond to a generalization of the traditional cutoffs (as we strived for above in multifactor models where the prior literature is much deeper).

Unlike the multifactor DFI algorithm that varied the number of misspecification levels depending on the number of factors (which was done to maintain a connection to Hu and Bentler, 1999), the one-factor DFI algorithm takes a more standardized approach. Each one-factor model fit with our Shiny application will test three levels of misspecification for one-factor models regardless of model size. Each of these levels is proportional to the number of items in the one-factor model that do not already have another residual covariance with another item – denoted as I – so that the levels of misspecification have constant meaning across all models to improve interpretation of the DFI cutoffs[8]. Mathematically, the number of residual correlations added in the data generation model for each level of misspecification is

Level-1: $\lfloor \lfloor I/2 \rfloor /3 + 0.50 \rfloor$

Level-2: $\lfloor 2(\lfloor I/2 \rfloor /3) \rfloor$

Level-3: $\lfloor I/2 \rfloor$

Where $\lfloor \quad \rfloor$ is the floor function that rounds all values down to the nearest integer. The three levels of misspecifications are ascending such that Level-3 is the most severe. The Level-1 misspecification implies that about one-third of the items in the data generation model have residual correlations, the Level-2 misspecification implies that about two-thirds of items in the data generation model have residual correlations, and the Level-3 misspecification implies that each item has one residual correlation with exactly one other item if there are an even number of items. If the number of items is odd, all but one item will have a residual correlation with exactly one other item in the Level-3 misspecification. When $I/2$ is divisible by 3, the proportion will be exactly one-third for Level-1 and exactly two-thirds for Level-2. Models with only four items

[8] Items that already have an error covariance in the model are ineligible for covariance misspecifications. Under local independence where there are no residual covariances, *I* will be equal to the number of items in the model.

will only have one level (a Level-2 misspecification with 4 items would be just-identified) and models with only five available items will only have two levels (three covariances cannot be added without reusing items).

When creating these models, the magnitude of residual correlations added to the data generation model are not limited in the same way as cross-loadings, so we do not base the magnitude of additional paths in the data generation model off the estimates in the empirical model. Rather, each residual correlation added in the data generation model is equal to 0.30. Residual correlations are added in the data generation starting with the items that have the lowest standardized loadings and additional residual correlations are added to item pairs in ascending order of their standardized loadings with no item receiving more than one residual correlation. That is, assuming no residual correlations in the empirical model for clarity, the first residual correlation in the data generation model would be added between the items with lowest and second lowest standardized loadings, the second residual correlation between the third and fourth lowest loadings, and third residual correlation between the fifth and sixth lowest loadings and so on. This pattern was chosen to provide a replicable series of steps so that the results are identical if fit repeatedly to the same model rather than for any particular methodological reason. An example of the one-factor algorithm in provided in the next subsection.

One-Factor Model Example

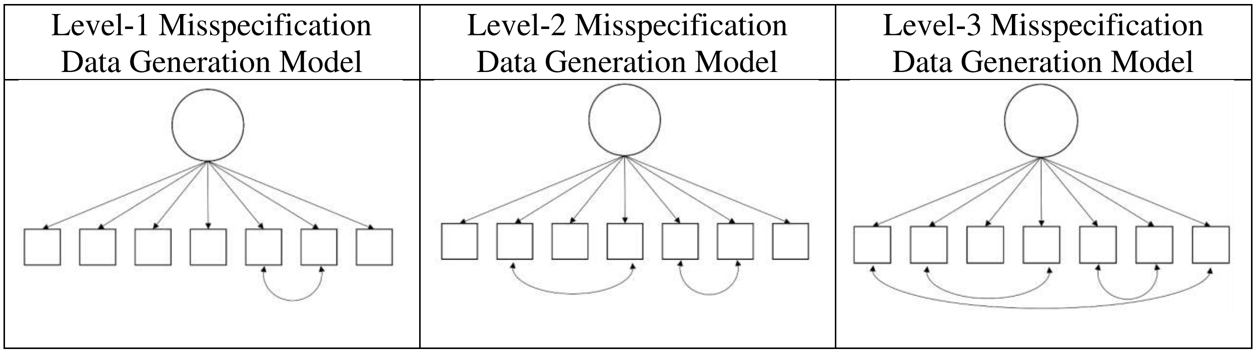
Example data come from the SAQ-7 scale data used in Field (2005) and that appear as an example of CFA on the popular UCLA Institute for Digital Research and Education (IDRE) Statistical Consulting website (https://stats.idre.ucla.edu/spss/seminars/introduction-to-factor-analysis/). The data contain 2,571 responses to seven items about anxiety surrounding learning to use the SPSS statistical software and the interest of the analysis is in determining whether this scale is unidimensional and measuring a single construct. The data were generated and some of the items are facetiously worded to engage introductory students to which the textbook is aimed; however, the data are openly available from several sources and are able to demonstrate implementation of our proposed approach to one-factor models.

We fit a one-factor model in M*plus* version 8.3 using maximum likelihood estimation and the standardized loadings and fit criteria are provided in Table 11. The model does not fit exactly given that the χ^2 test is significant ($\chi^2\left(14\right)=376.32, p<.01$). The sample size was quite large and we had high power to detect small misspecifications, so consulting fit indices may help determine whether the model may still have some utility or whether it is grossly misspecified. The RMSEA is .10 is above the traditional .06 cutoff; however, the degrees of freedom are small for this model (as they are for many one-factor models) and RMSEA may be inflated relative to traditional cutoffs (e.g., Kenny et al., 2015). The SRMR of .049 is below the traditional cutoff and the CFI of .906 is below the traditional cutoff but above the heuristic .90 cutoff that is sometimes considered acceptable for incremental indices. Based on traditional cutoffs, there may be some evidence that the model misspecifications are small enough to warrant considering the model further. However, it is unclear how sensitive the traditional cutoffs are to misspecifications in one-factor models because traditional cutoffs were derived from multifactor models. Furthermore, the model has more items per factor and lower factor reliability (H = .79) than the conditions used to derive traditional cutoffs.

Table 11
Standardized loadings and fit from one-factor CFA testing unidimensionality of the SAQ-7 scale

Item	Std. Loading	Fit	
1	.590	RMSEA	.100
2	-.553	RMSEA 90% CI	[.092, .109]
3	.672	SRMR	.049
4	.576	CFI	.906
5	.491	$\chi^2(14)$	376.32
6	.497		p <.01
7	.648		

To quantify the degree of misspecification more specifically for this one-factor model, we calculated the DFI cutoffs. The model has 7 items, meaning that the Level-1, Level-2, and Level-3 misspecifications will consist of one, two, and three additional 0.30 residual correlations in the data generation model, respectively. Figure 9 shows the path diagram for the data generation model at each level of misspecification. The data generation for a Level-1 misspecification includes an additional .30 residual correlation between the items with the two lowest standardized loadings (Item 5 and Item 6). The Level-2 misspecification adds another .30 residual correlation to the items with the next two lowest standardized loadings (Item 2 and Item 4). The Level-3 misspecification then adds a third .30 residual correlation to the items with the next two lowest standardized loadings (Item 1 and Item 7). Because there are an odd number of items, the item with the highest standardized loading (Item 3) does not receive a residual correlation in any of the data generation models.



Level-1 Misspecification Data Generation Model

Level-2 Misspecification Data Generation Model

Level-3 Misspecification Data Generation Model

Figure 9. Path diagrams for the data generation models used to derive Level-1, Level-2, and Level-3 misspecification DFI cutoffs for testing unidimensionality of the SAQ-7

Table 12 shows the DFI cutoffs for these model and data characteristics for the three standardized levels of misspecification. From the DFI cutoffs, we can see that the model misspecification is not consistent with a Level-1 or Level-2 misspecification as the empirical model fit index values are worse that these cutoffs for all three indices. The model fit indices are consistent with (or slightly better than) a Level-3 misspecification, indicating that the cumulative model misspecification is on par with every item but one having an omitted 0.30 residual correlation with another item. Similar to Cohen's *d* bins, Level-1 is intended to be the cutoff for a 'small' misspecification, Level-2 the cutoff for a 'medium' misspecification, and Level-3 the cutoff for a 'large' misspecification. So, the unidimensionality of the model is questionable given the proximity of the empirical model fit index values to the Level-3 DFI cutoffs.

Table 12
DFI cutoffs for testing unidimensional of SAQ-7 Scale

Misspecification Level	SRMR	RMSEA	CFI
1	.033	.064	.962
2	.044	.088	.927
3	.052	.107	.903

As a reminder, this does not mean that the model necessarily has omitted residual correlations – the omitted residual correlations included in the data generation model are merely one representative misspecification to help better understand the scaling of fit indices for these data and model characteristics. This would be interpreted as the cumulative misspecification in the model being consistent with three omitted residual correlations. However, the distribution and pattern of misfit is not necessarily the same as the representative misspecification used in the DFI simulations. Given the relatively poor fit of the model, discussion of possible modifications is considered in the next subsection.

Model Modification

DFI provides information about severity of misfit, not just whether fit is broadly classified as good or bad. This allows us to see that the model in the previous section does not just fit poorly, but that the fit indices are consistent with a rather large misspecification. In such situations, two perspectives emerge on how to proceed depending on where the analysis falls along the continuum of confirmatory to exploratory. If the model is confirmatory and altering the model is objectionable, researchers could ascertain whether they are comfortable with a level of misspecification just slightly below the Level-3 CFI cutoff. If the model is more exploratory, researchers can follow up possibly deficient fit to diagnose why the model fit is poor and whether there are clear ways in which fit would be improved (such modifications should, of course, be reported). As an analogy to analysis of variance, this could be considered as an omnibus-type test indicating that there is misfit somewhere and inspections of diagnostics would serve to locate the source of the misspecification.

As noted previously, the standardized residual matrix is a helpful source to locate such misfit. This matrix for the current model is,

	I1	I2	I3	I4	I5	I6	I7
I1	0						
I2	−.01	0					
I3	.04	−.01	0				
I4	.06	.01	.01	0			
I5	.04	.01	.02	−.01	0		
I6	−.08	.05	−.06	−.03	−.02	0	
I7	−.08	−.02	−.03	−.04	−.02	.19	0

Entries in this matrix indicate the difference between the model-implied correlation and the observed correlation for observed variables. Values of 0 indicate the model-implied correlation exactly reproduces the observed correlation, negative values indicate the model-implied correlation is larger than the observed correlation, and positive values indicate the model-implied correlation is smaller than the observed correlation. Most of the standardized residual elements are near 0 except the correlation between Item 6 and Item 7, which was off by 0.19 (observed correlation = 0.51; model-implied correlation = 0.32). This indicates that there is likely an unmodeled relation between these two items, which is corroborated by the modification indices which suggest that adding a residual covariance between Item 6 and Item 7 would reduce the χ^2 statistic by 322.78. Item 6 and Item 7 are both related to anxiety about computer literacy rather than about mathematical or statistical reasoning, so this may be responsible for the additional relationship above and beyond the latent variable.

The standardized estimates for the modified model with a residual correlation between Item 6 and Item 7 are shown in Table 13. The model still does not fit exactly ($\chi^2(13) = 66.77, p < .01$) but the fit indices are much improved (RMSEA = .040, SRMR = .021, CFI = .986). The DFI cutoffs can then be reassessed using the modified model.[9] Because the

[9] In this example, we recalculate the DFI cutoffs after modifying the model. We could see arguments against this practice such that the DFI cutoffs from the original model should be used instead. We could not determine a definitive answer to which should be preferred and additional consideration of this topic would be needed to determine which set of cutoffs is most appropriate. In general, for modifications that only involve different paths and

modified model has one existing residual correlation, there are only 5 possible items to which a residual correlation could be added in the data generation model. Therefore, only two levels of misspecification can be tested rather than the three levels that were tested in the original model. The DFI cutoffs for the modified model are shown in Table 14.

