DISSERTATION

Why one size does not fit all: Evaluating the validity of fixed cutoffs for model fit indices and developing new alternatives

---

Katharina Mechthild Maria Groskurth

University of Mannheim

Inaugural dissertation submitted in partial fulfillment of the requirements for the degree Doctor of Social Sciences in the Graduate School of Economic and Social Sciences at the University of Mannheim

December 7th, 2022

Supervisors:
Prof. Dr. Thorsten Meiser
Dr. Clemens Lechner

Dean of the Faculty of the School of Social Sciences:
Prof. Dr. Michael Diehl

Thesis Evaluators:
Prof. Dr. Thorsten Meiser
Prof. Dr. Edgar Erdfelder
Prof. Dr. Eldad Davidov

Date of Defense:
March 22nd, 2023

To my parents.

Table of Contents

Acknowledgments

This doctoral thesis would not have been possible without so many people's advice, support, and motivation—I wholeheartedly want to thank them all.

I would like to thank Clemens Lechner for his guidance, support, and enthusiasm. He always motivated and inspired me while giving me the time and freedom to develop my own ideas. I would also like to thank Thorsten Meiser for shaping, discussing, and challenging my research. He always encouraged me and truly extended my knowledge of psychometrics. I am grateful for his focus on necessary details while making pragmatic decisions. I am grateful for the continuous support of Matthias Bluemke, who aroused my enthusiasm for psychometrics—without him, this project would probably never have started. I learned so much from them—they all sustainably shaped my thinking.

I would also like to thank Edgar Erdfelder for discussing, challenging, and shaping my initial ideas in several CDSS workshops. I am grateful to Eldad Davidov, who guided me through his research and awakened my enthusiasm for measurement invariance. Many thanks to all dissertation committee members for taking the time and effort to evaluate this thesis.

I also want to thank all my colleagues and fellow students from the CDSS, GESIS, and the chair of Research Methods and Psychological Assessment at the University of Mannheim for all the stimulating discussions on various research topics and engaging feedback on preliminary results. It was a joy working and studying with them. I especially want to thank Nivedita Bhakta for working together on the second manuscript. I am grateful to all my colleagues and fellow students who became friends over the years. They made the time special.

I am truly grateful to my dearest friends, family, and David for always being at my side.

Abstract

Model evaluation is a central topic in structural equation modeling. Researchers commonly evaluate whether a model fits their data with fixed cutoffs for fit indices (e.g., CFI ≥ .95 for good model fit). Researchers apply the same fixed cutoffs in various empirical settings, even if those settings diverge from the simulated scenarios the cutoffs originated from. In this thesis, I outlined why this one-size-fits-all usage of fixed cutoffs is invalid and proposed alternative approaches for model evaluation.

In the first manuscript, I investigated the fit indices' sensitivity to misspecification in confirmatory factor models and their susceptibility to various model, data, and estimation characteristics in a large-scale simulation. Several characteristics (especially the factor correlation and the type of estimator) strongly influenced fit indices. They interacted in complex ways implying that cutoffs for fit indices are only valid for the context from which they originate. Based on the large-scale simulation, I developed two approaches to generate cutoffs tailored to empirical settings resembling the simulated scenarios. Researchers can read out scenario-specific cutoffs from large-scale tables. Alternatively, researchers can use regression formulae and plug in characteristics of interest to calculate scenario-specific cutoffs.

In the second manuscript, I reviewed and discussed all approaches to tailored cutoffs proposed in the literature. Based on the literature review, I developed a new approach that combines a Monte Carlo simulation with receiver operating characteristic (ROC) analysis. The so-called simulation-cum-ROC approach generates cutoffs for various fit indices tailored to the setting of interest. Uniquely, it guides researchers on which fit index best evaluates whether the model fits the data (or not) in the setting of interest.

In the third manuscript, I focused on a specific area in which binary decisions on model fit abound: measurement invariance testing. I developed effect size measures, so-called Measurement Invariance Violation Indices (MIVIs), for items and item sets that continuously quantify non-invariance (i.e., misfit) if identified by binary cutoffs. MIVIs quantify non-invariant parameter differences in units of the latent variable's pooled standard deviation.

This thesis demonstrated that cutoffs must be tailored to the setting of interest for valid model evaluation. I outlined and developed various approaches that differ in their flexibility to obtain scenario-specific cutoffs. Newly developed effect size measures allow researchers to continuously quantify misfit (i.e., non-invariance) in addition to cutoffs following the binary fit-misfit logic. This research is a step towards more valid model evaluation techniques.

Manuscripts

Manuscripts of This Thesis

The following three manuscripts are part of the cumulative thesis:

MANUSCRIPT I

Groskurth, K., Bluemke, M., & Lechner, C. M. (2022a). *Why we need to abandon fixed cutoffs for goodness-of-fit indices: A thorough simulation and possible solutions.* Invited revision at *Behavior Research Methods*.

MANUSCRIPT II

Groskurth, K., Bhaktha, N., & Lechner, C. M. (2022). *Making model judgments ROC(K)-solid: Tailored cutoffs for fit indices through simulation and ROC analysis in structural equation modeling.* Invited revision at *Psychological Methods*.

MANUSCRIPT III

Groskurth, K., Bluemke, M., & Lechner, C. M. (2022b). *Measurement Invariance Violation Indices (MIVIs): Effect sizes for non-invariance of items and item sets.* Manuscript in preparation.

The overarching goal of the thesis was to make model evaluation more valid in the structural equation modeling context, mostly conducted via fixed cutoffs for fit indices. In the synopsis, I first highlight the importance of model evaluation. Further, I contrast the standard practice of model evaluation (done via fixed cutoffs for fit indices) with its methodological criticism (claiming that those cutoffs are invalid in most instances), underlining the motivation of my thesis. I outline possible reasons for the gap between the current practice and methodological recommendations for model evaluation. Additionally, I outline potential solutions that might help to close the gap in reference to the three manuscripts of the present thesis. Then, I present the three manuscripts' main findings and newly developed approaches. Lastly, I discuss the three manuscripts, reflect on their contribution to the methodology and standard practice of model evaluation, and elaborate on directions for future research. I appended all three manuscripts to this thesis.