Table 13
Standardized loadings and fit for the modified SAQ-7 model featuring a residual correlation between Item 6 and Item 7

Item	Std. Est.	Fit	
1	.619	RMSEA	.040
2	-.558	RMSEA 90% CI	[.031, .050]
3	.694	SRMR	.021
4	.588	CFI	.986
5	.498	$\chi^2(13)$	66.77
6	.403		*p* <.01
7	.582		
Corr (6, 7)	.375		

Table 14
DFI cutoffs for testing modified model for SAQ-7 scale

Misspecification Level	SRMR	RMSEA	CFI
1	.032	.061	.967
2	.040	.082	.948
3	---	---	---

The level of misspecification present in the modified model appears to be relatively low and is consistent with one omitted residual correlation or less severe (i.e., the fit indices of the empirical model are better than the Level 1 DFI cutoffs). Furthermore, inspecting the standardized residual matrix reveals that there is no obvious source of misfit in the modified model,

that do not change the number of variables, there should not be too great a difference between the sets of cutoffs because the model characteristics should be quite close. This pattern is observed in this example whereby the Level-1 and Level-2 cutoffs in Tables 12 and 14 differ only in the third decimal place in all but one case. For more fundamental modifications that involve removing items or changing the number of factors such that there are differences in the model characteristics, it would seem more prudent to consider recalculating cutoffs such that they based on the modified model characteristics.

	I1	I2	I3	I4	I5	I6	I7
I1	0						
I2	.01	0					
I3	.01	.01	0				
I4	.03	.02	–.01	0			
I5	.02	.02	.00	–.02	0		
I6	–.03	.00	.01	.00	.02	0	
I7	–.05	–.07	.00	–.02	.01	.00	0

Also note that the modified model could equivalently be fit as a two-factor model with Factor 1 loading on Item 1 through Item 5 (math and statistics related items) and Factor 2 loading on Item 6 and 7 (computer related items). If the model were fit this way, the loglikelihood, χ^2 statistic, fit indices, and model-implied covariance matrices are all identical. We can run the modified two-factor model estimates through the Shiny application to get the DFI cutoffs when treating the modified model as a multifactor model. The DFI cutoffs from treating the model as multifactor rather than as one-factor with residual covariances is shown in Table 15 (recall that two-factor models can only test one level of misspecification in the application). The DFI cutoffs in Table 15 are not identical to those in Table 14 because the misspecification in the data generation model is slightly different (based on a cross-loading rather than residual correlation), but they are relatively close and lead to the same overall conclusion that the model fit is consistent with a Level-1 misspecification (or lower).

Table 15
DFI cutoffs from treating the modified model as a two-factor model rather than a one-factor model with correlated residuals

Misspecification Level	SRMR	RMSEA	CFI
1	.037	.076	.969

The equivalence of fit between the one-factor model with an error covariance and the two-factor model and the ease at which a researcher can consult modification indices and change their model further emphasizes the importance of establishing strong, evidence-based theory a priori. When modeling behavioral data, there will often be modifications that can improve fit because people are heterogeneous and psychological constructs are complex. We encourage researchers to remember the motivation for using a CFA model in the first place: to present evidence of validity. This validity is often based on the internal structure although other types of evidence of validity exist (AERA, APA, NCME, 2014). Severe misfit can arise for any number of reasons and a decision of how to best handle that (e.g., defend the scale as is, modify theory, or modify items) is likely best made with the help of other types of validity evidence.

Discussion

Global model fit indices are often treated as a fundamental source of validity evidence for psychological assessments, necessitating appropriate implementation. Many methodological studies have shown that the meaning of these fit indices change depending on data and model characteristics. The implication is that the threshold needed to achieve "good fit" with fixed cutoffs is inconsistent and models with certain characteristics (e.g., low factor reliability) have an arbitrary advantage relative to Hu and Bentler's traditional fixed cutoffs. To create more flexible and equitable cutoffs, previous recommendations have suggested simulation-based techniques that are custom tailored to the unique characteristics of the model being evaluated to derive custom cutoffs. Though this recommendation is insightful for circumventing undesirable properties of fixed cutoffs, it has failed to gain traction in empirical studies, presumably because the requisite quantitative training required to implement simulation-based techniques exceeds the quantitative training possessed by many empirical researchers. Alternatively, even if researchers have the requisite quantitative training, the process of programming a unique simulation for each model can be time intensive.

Our DFI approach aims to address these issues to improve the precision of fit index cutoffs in a user-friendly manner, making bespoke evaluation of psychometric models and validity evidence widely accessible. Our Shiny application creates simulation code based on model results and internally executes the simulation so that researchers without knowledge of simulation techniques can exploit modern computational resources and those with knowledge of simulation techniques can streamline the process. This overcomes barriers present with earlier proposals for simulation techniques to custom tailor cutoffs. The result is that DFI cutoffs are fully tailored to the data and model characteristics without demanding that users possess the ability to manually program their own simulations.

The DFI approach aims to provide a user-friendly alternative to the common practice of using fit indices as ad hoc hypothesis tests and to revert their use to effect sizes that quantify the magnitude of misspecification in the model. In this way, the DFI approach is not just about revising cutoffs but rather about changing how researchers interact with and use fit indices. Providing a set of cutoffs rather than a single binary decision point allows researchers to more openly acknowledge that the model does not fit exactly and contextualize potential misspecifications, similar to the valuable interplay of statistical and practical significance enjoyed in other statistical models. This hopefully will serve to dampen the rift between those who ardently support the χ^2 test and those who prefer fit indices. Rather than competing against one another in a methodological Battle Royale where only one method can emerge victorious, the DFI approach reinforces that the χ^2 test and fit indices are different metrics with different goals and can be used to complement – not replace – one another.

DFI vs. Equivalence Testing

Readers familiar with recent developments may note that equivalence testing (Yuan et al., 2016; Marcoulides & Yuan, 2018) has a similar goal to the proposed DFI method. The method proposes switching from a null hypothesis framework typically used for model fit evaluation to an equivalence testing framework whereby the hypothesis being tested is whether the discrepancy function is larger than a particular value – which can be defined by fit index values – rather than testing whether the model-implied covariance matrix is exactly equal to the observed covariance matrix. As a byproduct of this switch, Yuan et al. (2016) note the previous literature on how fit index values can carry different meaning in different contexts and provide a method by which traditional fit index cutoffs can be adjusted so that they change in accordance with data and model characteristics so that the level of misspecification used in the equivalence testing null hypothesis is equitable.

Rather than customized simulation studies, equivalence testing provides a correction factor based on results from a best subset regression involving interactions and polynomial terms of sample size and degrees of freedom. This approach has advantages over DFI in that it is faster to compute and there are no differences in the adjustment between different types of models (e.g., one-factor vs. multifactor). However, a possible limitation relative to DFI is that aspects other than sample size and degrees of freedom are not included in the equivalence testing correction factor despite aspects like the number of factors, the number of items per factor, and factor reliability being reported to affect fit index values.

As an anecdotal example using values in Table 8 that manipulated factor reliability, the RMSEA DFI cutoffs for N = 500 was .026 for H = .69 and .088 for H = .94. However, the "close fit" equivalence testing cutoff is .059 for both H = .69 and H = .94 because the equivalence testing correction factor does not include factor reliability. The fact that the cutoffs between DFI and equivalence testing are different anecdotally does not imply that either is superior and further investigation would be needed to compare the performance of these methods more comprehensively.

To give researchers the ability to make their own conclusions about the viability of DFI cutoffs and equivalence testing cutoffs, we modified the open-source R code provided by Yuan et al. (2016) and included an equivalence testing for RMSEA and CFI as part of Shiny application to make the method more accessible to researchers who wish to use it but who may not be comfortable with R functions or to researchers who want to study its properties further.

Limitations

Despite the advantages of DFI cutoffs we have illustrated in this paper, DFI cutoffs undoubtedly have limitations that are important to note and there is ample room for expansion and refinement of the approach. In particular,

1. In its current form, the DFI approach is only applicable to CFA models. The process described to create the data generation model does not necessarily apply to different modeling contexts. For instance, creating a data generation model by adding cross-loadings in latent growth models would not be meaningful because the empirical model in a latent growth context typically includes loadings from all growth factors to all repeated measures. Additional work would be needed to identify relevant misspecifications outside of CFA contexts (e.g., the ability to detect a moderate quadratic component or autocorrelation in residual variance might be more meaningful misspecifications in a latent growth model). The Shiny application we provide may still function for these models and provide results; however, the procedure we describe is not intended nor tested for these types of applications and results would, at this point, be invalid.

A clear future direction is to extend to logic of the DFI approach proposed here for CFA to other model types. Dynamizing measurement invariance cutoffs to accompany those suggested by Cheung and Rensvold (2002) and Chen (2007) would be one obvious next step with wide practical utility. Specialized approaches may also be needed for bifactor models because the approach of including additional cross-loadings in the data generation model may not function properly when all items already have cross-loadings by design (e.g., the formula for the maximum allowable cross-loading likely will be different). The same is true for second-order or hierarchical factor models where additional cross-loadings in the data generation model may have differential impact if applied to a first-order or second-order factor.

2. Though our intention is to improve model fit assessment by providing empirical researchers with custom cutoffs, an unintended consequence could be an increase in researcher degrees of freedom. With more options, researchers can more easily shop around for the fit criteria that aligns with the desired narrative. This could manifest, for instance, when a researcher has low factor reliability. Traditional fixed cutoffs would be more likely to return values indicating good fit, so the traditional cutoffs might be employed in such contexts. Rather than refining model fit assessment, the net effect could be a deluge of “good” fitting models because there are more paths that researchers could take to arrive at a conclusion of good fit.

It could be difficult to dissuade empirical researchers from using the traditional fixed cutoffs when it is against their interests or when a generation of researchers has committed fixed cutoffs to memory. The focus on quantification of misspecification with the DFI approach is intended to discourage this practice by making researchers more actively support their conclusion regarding fit, possibly by investigating local areas of strain and presenting additional types of validity arguments beyond the internal structure.

3. Global fit is the de facto modality for assessing model fit, but there is foundational theoretical criticism of this approach to model fit (e.g., Tomarken & Waller, 2003, 2005). For instance, even when data and model characteristics are identical, an SRMR of .05 can occur for different reasons. Because SRMR has global focus and averages over all elements of a covariance matrix, a model can achieve a .05 value by having several small discrepancies spread throughout the model or by having one large misspecification. The DFI approach does not address – nor does it attempt to address – the potential caveats within the broader global fit framework. Instead, the DFI approach is targeted at improving fit assessment within the confines of the global fit framework with which empirical researchers are familiar, but we fully acknowledge the deficiencies therein and the recent work that has noted the benefits of local fit inspections (e.g., Thommes, Rosseel, & Textor, 2018).

   Indeed, some might relate DFI cutoffs reliance on global fit to the portion of John Tukey's quote about the futility of precise answers to the wrong question. While an element of this sentiment exists, we would retort that model fit is sufficiently nebulous that current fixed cutoffs are an imprecise answer to an imperfect question while no consensus exists about what the right question even looks like. In this respect, having a precise answer to an imperfect question seems like an envious position to which model fit could aspire because, currently, the field does not have the right answer to any question when it comes to model fit.