Other Publications

I have published five manuscripts during my time at the *Graduate School of Economic and Social Sciences* and *GESIS – Leibniz Institute for the Social Sciences* that can be subsumed under the field of model evaluation but are not part of the cumulative thesis. They are listed below:

Groskurth, K., Nießen, D., Rammstedt, B., & Lechner, C. M. (2022). The Impulsive Behavior Short Scale–8 (I-8): A comprehensive validation of the English-language adaptation. *PLoS ONE, 17*(9), Article e0273801. https://doi.org/10.1371/journal.pone.0273801

Nießen, D., Schmidt, I., Groskurth, K., Kemper, C. J., Rammstedt, B., & Lechner, C. M. (2022). The Internal–External Locus of Control Short Scale–4 (IE-4): A comprehensive validation of the English-language adaptation. *PLoS ONE, 17*(7), Article e0271289. https://doi.org/10.1371/journal.pone.0271289

Nießen, D., Groskurth, K., Kemper, C. J., Rammstedt, B., & Lechner, C. M. (2022). The Optimism–Pessimism Short Scale–2 (SOP2): A comprehensive validation of the English-language adaptation. *Measurement Instruments for the Social Sciences, 4*, Article 1. https://doi.org/10.1186/s42409-021-00027-6

Groskurth, K., Nießen, D., Rammstedt, B., & Lechner, C. M. (2021). An English-language adaptation and validation of the Political Efficacy Short Scale (PESS). *Measurement Instruments for the Social Sciences, 3*, Article 1. https://doi.org/10.1186/s42409-020-00018-z

Lechner, C. M., Bhaktha, N., Groskurth, K., & Bluemke, M. (2021). Why ability point estimates can be pointless: A primer on using skill measures from large-scale assessments in secondary analyses. *Measurement Instruments for the Social Sciences, 3*, Article 2. https://doi.org/10.1186/s42409-020-00020-5

1 Introduction

1.1 The Importance of Model Evaluation

Nearly every science rooted in (post)positivism has an a priori mental representation or theoretical model of its entities and their relationships (i.e., the causal mechanisms; Della Porta & Keating, 2008). General examples of such theoretical models are Newton's law of universal gravitation (1687/1999), which explains the phenomenon of gravity on earth, and Durkheim's study of suicide (1897/1951), stating that socially integrated individuals are less likely to commit suicide. In psychology, theoretical models frequently include unobservable constructs (Borsboom et al., 2003), such as self-esteem, personality, or religious belief.

A fundamental principle of quantitative science is to corroborate or falsify theoretical models in empirical reality (Popper, 2002). Therefore, a theoretical model must be testable. In psychology, specific (and frequently multiple) items represent an unobservable construct (Borsboom et al., 2003). An individual's manifestation of the construct determines their answers to the items. The relation between a psychological construct and its items is a theoretical model by itself (i.e., the so-called measurement model, e.g., Borsboom et al., 2003), easily tested in empirical reality. Individuals need to provide answers to the items, and covariances between them make it possible to estimate the latent, unobserved variable representing the psychological construct (e.g., Bollen, 1989). For instance, items such as "I take a positive attitude toward myself" and "I feel that I have a number of good qualities" should operationalize self-esteem (Rosenberg, 1965).

Although a particular theoretical representation might seem logical or straightforward, empirical tests might suggest that the theoretical model is wrong (and potentially more or less complex than theoretically assumed, e.g., Popper, 2002).[1] It is essential, if not a central fundament of quantitative science, to evaluate theoretical models in empirical reality. This thesis's focus is the central and ubiquitous topic of model evaluation in the context of psychological constructs. I focus on structural equation models, particularly confirmatory factor analysis models, that originate from the classical test theory as opposed to models from

[1] Model tests in empirical reality might incorrectly suggest that a theoretical model is wrong—for example, as the tools to evaluate those models are fallible, where I elaborate on later.

the item response theory (for the different interpretations of models from the two theories, see Borsboom et al., 2003).

1.2 The Standard Practice of Model Evaluation

When using structural equation modeling (or, as it is the specific focus, confirmatory factor analysis), researchers typically evaluate theoretical models tested in the empirical reality with fit indices. Fit indices are essentially effect sizes for misfit; they should quantify how well a model (e.g., the theoretical model) approximates data (e.g., empirical data). As indicators of global (mis)fit, they evaluate the fit of the overall model (for an overview, see Schermelleh-Engel et al., 2003). Commonly used fit indices are the confirmatory fit index (CFI; Bentler, 1990), the root mean squared error of approximation (RMSEA; Steiger, 1990; see also Chen, 2007), and the standardized root mean squared residual (SRMR; Bentler, 1995; Hu & Bentler, 1999).

The common practice of model evaluation targets a binary fit-misfit decision instead of a continuous quantification of misfit (e.g., Jackson et al., 2009; McNeish & Wolf, 2021). Thus, researchers apply so-called cutoffs (i.e., decision thresholds) for fit indices. Commonly used cutoffs for fit indices indicating a well-fitting model are CFI $\geq$ .950, RMSEA $\leq$ .060, and SRMR $\leq$ .080 (Hu & Bentler, 1999). Not only overall but also nested model fit is evaluated similarly with common cutoffs for fit indices. Examining the comparability of models across groups is a prominent case of nested model fit testing (i.e., measurement invariance testing, e.g., Davidov et al., 2014; Millsap, 2011; Steenkamp & Baumgartner, 1998). Measurement invariance testing proceeds in a sequential fashion (Chen, 2007): When testing the equality of factor loadings in a model fitted across groups (i.e., metric invariance) with a reasonable sample size (total $N > 300$, equal cross-group sample sizes), researchers should reject the model with equal factor loadings if $\Delta\text{CFI} \leq -.010$ in combination with $\Delta\text{RMSEA} \geq .015$ or $\Delta\text{SRMR} \geq .030$ in comparison to the model without equal factor loadings (i.e., configural invariance). The same strategy and cutoffs apply to testing the equality of item intercepts (i.e., scalar invariance) or residual variances across groups (i.e., uniqueness invariance)—with the only exception that researchers should use $\Delta\text{SRMR} \geq .010$ as a cutoff.[1] For unequal cross-group sample sizes and a small total sample size (total $N \leq 300$), researchers should use less stringent cutoffs (i.e., reducing the absolute above-stated values by .005).