4. Missing data and deviations from normality are not considered in the simulation design that we outline. The data generation process assumes complete data that are multivariate normal. Fit indices can be affected by both of these aspects (Davey, 2005; Zhang & Savalei, 2020)

and the degree to which they are present could affect performance of the resulting DFI cutoffs. The current version only considers continuous items, but support for categorical items with weighted least squares would be another high-priority extension.

5. Pragmatically, a single set of cutoffs is simpler to interpret and enforce for editors, reviewers, and consumers of research broadly. Dynamically changing cutoffs could make evaluating research more difficult. The nature of dynamic cutoffs requires researchers to perform additional computations, meaning that there is a non-zero chance that the DFI cutoffs could be calculated incorrectly, which would be difficult to detect. Our hope is that the Shiny application accompanying this paper reduces the complexity of obtaining these values by removing essentially all the programmatic requirements. Nonetheless, the presence of additional steps beyond committing fixed values to memory naturally will correspond to at least some increase in instances of user error.

Concluding Remarks

Hu and Bentler (1999) is a seminal study in the field that was instrumental in shaping thinking about model fit and providing practical guidelines for assessing fit in empirical studies that was ultimately overgeneralized. These guidelines were based on a very narrow subspace of possible models and research in the intervening years has shown that fit indices beyond this narrow subspace behave differently. Our DFI approach and the associated Shiny application provides an accessible way to exploit benefits of simulation-based techniques to better tailor cutoffs to the data and model characteristics being evaluated. By assessing different magnitudes of misspecification, the DFI approach also helps return fit indices to their intended use of being effect size measures that quantify misfit rather being used as ad hoc hypothesis tests for whether misfit is present. Though there are clear opportunities for extension and refinement of the DFI approach beyond its current form, we hope that this paper helps to move the methodological literature away from lamenting about poor generalizability of fixed cutoffs and towards modern solutions that empirical researchers can adopt to evaluate their models more precisely and more accurately.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Validity. In *Standards for educational and psychological testing* (pp. 11–31). Washington, DC: American Psychological Association.

Anthoine, E., Moret, L., Regnault, A., Sébille, V., & Hardouin, J. B. (2014). Sample size used to validate a scale: a review of publications on newly-developed patient reported outcomes measures. *Health and Quality of Life Outcomes*, *12*(1), 1-10.

Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods*, *24*(1), 1-19.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences, 42*(5), 815–824.

Browne, M. W., MacCallum, R. C., Kim, C. T., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods*, *7*(4), 403-421.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230-258.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research, 36*(4), 462–494.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255.

Cohen, J. (1988). *Statistical power analysis for the social sciences*. Lawrence Earlbaum Associates.

Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2020). Avoid Cohen's 'small', 'medium', and 'large' for power analysis. *Trends in Cognitive Sciences, 24*(3), 200–207.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the" problem" of sample size: A clarification. *Psychological Bulletin*, *109*(3), 512-519.

Davey, A. (2005). Issues in evaluating model fit with missing data. *Structural Equation Modeling, 12*(4), 578–597.

Edwards, M. C., & Wirth, R. J. (2009). Measurement and the study of change. *Research in Human Development*, *6*(2-3), 74-96.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: rationale of two-index strategy revisited. *Structural Equation Modeling, 12*(3), 343–367.

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

Field, A. P. (2005). *Discovering statistics with SPSS* (2nd ed.). London: Sage.

Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F., & Borsboom, D. (2016). Measuring depression over time... Or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment*, *28*(11), 1354-1367.

Fried, E. I., & Nesse, R. M. (2015). Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Medicine*, *13*(1), 1-11.

Gomer, B., Jiang, G., & Yuan, K. H. (2019). New effect size measures for structural equation modeling. *Structural Equation Modeling*, *26*(3), 371-389.

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493-498.

Greiff, S., & Heene, M. (2017). Why psychological assessment needs to start worrying about model fit. *European Journal of Psychological Assessment, 33*(5), 313–317.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck & D. Sörbom (Eds.), *Structural equation modeling: Present and future—A festschrift in honor of Karl Jöreskog* (pp. 195–216). Scientific Software International.

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.

Harring, J. R., McNeish, D. M., & Hancock, G. R. (2017). Using phantom variables in structural equation modeling to assess model sensitivity to external misspecification. *Psychological Methods, 22*(4), 616–631.

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods, 16*(3), 319–336.

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

Hu, L., & 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.

Hu, L. T., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted?. *Psychological Bulletin*, *112*(2), 351-362.

Jackson, D. L. (2007). The effect of the number of observations per parameter in misspecified confirmatory factor analytic models. *Structural Equation Modeling, 14*(1), 48–76.

Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14*(1), 6–23.

Kane, M. (2013). The Argument-Based Approach to Validation. *School Psychology Review*, 42(4), 448–457.

Kang, Y., McNeish, D., & Hancock, G.R. (2016). The role of measurement quality on practical guidelines for assessing measurement and structural invariance. *Educational and Psychological Measurement. 76 (*4), 533-561.

Kearns, N. T., Blumenthal, H., Natesan, P., Zamboanga, B. L., Ham, L. S., & Cloutier, R. M. (2018). Development and initial psychometric validation of the Brief-Caffeine Expectancy Questionnaire (B-CaffEQ). *Psychological Assessment*, *30*(12), 1597-1611.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research, 44*(3), 486–507.

Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling, 10*(3), 333–351.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241-253.

Kuhfeld, M., & Soland, J. (2020). Avoiding bias from sum scores in growth estimates: An examination of IRT-based approaches to scoring longitudinal survey responses. *Psychological Methods*, advance online publication.

Lai, K., & Kelley, K. (2011). Accuracy in parameter estimation for targeted effects in structural equation modeling: Sample size planning for narrow confidence intervals. *Psychological Methods*, *16*(2), 127-148.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130–149.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490-504.

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research*, *36*(4), 611-637.

MacKinnon, D. P., & Pirlott, A. G. (2015). Statistical approaches for enhancing causal interpretation of the M to Y relation in mediation analysis. *Personality and Social Psychology Review*, *19*(1), 30-43.

Marcoulides, K. M., & Yuan, K. H. (2017). New ways to evaluate goodness of fit: A note on using equivalence testing to assess structural equation models. *Structural Equation Modeling*, *24*(1), 148-153.

Markland, D. (2007). The golden rule is that there are no golden rules: A commentary on Paul Barrett's recommendations for reporting model fit in structural equation modelling. *Personality and Individual Differences*, *42*(5), 851-858.

Marsh, H. W., Hau, K. T., & Grayson, D. (in press). Goodness of fit evaluation in structural equation modeling. In A. Maydeu-Olivares & J. McArdle (Eds.), *Psychometrics. A festschrift to Roderick P. McDonald*. Mahwah, NJ: Erlbaum Associates.

Marsh, H. W., Hau, K.-T., & 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.

Maydeu-Olivares, A. (2017). Assessing the size of model misfit in structural equation models. *Psychometrika*, *82*(3), 533-558.

McDonald, R. P. (2010). Structural models and the art of approximation. *Perspectives on Psychological Science*, *5*(6), 675-686.

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

McIntosh, C. N. (2012). Improving the evaluation of model fit in confirmatory factor analysis: A commentary on Gundy, CM, Fayers, PM, Groenvold, M., Petersen, M. Aa., Scott, NW, Sprangers, MAJ, Velikov, G., Aaronson, NK (2011). Comparing higher-order models for the EORTC QLQ-C30. Quality of Life Research. *Quality of Life Research*, *21*(9), 1619-1621.

McNeish, D., An, J., & Hancock, G. R. (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *Journal of Personality Assessment, 100*(1), 43–52.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*, 2287-2305.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806-834.

Miles, J., & Shevlin, M. (2007). A time and a place for incremental fit indices. *Personality and Individual Differences, 42*(5), 869–874.

Millsap, R. E. (2013). A simulation paradigm for evaluating model fit. In M. Edwards & R. MacCallum (Eds.), *Current issues in the theory and application of latent variable models* (pp. 165–182). New York: Routledge

Millsap, R. E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences*, *42*(5), 875-881.

Moshagen, M. (2012). The model size effect in SEM: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling*, *19*(1), 86-98.

Mulaik, S.A. (2009). *Foundations of factor analysis.* New York: McGraw-Hill.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*(4), 599-620.

Pornprasertmanit, S., Miller, P., Schoemann, A., Jorgensen, T. & Quick, C. (2020). simsem: SIMulated Structural Equation Modeling. R package version 0.5-15. Retrieved from http://CRAN.R-project.org/package=simsem

Pornprasertmanit, S., Wu, W., & Little, T. D. (2013). Using a Monte Carlo approach for nested model comparisons in structural equation modeling. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology* (pp. 187–197). New York: Springer.

Ropovik, I. (2015). A cautionary note on testing latent variable models. *Frontiers in Psychology*, 6.

Rosenbaum, P. R. (2002). Overt bias in observational studies. In P. R. Rosenbaum (Ed.), *Observational Studies* (pp. 71–104). Springer.

Rosenbaum, P. R. (2010). *Design of observational studies*. Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

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

Saris, W. E., Satorra, A., & Veld, W. M. van der. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16*(4), 561–582.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507–514.

Schönemann, P. H. (1981). Power as a function of communality in factor analysis. *Bulletin of the Psychonomic Society, 17*(1), 57–60.

Schneider, W. J. (2019). simstandard: Generate standardized data. R package version 0.3.0. https://CRAN.R-project.org/package=simstandard

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement, 79*(2), 310–334.

Shi, D., & Maydeu-Olivares, A. (2020). The effect of estimation methods on SEM fit indices. *Educational and Psychological Measurement*, *80*(3), 421-445.

Sivo, S. A., Fan, X., Witta, E. L., & Willse, J. T. (2006). The search for "optimal" cutoff properties: fit index criteria in structural equation modeling. *The Journal of Experimental Education, 74*(3), 267–288.

Slof-Op't Landt, M. C. T., Van Furth, E. F., Rebollo-Mesa, I., Bartels, M., van Beijsterveldt, C.

E. M., Slagboom, P. E., ... & Dolan, C. V. (2009). Sex differences in sum scores may be hard to interpret: the importance of measurement invariance. *Assessment*, *16*(4), 415-423.

Tanaka, J. S. (1987). " How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. *Child Development*, 134-146.

Thoemmes, F., Rosseel, Y., & Textor, J. (2018). Local fit evaluation of structural equation models using graphical criteria. *Psychological Methods, 23*(1), 27–41.

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with "well fitting" models. *Journal of Abnormal Psychology, 112*(4), 578–598.

Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology, 1*(1), 31–65.

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

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). New York, NY: Guilford.

Wu, H., & Browne, M. W. (2015). Quantifying adventitious error in a covariance structure as a random effect. *Psychometrika*, *80*(3), 571-600.

Yuan, K. H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling*, *23*(3), 319-330.

Yuan, K. H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research*, *40*(1), 115-148.

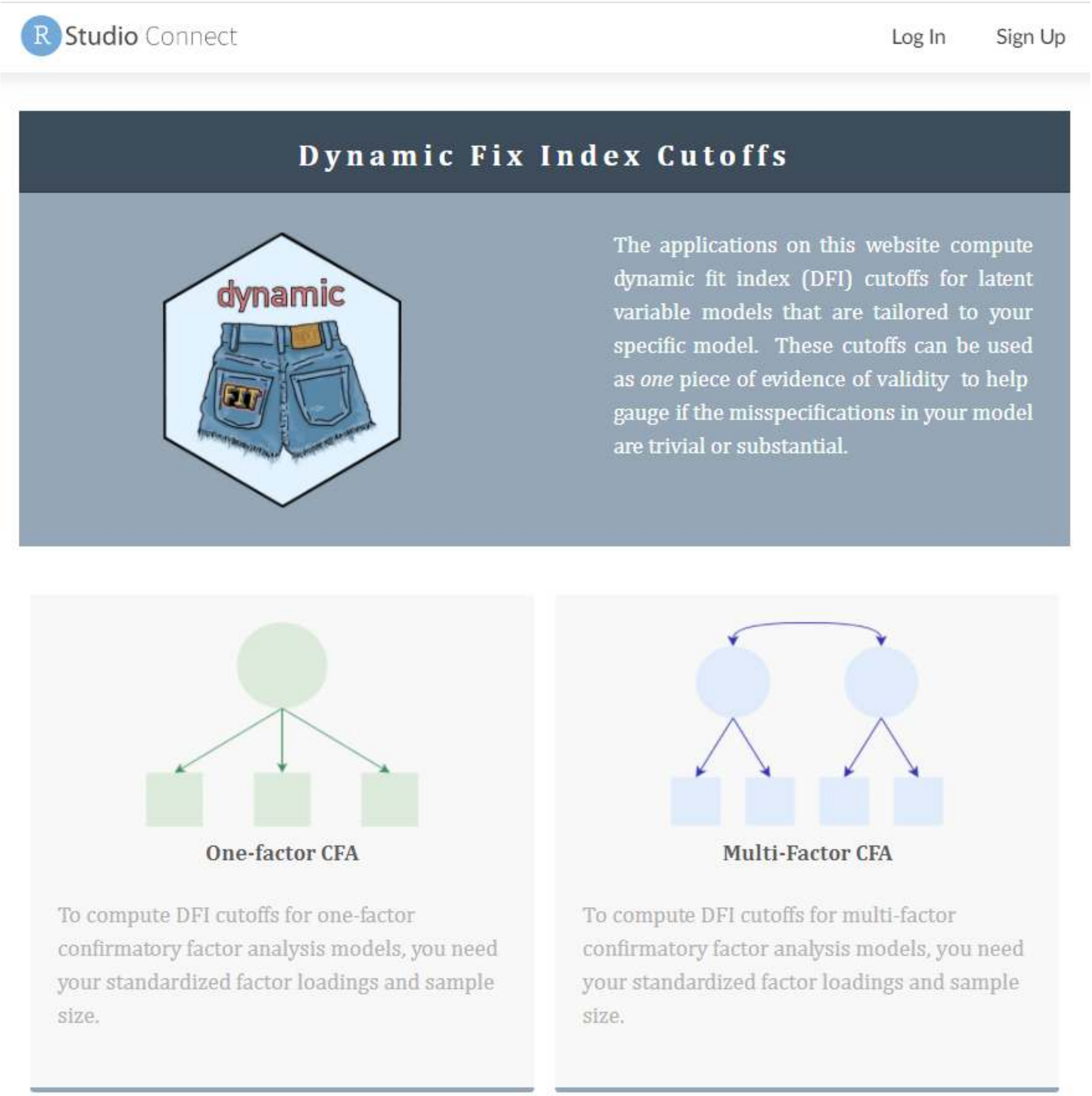
Yuan, K. H., & Bentler, P. M. (1999). F tests for mean and covariance structure analysis. *Journal of Educational and Behavioral Statistics*, *24*(3), 225-243.

Zhang, X., & Savalei, V. (2020). Examining the effect of missing data on RMSEA and CFI under normal theory full-information maximum likelihood. *Structural Equation Modeling*, *27*(2), 219-239.

Appendix A

Short Tutorial on Using Web-Based Application

The application website is located at www.dynamicfit.app and the landing page displays the models that are currently supported, one-factor CFA and multifactor CFA (there is also an application for equivalence testing, which is not pictured here). Clicking either image will open a software application specific to the model type of interest.



Dynamic Fix Index Cutoffs

The applications on this website compute dynamic fit index (DFI) cutoffs for latent variable models that are tailored to your specific model. These cutoffs can be used as *one* piece of evidence of validity to help gauge if the misspecifications in your model are trivial or substantial.

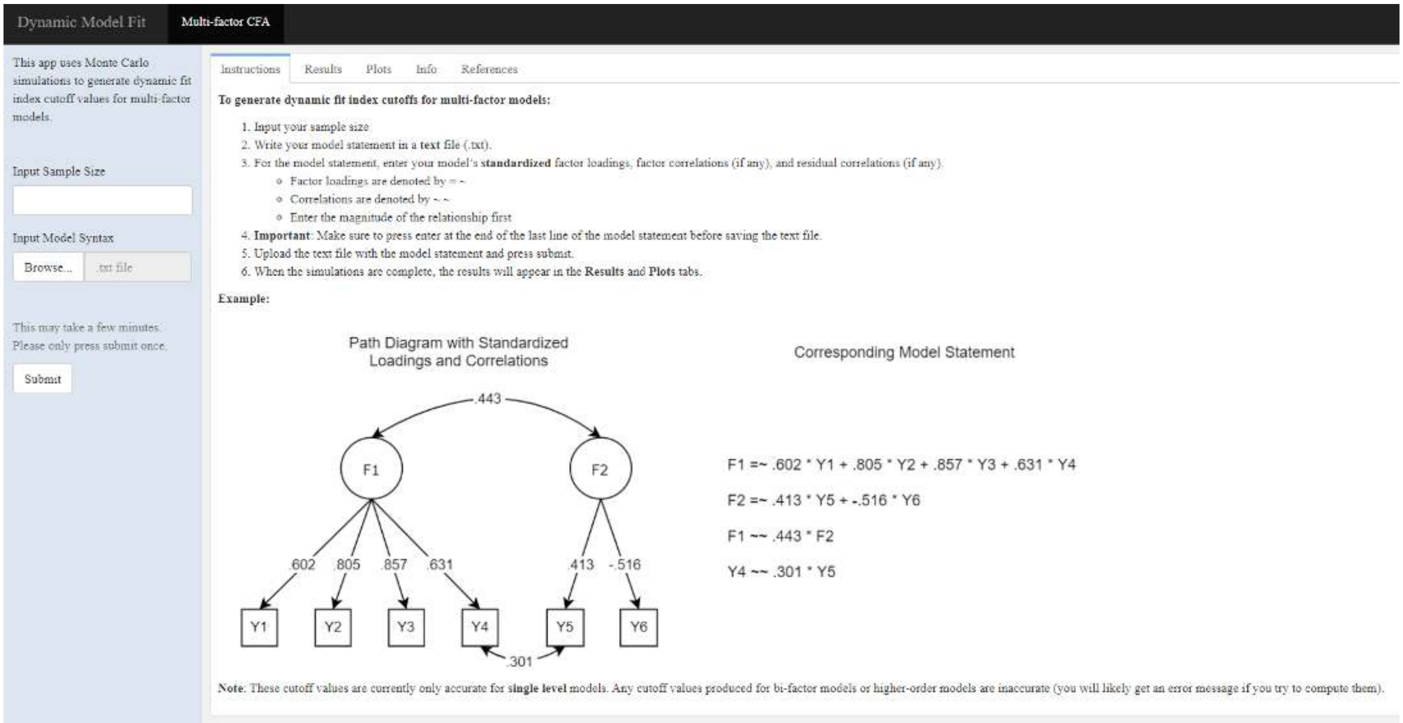
One-factor CFA

To compute DFI cutoffs for one-factor confirmatory factor analysis models, you need your standardized factor loadings and sample size.

Multi-Factor CFA

To compute DFI cutoffs for multi-factor confirmatory factor analysis models, you need your standardized factor loadings and sample size.

The homepage for each application gives instructions for how the application works along with a simple example. The general idea is that researchers provide their CFA model with the standardized estimates in lavaan notation in a .txt file and upload it on the left side of the page (if users are unfamiliar with lavaan syntax, the instructions page describes how to specify the models in this form). A sample model statement with a corresponding path diagram is included for clarity. The user also enters the sample size from their empirical dataset. After these two steps are done, users click the "Submit" button to begin calculations.



This app uses Monte Carlo simulations to generate dynamic fit index cutoff values for multi-factor models.

This may take a few minutes. Please only press submit once.

To generate dynamic fit index cutoffs for multi-factor models:

1. Input your sample size
2. Write your model statement in a **text** file (.txt).
3. For the model statement, enter your model's **standardized** factor loadings, factor correlations (if any), and residual correlations (if any).
   - Factor loadings are denoted by = ~
   - Correlations are denoted by ~ ~
   - Enter the magnitude of the relationship first
4. **Important**: Make sure to press enter at the end of the last line of the model statement before saving the text file.
5. Upload the text file with the model statement and press submit.
6. When the simulations are complete, the results will appear in the **Results** and **Plots** tabs.

Example:

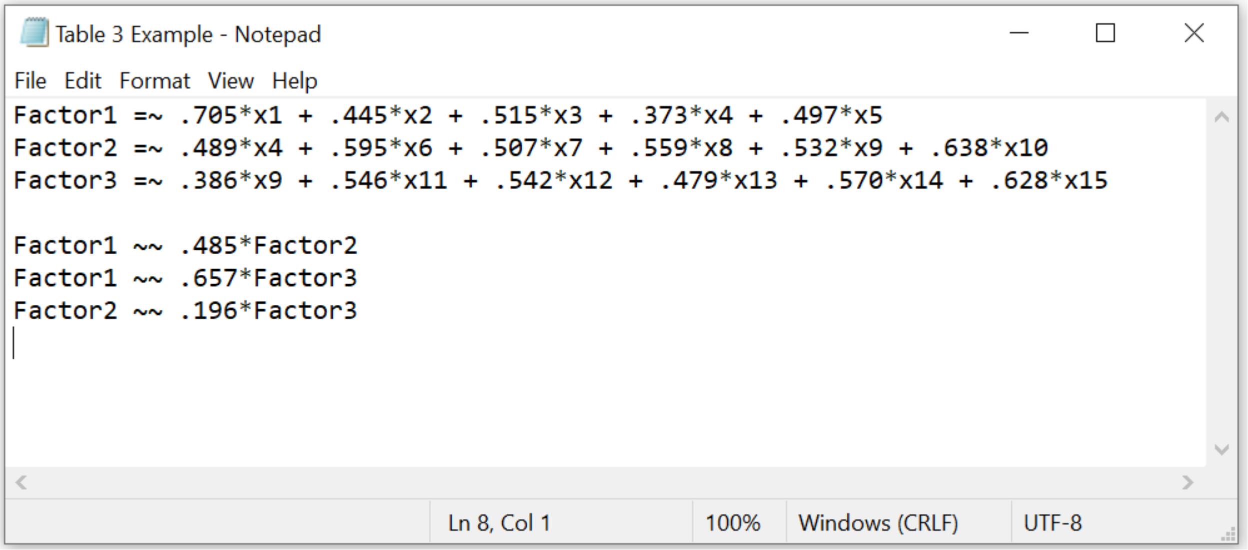
Path Diagram with Standardized Loadings and Correlations

Corresponding Model Statement

```
F1 =~ .602 * Y1 + .805 * Y2 + .857 * Y3 + .631 * Y4

F2 =~ .413 * Y5 + -.516 * Y6

F1 ~~ .443 * F2

Y4 ~~ .301 * Y5
```

Note: These cutoff values are currently only accurate for **single level** models. Any cutoff values produced for bi-factor models or higher-order models are inaccurate (you will likely get an error message if you try to compute them).