[1] SRMR turned out to be more sensitive to metric non-invariance instead of scalar or uniqueness non-invariance (Chen, 2007).

But where do these cutoffs come from? Cutoffs commonly originate from simulation studies (e.g., Chen, 2007; Hu & Bentler, 1999).[1] In these simulation studies, researchers specify the "true" model, which they never know in empirical reality. From this true model, more precisely called the data-generating or population model, they repeatedly sample data via a Monte Carlo simulation (e.g., Mooney, 1997). Then, a model structurally identical to the population model (i.e., the analysis model) is fit to each sampled data. Researchers save values of fit indices each time they fit the analysis model to the sampled data. This way, they obtain distributions of fit indices. A specific value of the resulting fit index distribution can serve as a cutoff. If higher values point to worse fit (e.g., RMSEA or SRMR), the fit index value at the 95th percentile is a commonly chosen cutoff. If lower values point to worse fit (e.g., CFI), the fit index value at the 5th percentile is a commonly chosen cutoff. Those cutoffs have a 5% Type I error rate of falsely rejecting a correctly specified model. Researchers can also include a condition where the analysis model is still the same, but the population model differs. The analysis model severely diverges from that population model (i.e., the analysis model is misspecified). Misspecification (or "severe divergence") refers to the failure of the analysis model to capture relevant complexities of the population model (e.g., specific parameters or factors). Thus, defining a model as misspecified is a subjective but considerate decision. Researchers then simulate data from the population model and fit the analysis model to that data. Including a condition with a misspecified model allows them to evaluate Type II error rates (i.e., how many misspecified models the cutoff wrongly accepts) in addition to Type I error rates. Type II error rates should be low (i.e., close to 0%) at the previously chosen cutoff.

Thus, when a theoretical model is tested in empirical reality, an analysis model, representing the theoretical model, is fit to empirical data. Once the model is fit to the data, standard statistical programs already provide several fit indices (such as Mplus, Muthén & Muthén, 1998-2017, or the lavaan package in R, Rosseel, 2012). If empirical values of fit indices pass the cutoffs, originating from simulation studies (e.g., Chen, 2007; Hu & Bentler, 1999), the researcher accepts the model. Empirical evidence favors the model as it fits to data

[1] Simulations generating distributions for fit indices are especially useful if the fit index distribution is unknown. The χ^2 test statistic follows a χ^2 distribution (if distributional assumptions hold, e.g., Schermelleh-Engel et al., 2003). If fit indices incorporate the χ^2 test statistic (e.g., RMSEA), their distributions can be inferred (e.g., Moshagen & Erdfelder, 2016). If fit indices do not incorporate the χ^2 test statistic (e.g., SRMR), their distributions remain unknown. Simulations allow to generate fit index distributions if fit indices do not follow a known distribution or if assumptions are violated so that fit indices do not follow a known distribution anymore (e.g., the χ^2 test statistic does not follow a χ^2 distribution anymore, if the items are not multivariate normal).

generated from a still unknown population model. More technically, empirical evidence favors the null hypothesis, H_0, stating that the model is identical to the population model (Neyman & Pearson, 1928, 1933; see also Biau et al., 2010; Moshagen & Erdfelder, 2016; Perezgonzalez, 2015). The model is assumed to be correctly specified. If empirical values of fit indices fail the cutoffs, the researcher rejects the model. Empirical evidence does not favor the model as it does not fit to data generated from a still unknown population model. More technically, empirical evidence favors the alternative hypothesis, H_1, stating that the model is not identical to the population model to a specific degree of intolerable misspecification (Neyman & Pearson, 1928, 1933; see also Biau et al., 2010; Moshagen & Erdfelder, 2016; Perezgonzalez, 2015). The model is assumed to be misspecified.[1]

Crucially, researchers use cutoffs, derived once in a simulation, across diverse empirical settings (e.g., Jackson et al., 2009; McNeish & Wolf, 2021). Thus, the usage of cutoffs is independent of the specific empirical setting, which is why I call them fixed cutoffs. Several textbooks (e.g., Kline, 2016) promote this common way of model evaluation. It is ubiquitously used, easy to apply, follows an objective principle, and is accepted effectively by all journals (e.g., Jackson et al., 2009).

1.3 Methodological Criticism of the Standard Practice of Model Evaluation

Evaluating model fit using fixed cutoffs across various empirical settings sounds too good to be true, and it certainly is. Many methodologists heavily criticize this standard practice of model evaluation (e.g., Heene et al., 2011; Markland, 2007; Marsh et al., 2004; McNeish & Wolf, 2021; Niemand & Mai, 2018; Nye & Drasgow, 2011a). The core of their criticism is that the one-size-fits-all logic of fixed cutoffs is fallible: Values of fit indices, and accordingly their cutoffs, depend not only on misfit but also on other characteristics of the empirical setting. Values of fit indices vary with different model, data, and estimation characteristics (for an overview, see Groskurth, Bluemke, & Lechner, 2022a; Niemand & Mai, 2018). For instance, it has been shown in many studies that model misspecification is harder to detect with lower factor loadings (Beierl et al., 2018; Groskurth, Bluemke, & Lechner, 2022a; Hancock &

[1] Methodological recommendations partly exist on which fit indices researchers should rely on (e.g., Chen, 2007; Mai et al., 2021). In practice, the preference for fit indices remains vague; researchers rely on all fit indices equally (e.g., Jackson et al., 2009).

Mueller, 2011; Heene et al., 2011, McNeish et al., 2018; Shi et al., 2018, 2019; Shi & Maydeu-Olivares, 2020).

It follows that fixed cutoffs are only valid for empirical settings close to the simulated scenarios from which the cutoffs originate. To derive cutoffs via a simulation, researchers cannot include all possible model, data, and estimation characteristics that might occur in the empirical reality but instead focus on a tiny subset of them (e.g., sample size and response distribution in Hu & Bentler, 1999). Thus, cutoffs are derived based on particular simulation scenarios but applied in empirical settings that may be considerably different from these simulation scenarios.

Applying the same fixed cutoffs to diverse empirical settings, which differ from the initial simulated scenarios, can lead to wrong conclusions in model evaluation (e.g., McNeish & Wolf, 2021). Theoretical models might be rejected (or accepted) in empirical settings not because they are invalid (or valid) models of psychological constructs but because the methods to evaluate these models are invalid. Relying on decision criteria that can lead to wrong conclusions undoubtedly harms moving forward in science.

1.4 Closing the Gap Between the Standard Practice of Model Evaluation and Its Methodological Criticism

1.4.1 Defining the Gap

Given the large gap between the standard practice of model evaluation and related methodological recommendations, why has the standard practice of model evaluation not substantially changed yet? I can only hypothesize about possible answers to this central question. Three reasons crossed my mind why fixed cutoffs have not yet been replaced: One reason might be that it is not entirely clear to applied researchers how invalid fixed cutoffs are. Another reason might be that there are no easy-to-apply alternatives to fixed cutoffs. A final reason might be that fixed cutoffs seem relatively tolerant. Researchers might be afraid that their model will fail when using alternatives (which can hamper publication as a result; Flora, 2020). In addition to binary cutoffs, a new, continuous view on (mis)fit is potentially needed. I elaborate on the three reasons in the following.

1.4.2 Understanding the Gap

1.4.2.1 Is There an Unawareness About the Limited Generalizability of Fixed Cutoffs?

Studies evaluating the sensitivity of fit indices to model misspecification and their susceptibility to model, data, and estimation characteristics commonly focus on a single influential characteristic (e.g., factor loadings; Heene et al., 2011). This might lead to the assumption that one can track those influences and explain the absolute fit index value one observes. For instance, one might argue that the model fits well, although one observes a "bad" fit index value. One might explain that the "bad" fit index value results from low factor loadings but not from model misfit (e.g., Heene et al., 2011). Can you be sure that the fit index value became "bad" because of low factor loadings but not model misfit? Or might it indeed become "bad" because of the model misfit, as, for instance, the chosen estimator leads to a greater fit index sensitivity to misfit (e.g., Xia & Yang, 2019)? This illustration shows that influences of model, data, and estimation characteristics on fit indices are multitude and might interact in complex ways.

Previous research has identified several characteristics influencing fit indices. Those characteristics are the sample size (e.g., DiStefano et al., 2019), type of estimator (e.g., Xia & Yang, 2019), number of items (e.g., Kenny & McCoach, 2003), number and distribution of response options (e.g., Xia & Yang, 2018), magnitude of factor loadings (e.g., Heene et al., 2011), and factor correlation (e.g., Beauducel & Wittmann, 2005).

However, there has not been any integrative study that includes all previously known characteristics influencing fit indices. How fit indices react to the joint influences of all those characteristics remains unclear. It remains unknown how those influences attenuate or aggravate when investigated in tandem. Investigating several influences in tandem also provides a holistic picture of the relative importance of those influences. Such a holistic study may underline even more strongly than previous ones that fit indices depend on several characteristics that may interact in complex ways. It may push applied researchers toward abandoning fixed cutoffs for fit indices. Consequently, one goal of this thesis was to replicate and extend prior knowledge on the fit indices' sensitivity to misspecification and their susceptibility to several model, data, and estimation characteristics by investigating many characteristics in tandem.

The first manuscript (Groskurth, Bluemke, & Lechner, 2022a) aimed at this goal: It presented the most in-depth simulation study on fit indices for confirmatory factor analysis

models to date. By including several characteristics into one simulation study (that were primarily studied in isolation before), I replicated knowledge: Fit indices were not only sensitive to model misspecification (as they should be) but also susceptible to a wide range of model, data, and estimation characteristics (e.g., type of estimator and loading magnitude). By considering several characteristics in tandem, I also unraveled patterns that would otherwise remain unnoticed: The factor correlation (of analysis models with unmodeled cross-loadings that were present in the population) moderated several effects (e.g., the influence of the type of estimator on fit indices). Model, data, and estimation characteristics interacted in complex and highly unpredictable ways.

Recognizing the susceptibility of fit indices to several model, data, and estimation characteristics—and, in turn, the invalidity of fixed cutoffs when applied to empirical settings different from the initial simulated scenarios—is just one part that may lead to a change of the status quo in model evaluation. Methodologists must provide easy-to-apply alternatives to fixed cutoffs to help researchers entirely abandon fixed cutoffs.

1.4.2.2 Are There No Alternatives to Fixed Cutoffs?

A recent suggestion for a more valid model evaluation was to generate tailored cutoffs as an alternative to fixed cutoffs (e.g., Millsap, 2013). These cutoffs are generated specifically for the setting of interest. They are, thus, a natural solution to the pressing problems of fixed cutoffs that methodologists have pointed out.