The corresponding .txt file for the example shown in Table 3 is

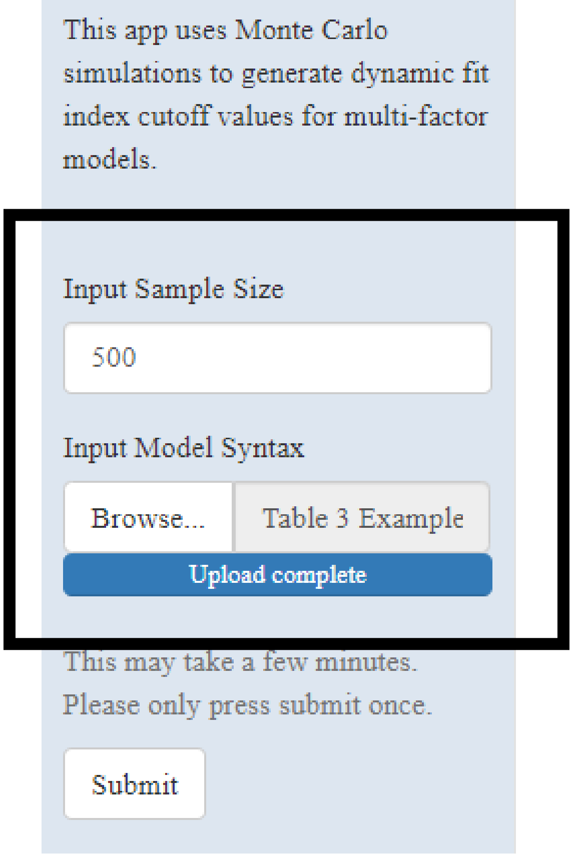


```
Factor1 =~ .705*x1 + .445*x2 + .515*x3 + .373*x4 + .497*x5
Factor2 =~ .489*x4 + .595*x6 + .507*x7 + .559*x8 + .532*x9 + .638*x10
Factor3 =~ .386*x9 + .546*x11 + .542*x12 + .479*x13 + .570*x14 + .628*x15

Factor1 ~~ .485*Factor2
Factor1 ~~ .657*Factor3
Factor2 ~~ .196*Factor3
```

And is presented in text here for users that wish to copy and paste the model statement into a text file:

```
Factor1 =~ .705*x1 + .445*x2 + .515*x3 + .373*x4 + .497*x5
Factor2 =~ .489*x4 + .595*x6 + .507*x7 + .559*x8 + .532*x9 + .638*x10
Factor3 =~ .386*x9 + .546*x11 + .542*x12 + .479*x13 + .570*x14 + .628*x15

Factor1 ~~ .485*Factor2
Factor1 ~~ .657*Factor3
Factor2 ~~ .196*Factor3
```

This file is uploaded and the sample size is set to 500.



This app uses Monte Carlo simulations to generate dynamic fit index cutoff values for multi-factor models.

Input Sample Size

500

Input Model Syntax

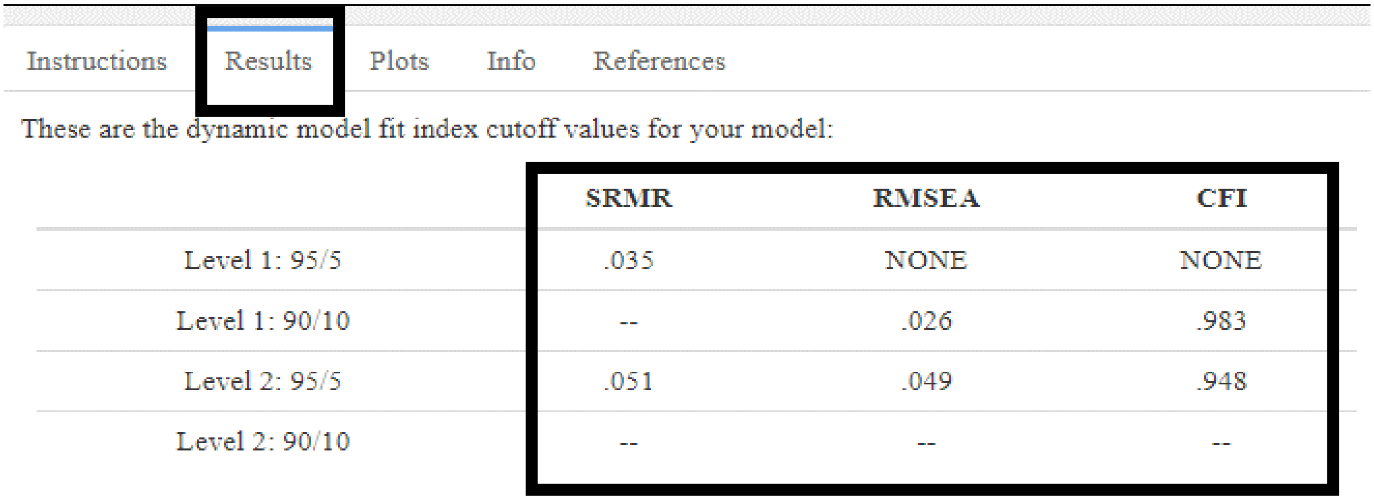
Table 3 Example

This may take a few minutes. Please only press submit once.

After clicking submit, a progress bar will appear at the top of the page to indicate that the application is running to (a) parse the .txt file to determine which misspecification to include in the data generation model, (b) create code for the data generation model, (c) generate 500 replication from the data generation model, (d) fit the empirical model to each, (e) generate 500 datasets from the empirical model, (f) fit the empirical model to each generated dataset, and (g) summarize the fit index values to determine the dynamic cutoffs for SRMR, RMSEA, and CFI.

Once the process is complete, the progress bar will disappear and clicking on the "Results" tab will show the DFI cutoffs from the simulation. For the multifactor CFA application, the number of misspecifications will vary by model size. In this example, the model has 3 factors, so two different levels of misspecification are shown. The Level-1 rows correspond to a data generation model with one additional cross-loading and the Level-2 rows correspond to a data generation model with two additional cross-loadings. For the one-factor CFA application, 3 levels will always be shown unless the number of items is too small to support 3 levels.

Hu and Bentler (1999) did not use a consistent criterion to determine where the cutoff should be relative to the correct and misspecified model distributions. From the results in their paper, the common approaches were (a) to reject 95% of misspecified models while rejecting no more than 5% of true models or (b) to reject 90% of misspecified models while rejecting no more than 10% of true models. The application will show the cutoffs satisfying Criterion A for each level of misspecification if possible. If Criterion A cannot be met, then it will show the cutoffs that satisfy Criterion B. If neither can be satisfied, all cells will read "NONE".

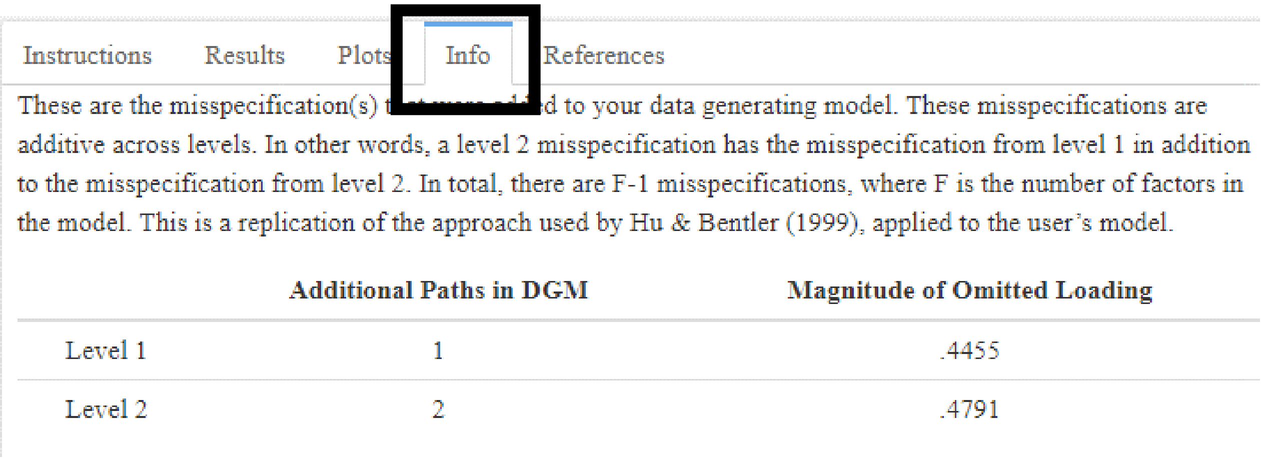


These are the dynamic model fit index cutoff values for your model:

	SRMR	RMSEA	CFI
Level 1: 95/5	.035	NONE	NONE
Level 1: 90/10	--	.026	.983
Level 2: 95/5	.051	.049	.948
Level 2: 90/10	--	--	--

The “Info” tab in the multifactor CFA application will show which misspecifications were added for each level of misspecification. In this example, the Level-1 data generation model contained a loading with magnitude .4455 that was not present in the empirical model. The Level-2 data generation model contained two loadings with magnitude .4455 and .4791 in the data generation model that were not present in the empirical model.

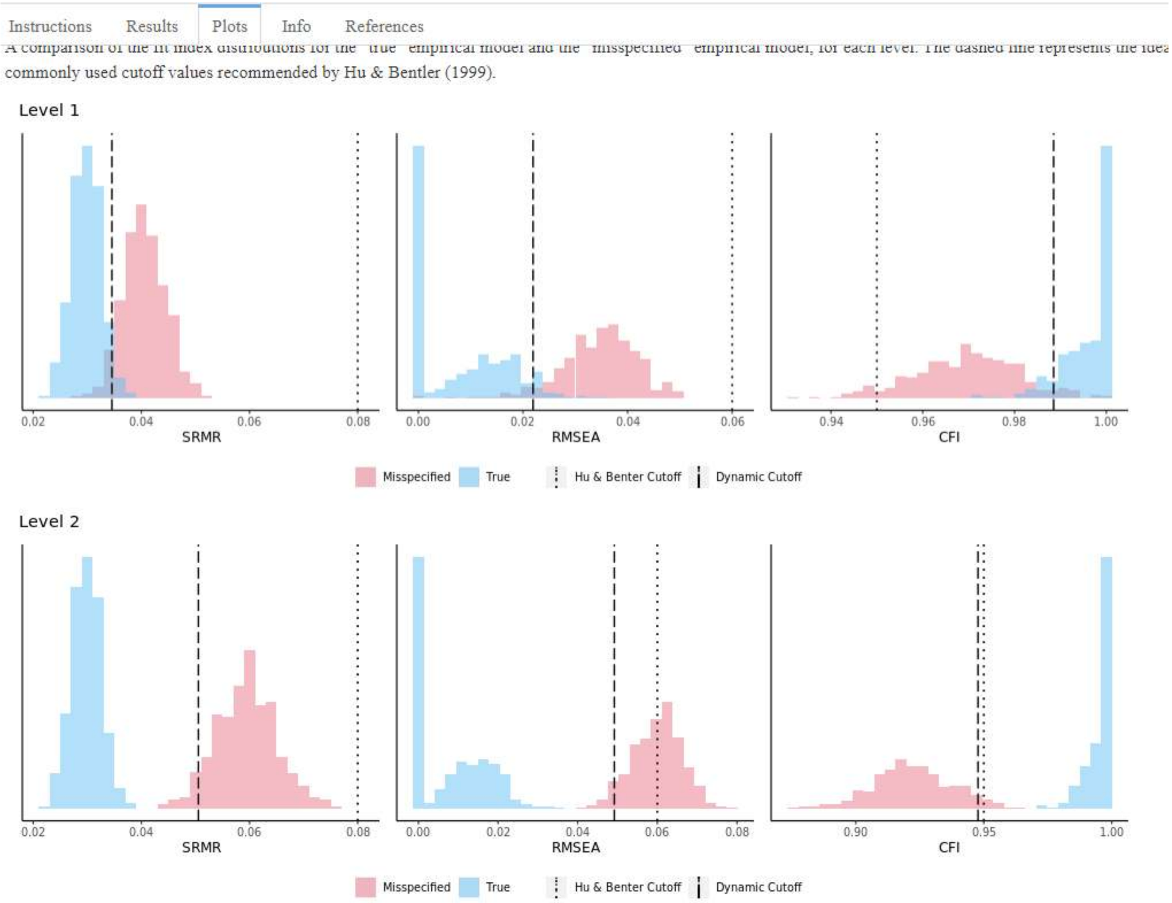
The one-factor CFA tab does not have an “Info” tab because the additional paths in the data generation model are more standardized and always include residual correlations equal to .30 between one-third of items for Level-1, two-thirds of items for Level-2, and all items for Level-3.