Currently, there are five principal approaches to generating tailored cutoffs that differ in their flexibility to accommodate different fit indices and settings of interest. I ordered them from least to most flexible. The first approach is the table-based one. Similar to looking up critical values of *z*-scores or *t*-statistics, researchers select the cutoff that best matches their empirical setting from scenario-specific cutoff tables (Groskurth, Bluemke, & Lechner, 2022a). A second approach is χ^2 distribution-based. Relying on statistical assumptions of the χ^2 distribution without and with misspecification, researchers generate tailored cutoffs for fit indices that include the χ^2 test statistic (Moshagen & Erdfelder, 2016). A third approach is regression-based. Researchers generate tailored cutoffs based on meta-regression results from a prior simulation study (Nye & Drasgow, 2011a; see also Groskurth, Bluemke, & Lechner, 2022a). In the simulation-based approach, the fourth approach, researchers fit the model of interest to data repeatedly sampled from a known population model. They generate tailored cutoffs based on the resulting values of fit indices (McNeish & Wolf, 2021; Millsap, 2013;

Niemand & Mai, 2018; Pornprasertmanit, 2014). In essence, the simulation-based approach is a parametric bootstrap approach. It samples (i.e., simulates) data based on parameters (i.e., from the population model). Differently, the non-parametric bootstrap approach samples data based on transformed empirical data as a fifth approach. In essence, researchers fit the model of interest to data repeatedly sampled from transformed empirical data. Crucially, they transform empirical data as if the model has generated it. Researchers generate tailored cutoffs based on the resulting values of fit indices (Bollen & Stine, 1992; Kim & Millsap, 2014).

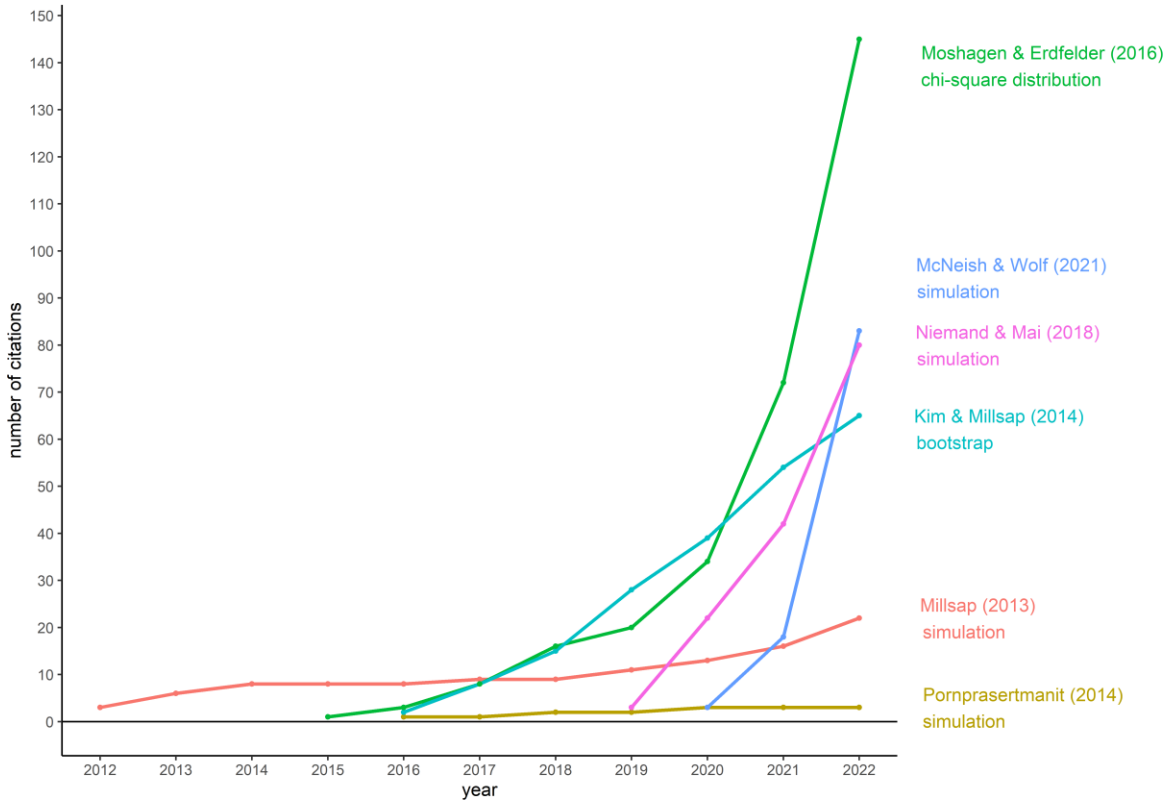
The above summary of tailored cutoff approaches illuminated the emerging status of this strand of research: There has been only a couple of approaches, which mainly began to emerge one decade ago (i.e., Bollen & Stine, 1992; Kim & Millsap, 2014; Millsap, 2013; Pornprasertmanit, 2014), with the recent emergence of new approaches around 2020 (e.g., McNeish & Wolf, 2021; Niemand & Mai, 2018). In the history of model evaluation, approaches of tailored cutoffs for fit indices can be considered relatively new. To analyze the interest in tailored cutoff approaches, I quantified the cumulative increase of citations in GoogleScholar across the years since seminal articles were published. I included the following seminal articles: Kim and Millsap (2014, bootstrap), McNeish and Wolf (2021, simulation), Millsap (2013, simulation), Moshagen and Erdfelder (2016, χ^2 distribution), Niemand and Mai (2018, simulation), and Pornprasertmanit (2014, simulation).[1] I did not include Nye and Drasgow (2011a, regression) and Groskurth, Bluemke, and Lechner (2022a, table and regression) as their focus was on the susceptibility of fit indices to the setting of interest. I also omitted Bollen and Stine (1992, bootstrap). That article had around 1,600 citations at the time of this writing. Thus, it raised lots of interest; much more than any other article included (receiving only up to around 150 citations)—which is why I considered it an outlier here. Figure 1.1 shows the interest in seminal articles on tailored cutoff approaches quantified through the cumulative increase of citations in GoogleScholar.