These are the misspecification(s) that were added to your data generating model. These misspecifications are additive across levels. In other words, a level 2 misspecification has the misspecification from level 1 in addition to the misspecification from level 2. In total, there are F-1 misspecifications, where F is the number of factors in the model. This is a replication of the approach used by Hu & Bentler (1999), applied to the user’s model.

	Additional Paths in DGM	Magnitude of Omitted Loading
Level 1	1	.4455
Level 2	2	.4791

The “Plots” tab shows the fit index distributions for the correct and misspecified models for each level of misspecification to show the overlap between the distributions and how the DFI cutoffs were derived. A reference line is also added to facilitate comparing the DFI cutoffs to the traditional Hu and Bentler (1999) cutoffs.



A comparison of the fit index distributions for the "true" empirical model and the "misspecified" empirical model, for each level. The dashed line represents the idea
commonly used cutoff values recommended by Hu & Bentler (1999).

Appendix B

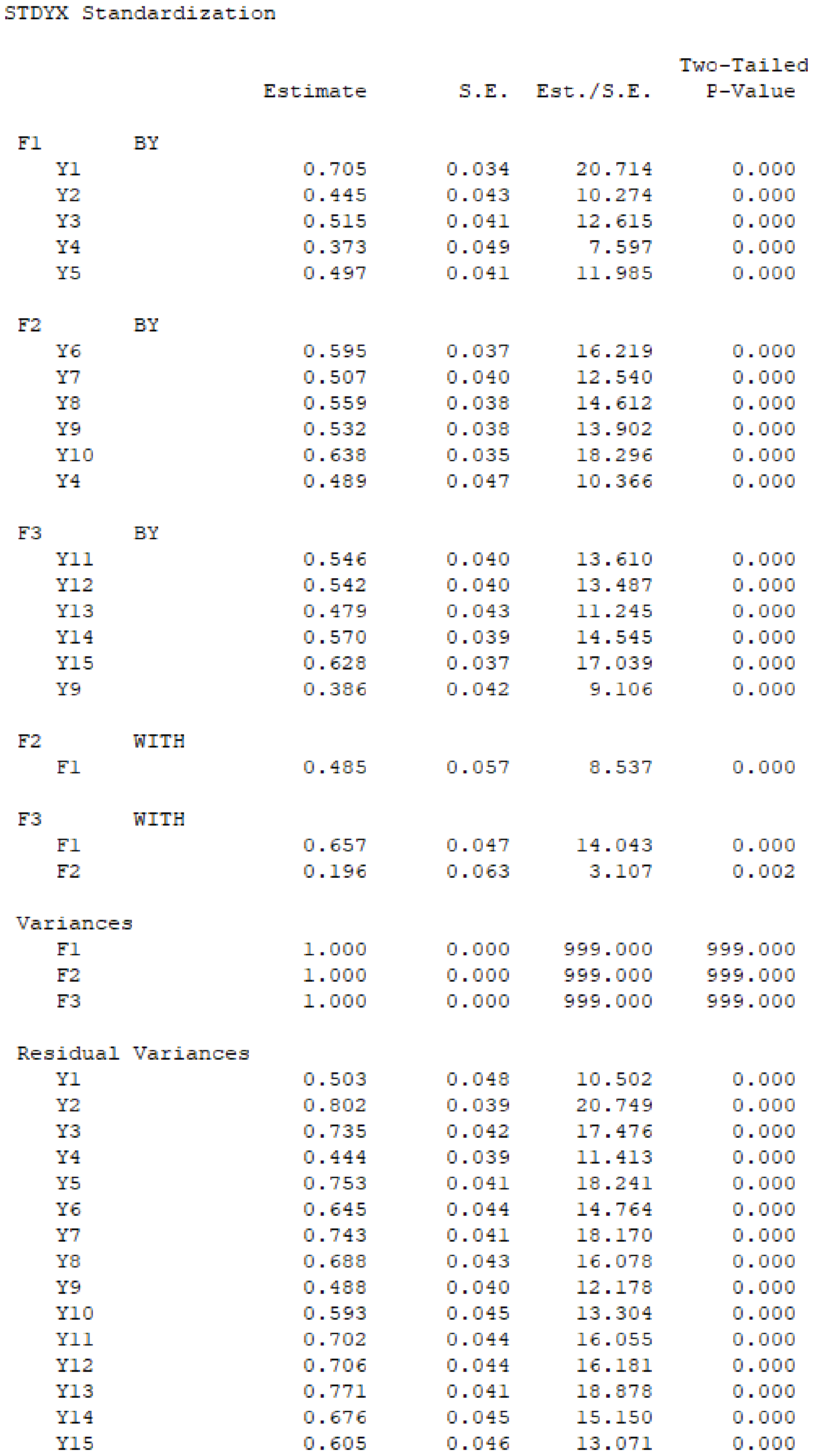
Implementation in Mplus

This appendix walks through how to obtain dynamic cutoffs from the “Simulated Data Example” section of the main text that uses the first replication from the $N = 500$, $H = 0.69$ condition from the Complex model. We will go through each of the 5 steps and show M*plus* code and screenshots to explain how to obtain the DFI cutoffs and how to interpret them. M*plus* Version 8.3 was used throughout this document.

Step 1: Fit the Empirical Model

The path diagram for the empirical model was shown in Figure 4 in the main text. M*plus* code for the model is shown below. The latent variables are assigned scale by constraining the factor variances to 1 and the first loading of each item is freely estimated. The last line requests that standardized loadings be output since they will be used in subsequent steps. The model is a standard CFA model with no mean structure, so we will not go through each line of the code as we assume that M*plus* users are familiar with the basic setup of CFA model in the program. The standardized estimates from this code as reported in Table 3 of the main text are shown on the next page

```
DATA: FILE IS Complex Rep1 .55 Loadings.csv;
VARIABLE: NAMES ARE y1-y15;

ANALYSIS:
ESTIMATOR = ML;
MODEL = NOMEANSTRUCTURE;
INFORMATION = EXPECTED;

MODEL:
f1 BY y1* y2-y5;
f2 BY y6* y7-y10 y4;
f3 BY y11* y12-y15 y9;

f1@1;
f2@1;
f3@1;

OUTPUT: STDYX;
```



```
STDYX Standardization

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

 F1       BY
    Y1                 0.705      0.034     20.714      0.000
    Y2                 0.445      0.043     10.274      0.000
    Y3                 0.515      0.041     12.615      0.000
    Y4                 0.373      0.049      7.597      0.000
    Y5                 0.497      0.041     11.985      0.000

 F2       BY
    Y6                 0.595      0.037     16.219      0.000
    Y7                 0.507      0.040     12.540      0.000
    Y8                 0.559      0.038     14.612      0.000
    Y9                 0.532      0.038     13.902      0.000
    Y10                0.638      0.035     18.296      0.000
    Y4                 0.489      0.047     10.366      0.000

 F3       BY
    Y11                0.546      0.040     13.610      0.000
    Y12                0.542      0.040     13.487      0.000
    Y13                0.479      0.043     11.245      0.000
    Y14                0.570      0.039     14.545      0.000
    Y15                0.628      0.037     17.039      0.000
    Y9                 0.386      0.042      9.106      0.000

 F2       WITH
    F1                 0.485      0.057      8.537      0.000

 F3       WITH
    F1                 0.657      0.047     14.043      0.000
    F2                 0.196      0.063      3.107      0.002

 Variances
    F1                 1.000      0.000    999.000    999.000
    F2                 1.000      0.000    999.000    999.000
    F3                 1.000      0.000    999.000    999.000

 Residual Variances
    Y1                 0.503      0.048     10.502      0.000
    Y2                 0.802      0.039     20.749      0.000
    Y3                 0.735      0.042     17.476      0.000
    Y4                 0.444      0.039     11.413      0.000
    Y5                 0.753      0.041     18.241      0.000
    Y6                 0.645      0.044     14.764      0.000
    Y7                 0.743      0.041     18.170      0.000
    Y8                 0.688      0.043     16.078      0.000
    Y9                 0.488      0.040     12.178      0.000
    Y10                0.593      0.045     13.304      0.000
    Y11                0.702      0.044     16.055      0.000
    Y12                0.706      0.044     16.181      0.000
    Y13                0.771      0.041     18.878      0.000
    Y14                0.676      0.045     15.150      0.000
    Y15                0.605      0.046     13.071      0.000
```

Step 2: Create the Data Generation Model

The main text describes the process by which the data generation model is created, which corresponds to path diagram in Figure 5. In this data, the selected item was Item 2 and the associated cross-loading is 0.445. As noted in the main text, this cross-loading was below the maximum allowable value so a loading from Factor 2 to Item 2 with value of 0.445 will be added to the data generation model. In doing so, we need to change the standardized residual variance of Item 2 to reflect the additional variance explained by the cross-loading to ensure that the total variance of Item 2 remains equal to 1. The total explained variance for Item 2 is now

$$0.445^2 + 0.445^2 + 2\left(.445 \times .485 \times .445\right) = 0.588$$

meaning that the standardized residual variance should be $1 - .588 = .412$. All other standardized residual variances can remain at the value shown in the above output.

Step 3: Generate Data and Fit the Empirical Model

M*plus* code for creating the data generation model and fitting the empirical model to the generated data is presented below. For readers unfamiliar with the MONTECARLO utility in M*plus*, unlike a traditional analysis, no data are read into the program. First, users tell M*plus* the names of the variables they want to generate in each dataset, the number of observations to generate per dataset, and the number of unique datasets to generate. Then, users provide a model from which to generate data in the `MODEL MONTECARLO` command (as determined in Step 2). The model is written out just like any other M*plus* model except that users must give M*plus* population values from which to generate data. These population values come after an asterisk and must be provided for every parameter in the model. The population values are taken from the standardized estimates in Step 1 (with the exception of the residual variance for the item which contains an additional cross-loading, as mentioned in Step 2). Then, the `MODEL` command contains the standard M*plus* code for the original empirical model, which will be fit to the generated datasets. The M*plus* code below essentially says:

1. The `MONTECARLO` command tells M*plus* to create 1000 unique datasets such that each dataset has 500 people and 15 continuous variables labeled `y1` to `y15`
2. The values of the generated variables are randomly drawn from the model specified in the `MODEL MONTECARLO` command, which takes its values from the empirical model estimates with one additional cross-loading set equal to the weakest loading item.
3. The `MODEL` command then tells M*plus* to fit the empirical model to each generated dataset.