For most publications (i.e., McNeish & Wolf, 2021; Moshagen & Erdfelder, 2016; Niemand & Mai, 2018), the interest has exponentially increased since 2020, although publications dated back some years ago (i.e., Moshagen & Erdfelder, 2016). Likewise, interest in Millsap's (2013) simulation-based approach has started rising recently. Only Kim and Millsap's (2014) bootstrap approach to tailored cutoffs has raised interest continuously since

[1] I initially intended to analyze the number of citations of seminal articles in PsycIndex that include keywords such as "dynamic fit index" or "tailored cutoffs." Because these keywords only led to one article (McNeish & Wolf, 2021), I decided to analyze the number of citations of seminal articles in GoogleScholar.

its publication date. Pornprasertmanit's (2014) doctoral thesis, including the simulation-based approach, is the only publication developing a tailored cutoff approach that has no rise in interest yet. In sum, researchers recently showed a willingness to discuss, adopt, and expand tailored cutoffs as an alternative to fixed cutoffs.

Figure 1.1: *Cumulative Increase of Citations for Seminal Tailored Cutoff Articles in GoogleScholar*



Note. I created the figure on November 25th, 2022. Thus, data for 2022 does not include the entire year.

Many methodologists provide automated tools for their tailored cutoff approaches, such as R functions, shiny apps, and websites (Jak et al., 2021; Jobst et al., 2021; Mai et al., 2021; McNeish & Wolf, 2021; Moshagen & Erdfelder, 2016; Niemand & Mai, 2018; Pornprasertmanit et al., 2021; Schmalbach et al., 2019). These tools enable researchers to derive tailored cutoffs easily. Researchers must not have any experience in, for instance, simulating data to generate tailored cutoffs. They must plug in their model, data, and estimation characteristics on the website/function/shiny app, which automatically does the calculation in the background.

My former elaborations may let readers think there are multiple (or even enough) approaches to start with the integration of tailored cutoffs into the common practice of model evaluation. Thanks to the automated tools, the approaches seem easy to apply. Although tailored cutoff approaches get more and more recognition (as quantified through the number of citations, see Figure 1.1), using fixed cutoffs for model evaluation is still the dominant practice. Applied research rarely uses tailored cutoffs. This is evident from the absolute number of citations in Figure 1.1: Each seminal article on tailored cutoff approaches reached no more than 150 citations in GoogleScholar since their publication, except for Bollen and Stine's (1992) article reaching around 1,600 citations. Hu and Bentler's (1999) seminal article on fixed cutoffs has already received around 9,500 additional citations in 2022 alone (and more than 97,000 citations since its publication) in GoogleScholar at the time of this writing. So, what could convince researchers to apply tailored cutoffs instead of fixed cutoffs?

To my knowledge, there has not yet been any study that summarizes and compares all approaches to tailored cutoffs. Such integration of approaches could be beneficial for applied researchers: It makes aware of currently available approaches to tailored cutoffs. Further, researchers learn about those approaches and their advantages, disadvantages, and limitations in a very condensed way—and they learn in which empirical settings the different approaches are helpful. Integration of approaches enables researchers to select the best approach for their empirical setting. Further, integration of approaches helps to move forward methodologically: Discussing which approaches of tailored cutoffs are developed so far, as well as the possibilities and limits of their applicability, helps identifying research gaps for improvement of approaches or suggesting new approaches. Thus, one goal of this thesis was to summarize and evaluate tailored cutoff approaches which was realized in the second manuscript (Groskurth, Bhaktha, & Lechner, 2022).

Another goal was to build on and extend existing approaches to tailored cutoffs. In the first manuscript (Groskurth, Bluemke, & Lechner, 2022a), I suggested table- and regression-based approaches to tailored cutoffs for a wider variety of scenarios than previously suggested (e.g., Nye & Drasgow, 2011a). In the second manuscript (Groskurth, Bhaktha, & Lechner, 2022), I developed a tailored cutoff approach that includes a feature not included in existing approaches: It provides guidance on which fit index to rely on in the setting of interest. The so-called simulation-cum-ROC approach combines a Monte Carlo simulation with receiver operating characteristic (ROC) analysis. Notably, the approach generates tailored cutoffs while identifying well-performing fit indices in the given setting. Thus, it evaluates the fit indices' performance in discriminating between correctly specified and misspecified models.

So far, I have outlined (and developed) different approaches to tailored cutoffs and thoroughly investigated the susceptibility of fit indices. If the problem of fixed cutoffs for fit indices is made entirely clear, and easy-to-apply alternatives to fixed cutoffs are well-known, what might still hinder applied researchers from changing the status quo of model evaluation? Researchers might be concerned that their model will fail when applying tailored instead of fixed cutoffs. From recent articles (e.g., Groskurth, Bhaktha, & Lechner, 2022; McNeish & Wolf, 2021), tailored cutoffs seem stricter, that is, more inclined to reject models than fixed cutoffs. It is well-known (as part of the replication crisis) that manuscripts have a higher chance of getting published in most journals if the models of interest are accepted (e.g., Flora, 2020). Thus, to entirely change the status quo of model evaluation, the fear of model failure must shrink—potentially through quantifying the impact of (mis)fit in addition to defining (mis)fit in a binary way.

1.4.2.3 Is the Quantification of Misfit Needed?

Current examples in articles on tailored cutoffs (e.g., Groskurth, Bhaktha, & Lechner, 2022; McNeish & Wolf, 2021) could quickly raise the impression that tailored cutoffs are more inclined to reject models than fixed cutoffs.[1] For instance, in all of McNeish and Wolf's (2021) empirical examples evaluating either one-, two-, or multi-factor models, tailored cutoffs (generated to detect minor levels of misspecification) for CFI were not below .962, for RMSEA

[1] Mainly, tailored cutoffs are derived from exactly fitting models (e.g., Bollen & Stine, 1992; McNeish & Wolf, 2021; Nye & Drasgow, 2011a; cf. Kim & Millsap, 2014; Millsap, 2013; Pornprasertmanit, 2014). Cutoffs are generated under the premise that analysis models are only acceptable if they capture all parameters from the population model (Millsap, 2007). I discussed approximately fitting models, as an alternative, in more depth in Chapter 3 of this thesis as well as in the second manuscript (Groskurth, Bhaktha, & Lechner, 2022).

not above .076, for SRMR not above .037. Similarly, in Groskurth, Bhaktha, and Lechner's (2022) examples evaluating two-factor models, tailored cutoffs for CFI were not below .984, for RMSEA not above .047, and for SRMR not above .031. In both studies (Groskurth, Bhaktha, & Lechner, 2022; McNeish & Wolf, 2021), tailored cutoffs were generally stricter than Hu and Bentler's (1999) fixed cutoffs (i.e., CFI around .950, RMSEA around .060, SRMR around .080).

If the model of interest fails (i.e., must be rejected), researchers may be unsure how to proceed. There are many ways to handle model misfit: The most common way is to employ local modifications by freely estimating certain misspecified parameters (identified via modification indices and the expected parameter change, e.g., Kaplan, 1991; Podsakoff et al., 2003; Saris et al. 2009; Whittaker, 2012). Alternatively, researchers theoretically define a completely new model. All these approaches mainly focus on refining or changing the existing model. Crucially, they all dismiss an essential step that could lead to the conclusion of retaining the model: They do not thoroughly analyze misfit and its potential implication. The central question remains unanswered: Is misfit large enough to be detrimental for further analysis?

For such further analysis, observed descriptive statistics (e.g., mean scores) or observed (co)variances (e.g., regression coefficients) are often of interest (McNeish & Wolf, 2020; Musil et al., 1998; Widaman & Revelle, 2022)—especially in a relational way (e.g., Chen, 2008): The core of cross-cultural research is to compare group statistics or predictors of various constructs (such as self-esteem) across cultures, countries, or ethnic groups (Berry et al., 2002; e.g., Chen, 2008). In clinical research, descriptive statistics (e.g., the incidence of depressive disorders) of the treatment group (e.g., who went through a prevention program) compared to the control group (e.g., without prevention program) are regularly of interest (e.g., Cuijpers et al., 2021). Similarly, common questions in developmental research are concerned with investigating differences (e.g., in personality) across cohort or age groups (e.g., Roberts et al., 2006). I could extend this list of examples to more areas within psychology (and, more generally, the social sciences), underlining that many studies aim at group comparisons. Such comparisons require measurement invariance (i.e., equal item-to-factor structure and model parameters across groups); non-invariance (i.e., unequal item-to-factor structure or model parameters across groups) disturbs such comparisons (e.g., Chen, 2008; Steinmetz, 2013). The central question remains: Is this misfit (i.e., non-invariance) large enough to be detrimental for further analysis (i.e., group comparisons)?

So-called effect size measures continuously quantify the extent of non-invariance present in the model.[1] In particular, they quantify how strongly specific parameters differ across groups. Effect size measures for measurement invariance tests abound in the item response theory framework (Meade, 2010). Within the classical test theory framework, researchers have started developing effect size measures for measurement invariance tests roughly a decade ago (Millsap & Olivera-Aguilar, 2012; Nye & Drasgow, 2011b; Oberski, 2014; Oberski et al., 2015; Pornprasertmanit, 2022). However, some existing effect size measures are complex to apply, requiring extra statistical packages (Dueber, 2019; Nye & Drasgow, 2011b; Oberski, 2014). Other existing effect size measures focus only on non-invariant parameter differences of single items (Millsap & Olivera-Aguilar, 2012; Pornprasertmanit, 2022; likewise, the modification index and expected parameter change) and not additionally on those of the complete item set, making it unable to investigate compensation or aggregation effects. Non-invariant parameter differences can compensate for each other if the pattern of non-invariance is mixed (Chen, 2008). The pattern of non-invariance is mixed if one group has higher parameter values on some non-invariant items but lower on others compared to another group. Non-invariant parameter differences can sum to zero so that, for instance, mean score differences across groups remain unbiased despite non-invariance being present. Differently, the pattern of non-invariance is uniform if one group always has higher parameter values on all non-invariant items than another group (Chen, 2008). Thus, non-invariant parameter differences aggregate so that, for instance, non-invariance impacts mean score differences across groups.

Thus, another goal of this thesis was to derive easy-to-apply effect size measures that quantify misfit (i.e., non-invariance) in multi-group confirmatory factor analysis for items and item sets. The third manuscript (Groskurth, Bluemke, & Lechner, 2022b) suggested so-called Measurement Invariance Violation Indices (MIVIs) as effect size measures for non-invariant parameter differences in items and item sets. MIVIs quantify the loading, intercept, or uniqueness differences of non-invariant items in units of the pooled standard deviation of the latent variable (either per item or as an average for item sets). They help decide on keeping non-invariant items or dropping them from the item set. Further, they assist in evaluating the quality of a questionnaire in a new research context. MIVIs are also helpful in assessing the

[1] Fit indices cannot be used as effect size measures here. They vary depending on the characteristics of the empirical setting (as outlined throughout this thesis). Further, they were already used in hypothesis testing when evaluating the fit of the invariance model with binary cutoffs. They should not be used as effect size measures simultaneously (Gomer et al., 2019).

amount of bias due to non-invariance in simple proxies such as observed (co)variances or mean scores.

By providing valuable tools to quantify misfit and its impact on simple statistics continuously, I wanted to provide a different, practical view on misfit and enrich binary decisions on model fit from a quantitative perspective. Using effect size measures to quantify misfit (here: non-invariance) connects to the replication crisis, where abandoning binary fit-misfit decisions is hotly debated (e.g., Cumming, 2014; Flora, 2020; Savalei & Dunn, 2015). In other words, using effect size measures to quantify misfit also exhibits a change of perspectives: It implies moving away from (solely) binary decisions to (additional) continuous quantifications.