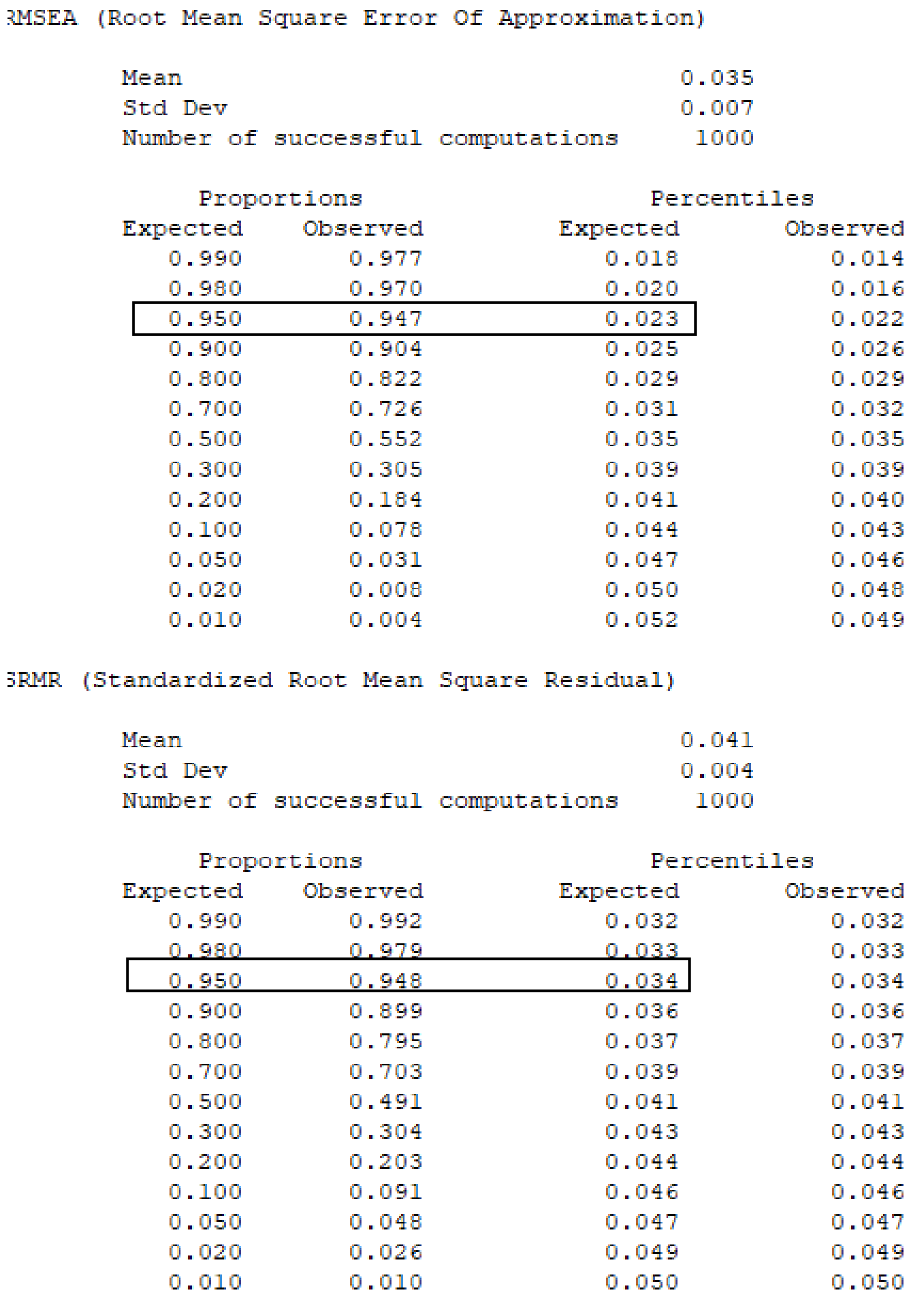
```
MONTECARLO:
!Generate 15 continuous variables;
NAMES = y1-y15;
!make the sample size 500 in each dataset;
NOBS = 500;
!create 1000 different datasets;
NREPS = 1000;
!set Seed for reproducible results;
SEED=1981;

!Remove estimates of item mean parameters;
ANALYSIS:
MODEL=NOMEANSTRUCTURE;
INFORMATION=EXPECTED;

!Data generation model used to generate the 1000 datasets;
MODEL MONTECARLO:

!Standardized Loadings for Factor 1;
f1 BY y1*.705 y2*.445 y3*.515 y4*.373 y5*.497;
!Standardized Loadings for Factor 2;
f2 BY y6*.595 y7*.507 y8*.559 y9*.532 y10*.638 y4*.489;
!Standardized loadings for Factor 3;
f3 BY y11*.546 y12*.542 y13*.479 y14*.570 y15*.628 y9*.386;

!Additional factor loading for weakest item, not present in empirical model;
f2 BY y2*.445;

!Factor Variances;
f1-f3*1;
!Factor covariances;
f1 WITH f2*.485;
f1 WITH f3*.657;
f2 WITH f3*.196;

!Standardized Residual variances ;
y1*.503;
y2*.412;
y3*.735;
y4*.444;
y5*.753;
y6*.645;
y7*.743;
y8*.688;
y9*.488;
y10*.593;
y11*.702;
y12*.706;
y13*.771;
y14*.676;
y15*.605;

!Empirical Model to be fit to each generated dataset;
MODEL:
f1 BY y1* y2-y5;
f2 BY y6* y7-y10 y4;
f3 BY y11* y12-y15 y9;
y1-y15;
f1-f3@1;
```

Step 4: Locate the 5th Percentile of the Misspecified Fit Index Distribution

This Monte Carlo simulation generates 1000 different datasets and each one is analyzed with the empirical model, meaning that there are 1000 different sets of output. Fortunately, M*plus* parses through each of these 1000 outputs automatically and summarizes the results in a single output file. The output contains outcomes that are commonly of interest in Monte Carlo simulation studies such as parameter estimate bias, confidence interval coverage, and – most importantly to this paper – distributions of fit indices. Not all fit indices are tracked by M*plus*, but the mean, standard deviation, and percentiles of the SRMR and RMSEA distributions are reported automatically.

The output from the analysis of this data is shown below for the RMSEA and the SRMR. Version 8.3 of M*plus* does not report the CFI for Monte Carlo studies. The leftmost column lists proportions and the rightmost column lists the fit index percentile value associated with that proportion. The proportions are one minus the percentile rank. We are interested in the 5th percentile as this identifies the fit index values to which 95% of misspecified model fit index values are greater than or equal. The 5th percentile can be located by looking at the Expected Percentile column that corresponds to the 0.950 Expected Proportion row.

In this case, the associated RMSEA value is 0.023 and the associated SRMR is 0.034, meaning that 95% of models with an omitted cross-loading yield an RMSEA value of 0.023 or higher and an SRMR of 0.034 or higher. In other words, using these values as cutoffs would yield 95% probability to accurately reject this model. Step 5 then will ensure that these values do not over-reject true models.



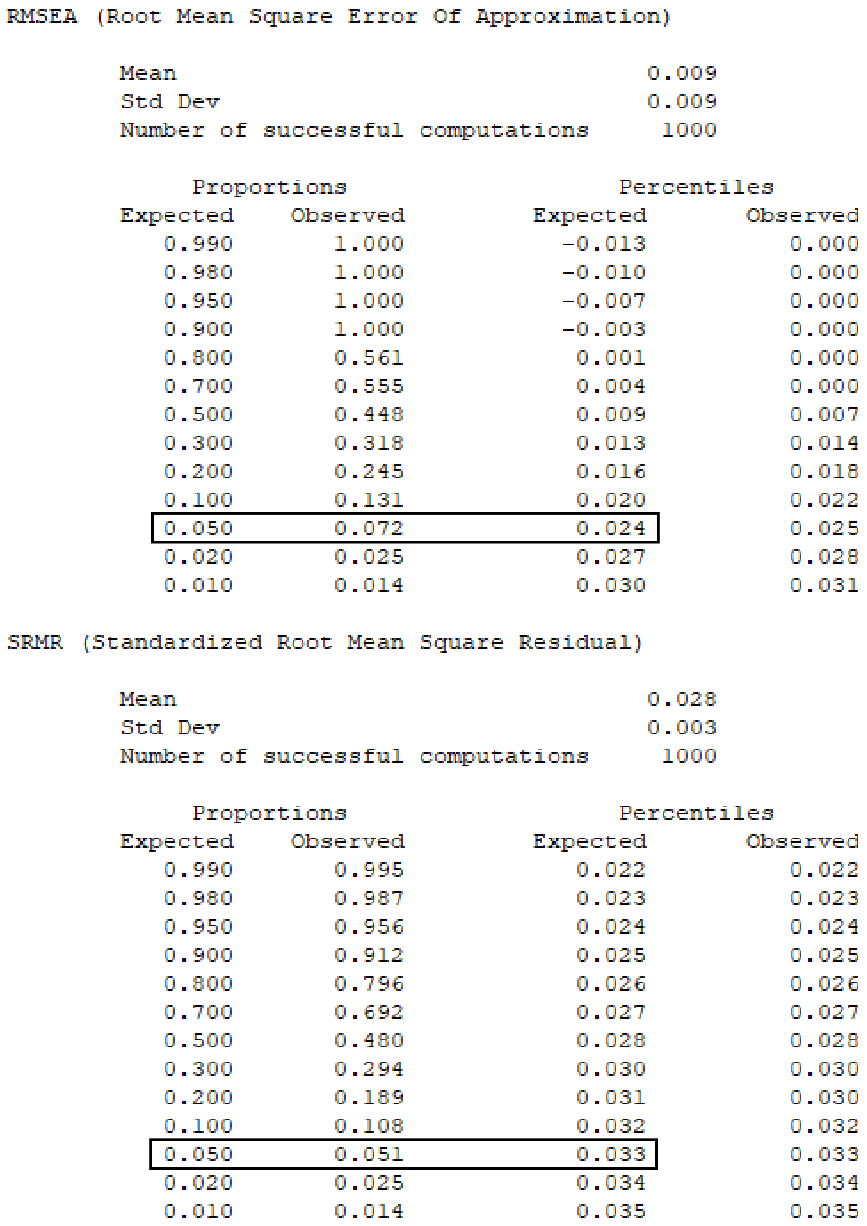
```
RMSEA (Root Mean Square Error Of Approximation)

       Mean                                  0.035
       Std Dev                               0.007
       Number of successful computations      1000

            Proportions                   Percentiles
       Expected    Observed          Expected       Observed
          0.990       0.977             0.018          0.014
          0.980       0.970             0.020          0.016
          0.950       0.947             0.023          0.022
          0.900       0.904             0.025          0.026
          0.800       0.822             0.029          0.029
          0.700       0.726             0.031          0.032
          0.500       0.552             0.035          0.035
          0.300       0.305             0.039          0.039
          0.200       0.184             0.041          0.040
          0.100       0.078             0.044          0.043
          0.050       0.031             0.047          0.046
          0.020       0.008             0.050          0.048
          0.010       0.004             0.052          0.049

SRMR (Standardized Root Mean Square Residual)

       Mean                                  0.041
       Std Dev                               0.004
       Number of successful computations      1000

            Proportions                   Percentiles
       Expected    Observed          Expected       Observed
          0.990       0.992             0.032          0.032
          0.980       0.979             0.033          0.033
          0.950       0.948             0.034          0.034
          0.900       0.899             0.036          0.036
          0.800       0.795             0.037          0.037
          0.700       0.703             0.039          0.039
          0.500       0.491             0.041          0.041
          0.300       0.304             0.043          0.043
          0.200       0.203             0.044          0.044
          0.100       0.091             0.046          0.046
          0.050       0.048             0.047          0.047
          0.020       0.026             0.049          0.049
          0.010       0.010             0.050          0.050
```

Step 5: Repeat Using the Empirical Model as the Data Generation Model

Step 4 showed that an RMSEA cutoff of 0.023 and a SRMR cutoff of 0.034 had a 95% probability to detect a misspecified cross-loading. However, we also need to determine whether these cutoffs do erroneously reject good models. We repeat Step 2 through Step 4 but make the data generation model equal to the empirical data (i.e., the data generation model is the model shown in Figure 4 in the main text and there are no additional cross-loadings included). The M*plus* code for the data generation model in Step 5 would be almost identical as Step 3 except that it would omit `f2 BY y2*.445` and reset the standardized residual variance of Item 2 to the value from the empirical analysis (i.e., `y2*.802` rather than `y2*.412`). The M*plus* code for this step is shown below.

```
MONTECARLO:
!Generate 15 continuous variables;
NAMES = y1-y15;
!make the sample size 500 in each dataset;
NOBS = 500;
!create 1000 different datasets;
NREPS = 1000;
!set Seed for reproducible results;
SEED=1981;

!Remove estimates of item mean parameters;
ANALYSIS:
MODEL=NOMEANSTRUCTURE;
INFORMATION=EXPECTED;

!Data generation model used to generate the 1000 datasets;
MODEL MONTECARLO:

!Standardized Loadings for Factor 1;
f1 BY y1*.705 y2*.445 y3*.515 y4*.373 y5*.497;
!Standardized Loadings for Factor 2;
f2 BY y6*.595 y7*.507 y8*.559 y9*.532 y10*.638 y4*.489;
!Standardized loadings for Factor 3;
f3 BY y11*.546 y12*.542 y13*.479 y14*.570 y15*.628 y9*.386;

!Factor Variances;
f1-f3*1;
!Factor covariances;
f1 WITH f2*.485;
f1 WITH f3*.657;
f2 WITH f3*.196;

!Standardized Residual variances ;
y1*.503;
y2*.802;
y3*.735;
y4*.444;
y5*.753;
y6*.645;
y7*.743;
y8*.688;
y9*.488;
y10*.593;
y11*.702;
y12*.706;
y13*.771;
y14*.676;
y15*.605;

!Empirical Model to be fit to each generated dataset;
MODEL:
f1 BY y1* y2-y5;
f2 BY y6* y7-y10 y4;
f3 BY y11* y12-y15 y9;
y1-y15;
f1-f3@1;
```

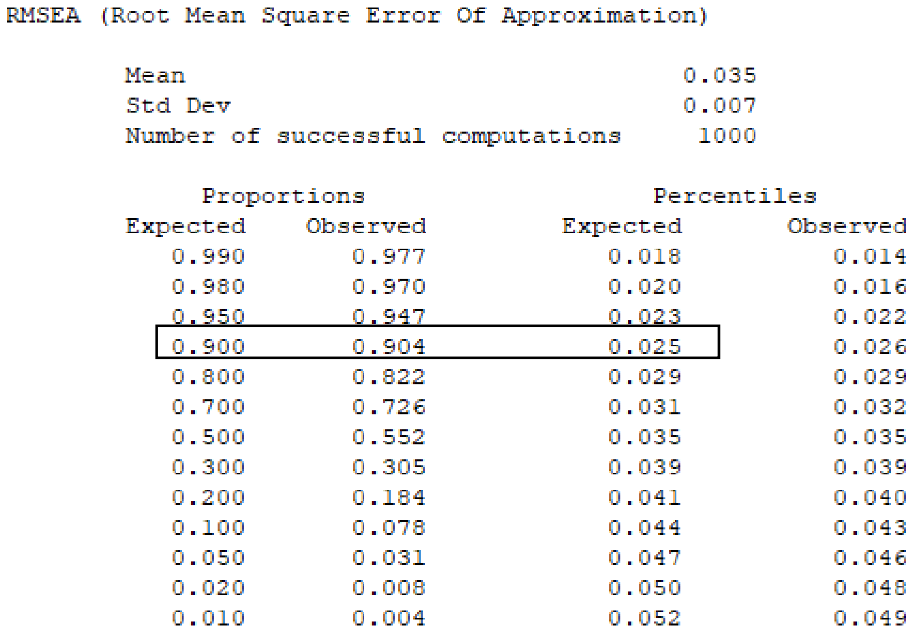
The M*plus* output from running this Monte Carlo code where the data generation model equal to the empirical model is shown below. Since this is the correct model, we are interested in the 95th percentile of the distribution. The 95th percentile can be located by looking at the Expected Percentile column corresponding to the 0.050 Expected Proportion row. The 95th percentile of the SRMR distribution is 0.033, which is below the 0.034 value from Step 4 meaning that an SRMR cutoff of 0.034 can distinguish between true and misspecified models. Unfortunately, the 95th percentile of the RMSEA distribution is 0.024, which is larger than the value obtained in Step 4, meaning that using 0.023 as the cutoff would reject more than 5% of true models.



```
RMSEA (Root Mean Square Error Of Approximation)

        Mean                                   0.009
        Std Dev                                0.009
        Number of successful computations       1000

             Proportions                  Percentiles
        Expected    Observed           Expected       Observed
           0.990       1.000             -0.013          0.000
           0.980       1.000             -0.010          0.000
           0.950       1.000             -0.007          0.000
           0.900       1.000             -0.003          0.000
           0.800       0.561              0.001          0.000
           0.700       0.555              0.004          0.000
           0.500       0.448              0.009          0.007
           0.300       0.318              0.013          0.014
           0.200       0.245              0.016          0.018
           0.100       0.131              0.020          0.022
           0.050       0.072              0.024          0.025
           0.020       0.025              0.027          0.028
           0.010       0.014              0.030          0.031

SRMR (Standardized Root Mean Square Residual)

        Mean                                   0.028
        Std Dev                                0.003
        Number of successful computations       1000

             Proportions                  Percentiles
        Expected    Observed           Expected       Observed
           0.990       0.995              0.022          0.022
           0.980       0.987              0.023          0.023
           0.950       0.956              0.024          0.024
           0.900       0.912              0.025          0.025
           0.800       0.796              0.026          0.026
           0.700       0.692              0.027          0.027
           0.500       0.480              0.028          0.028
           0.300       0.294              0.030          0.030
           0.200       0.189              0.031          0.030
           0.100       0.108              0.032          0.032
           0.050       0.051              0.033          0.033
           0.020       0.025              0.034          0.034
           0.010       0.014              0.035          0.035
```

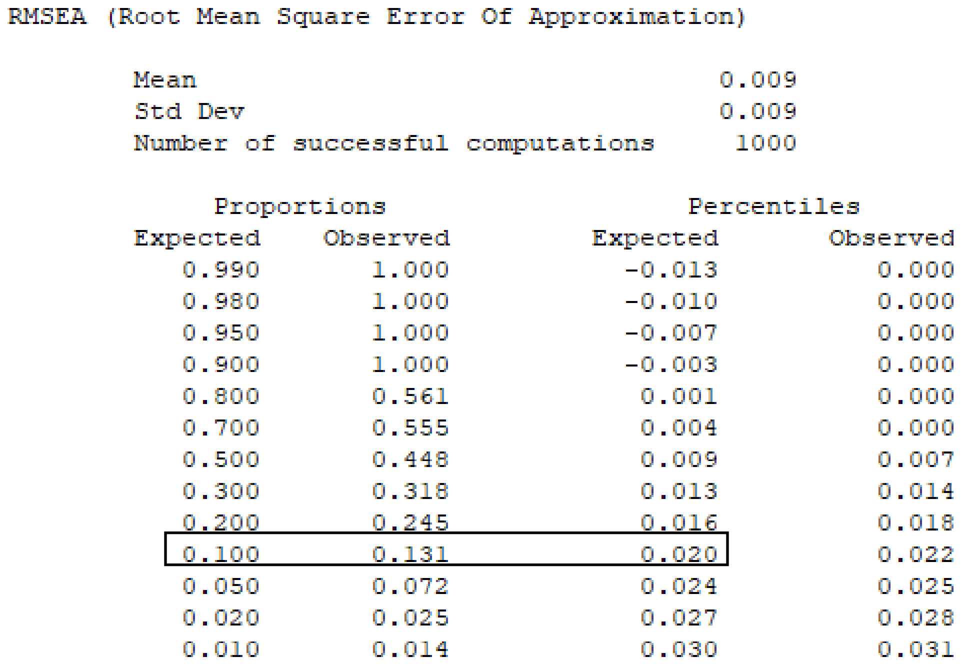
Step 5b: Assess a 10% Threshold

If more than 5% of correct models are rejected, the next goal is to assess a 10% threshold for rejecting correct models and a 90% threshold for rejecting misspecified models. To do so, we determine whether the 90th percentile of the true model distribution is below the 10% percentile of the misspecified model distribution. To follow this strategy for the current example, no code or simulations need to be rerun. Instead, we simply need to reference different rows of the previous simulations. In the misspecified model simulation, the 10th percentile of the distribution appears in the Expected Percentile column of the 0.900 Expected Proportion row, which is 0.025 for RMSEA. We do not need to look up this value for SRMR since both error rates were already 5% or below.



```
RMSEA (Root Mean Square Error Of Approximation)

        Mean                                    0.035
        Std Dev                                 0.007
        Number of successful computations        1000

             Proportions                   Percentiles
        Expected    Observed          Expected       Observed
           0.990       0.977             0.018          0.014
           0.980       0.970             0.020          0.016
           0.950       0.947             0.023          0.022
           0.900       0.904             0.025          0.026
           0.800       0.822             0.029          0.029
           0.700       0.726             0.031          0.032
           0.500       0.552             0.035          0.035
           0.300       0.305             0.039          0.039
           0.200       0.184             0.041          0.040
           0.100       0.078             0.044          0.043
           0.050       0.031             0.047          0.046
           0.020       0.008             0.050          0.048
           0.010       0.004             0.052          0.049
```

The 90th percentile of the correct distribution appears in in the Expected Percentile column of the 0.100 Expected Proportion row, which is 0.020. Now, the 90th percentile of the correct model distribution is below the 10th percentile of the misspecified model distribution so using a cutoff of 0.025 is unambiguous for the specified threshold.



```
RMSEA (Root Mean Square Error Of Approximation)

        Mean                                0.009
        Std Dev                             0.009
        Number of successful computations    1000

             Proportions                  Percentiles
        Expected    Observed         Expected        Observed
           0.990       1.000           -0.013           0.000
           0.980       1.000           -0.010           0.000
           0.950       1.000           -0.007           0.000
           0.900       1.000           -0.003           0.000
           0.800       0.561            0.001           0.000
           0.700       0.555            0.004           0.000
           0.500       0.448            0.009           0.007
           0.300       0.318            0.013           0.014
           0.200       0.245            0.016           0.018
           0.100       0.131            0.020           0.022
           0.050       0.072            0.024           0.025
           0.020       0.025            0.027           0.028
           0.010       0.014            0.030           0.031
```

Therefore, the dynamic RMSEA cutoff is 0.025 and the dynamic cutoff for SRMR is 0.034. There is a .001 difference in the RMSEA and SRMR cutoffs compared to what is reported in the paper since there is some Monte Carlo error between different randomly generated sets of data.