1.4.3 Goals of the Present Research

This thesis aimed to provide insights and tools that hopefully aid in changing the status quo of model evaluation in structural equation modeling, particularly confirmatory factor analysis. I approached this overarching goal by pursuing two strategies: (1) As a first strategy, I investigated which model, data, and estimation characteristics influence fit indices and how they influence them. I mainly pursued this strategy in the first manuscript (Groskurth, Bluemke, & Lechner, 2022a) by conducting a large-scale simulation study. (2) As a second strategy, I developed novel approaches for model evaluation that do not rely on fixed cutoffs. I mainly pursued this strategy in the second manuscript (Groskurth, Bhaktha, & Lechner, 2022) by developing the simulation-cum-ROC approach that generates cutoffs tailored to the setting of interest. Additionally, the table- and regression-based approaches to tailored cutoffs in the first manuscript belonged to that strategy (Groskurth, Bluemke, & Lechner, 2022a). I also primarily followed the second strategy in the third manuscript (Groskurth, Bluemke, & Lechner, 2022b), developing effect size measures (so-called Measurement Invariance Violations Indices) to quantify misfit (i.e., non-invariance) in addition to binary fit-misfit decisions. Ultimately, I hope the provided insights and tools help to move forward in abandoning fixed cutoffs for more valid model evaluation. I discuss the main gist of all three manuscripts below. Figure 5.1 in Chapter 5 visualizes the integrative concept of this thesis. Further, Chapter 5 includes an empirical example illustrating all approaches developed in this thesis.

2 The Present Research

2.1 Manuscript I: Problems of Fixed Cutoffs for Fit Indices

Groskurth, K., Bluemke, M., & Lechner, C. M. (2022a). *Why we need to abandon fixed cutoffs for goodness-of-fit indices: A thorough simulation and possible solutions.* Invited revision at *Behavior Research Methods*.

2.1.1 Motivation: Why Are Fixed Cutoffs Invalid, and What Are Open Questions?

Methodologists have long criticized the use of fixed cutoffs for fit indices. Fixed cutoffs are invalid when applied in empirical settings which were not part of the simulation scenarios from which the cutoffs originated. This conclusion stems from the finding that fit indices are susceptible to several model, data, and estimation characteristics. These are the sample size (e.g., DiStefano et al., 2019), type of estimator (e.g., Xia & Yang, 2019), number of items (e.g., Kenny & McCoach, 2003), number and distribution of response options (e.g., Xia & Yang, 2018), magnitude of factor loadings (e.g., Heene et al., 2011), and factor correlation (e.g., Beauducel & Wittmann, 2005). Prior simulation studies mainly focused on influences of specific characteristics on fit indices, a large-scale simulation study investigating all influences in tandem is lacking. It remains unclear how susceptible fit indices are to the joint influences of these characteristics and interactions thereof (e.g., sample size × number of response options). Such a large-scale simulation study allows for investigating which influences replicate while controlling the impact of other characteristics. Investigating joint influences might unravel new patterns (i.e., substantial interactions) that remain hidden when focusing on single characteristics. Such a study allows for comparing the relative strength of influences on fit indices. It can identify the most important influences (among many).

Two questions guided the present simulation study: Which model, data, and estimation characteristics influence values of fit indices the most? How do those characteristics influence the sensitivity of fit indices to detect misspecification? The first question concerns the *susceptibility* of fit indices to model, data, and estimation characteristics. The second question concerns the *sensitivity* of fit indices to detect misspecification.

2.1.2 Method: Simulation Study

The present Monte Carlo simulation study contained several scenarios. Each scenario comprised a population model with various characteristics, from which I sampled data in various sizes, and a correctly specified or misspecified analysis model fit to the data with different estimators. I only included confirmatory factor analysis models. I conducted two separate simulations, investigating two combinations of correctly specified and misspecified models. Either the factor dimensionality or cross-loadings were correctly specified or misspecified. I analyzed these two simulations separately.

In the first simulation, the population model was either a one-factor or correlated two-factor model. Factor correlations were r = .70, .50, or .30 to include different magnitudes of misspecification (i.e., a difference of .30, .50, or .70, when viewed from two perfectly correlating factors r = 1, which essentially subsume to a single factor). I fit a one-factor analysis model to data from both population models (i.e., one-factor and correlated two-factor). Thus, each analysis model was either correctly specified or misspecified regarding factor dimensionality.

In the second simulation, the population model was a two-factor model, either without or with cross-loadings. To include different magnitudes and proportions of misspecification, 17% or 33% of all items had cross-loadings with a standardized size of .20 or .30. Cross-loadings were only present on one of the two factors. I fit a two-factor analysis model without any cross-loadings to data generated from both population models (i.e., without and with cross-loadings). Thus, each analysis model was either correctly specified or misspecified regarding cross-loadings.

In both simulations, I varied six different model, data, and estimation characteristics: number of items (6, 12), number of response options (3, 5, 7), distribution of responses (symmetric, asymmetric), loading magnitude (.40, .60, .80), sample size (200, 500, 2,000), and type of estimator. I included the following types of estimators: maximum likelihood, ML, its robust variant, MLR, (constructed for continuously measured items), diagonally weighted least squares, DWLS, and its means-and-variances adjusted variant, WLSMV (constructed for categorically measured items; Li, 2016). In the simulation for models without and with cross-loadings, I also varied the factor correlation (.00, .30). Depending on the factor correlation, factors were forced to be uncorrelated (factor correlation = .00) or allowed to correlate (factor correlation = .30) in both population and analysis models. In the simulation regarding factor

dimensionality, the factor correlation confounded with the misspecification and was, thus, not considered an independent characteristic. I sampled data 1,000 times from each scenario, which resulted in n = 1,728,000 for the first simulation regarding factor dimensionality and n = 4,320,000 for the second simulation regarding cross-loadings. The final analysis contained fit indices for N = 5,956,844 converged models (about 2% of the models did not converge for various reasons).

I considered CFI (Bentler, 1990), RMSEA (Steiger, 1990; see also Chen, 2007), and SRMR (Bentler, 1995; Hu & Bentler, 1999) as fit indices. I also investigated influences on the χ^2 test statistic (Bollen, 1989) and its prominent variant, which is the χ^2 test statistic divided by the model's degrees of freedom (χ^2/df). Essentially, the χ^2 test statistic is a formal test statistic rather than a fit index, although often used as the latter (Jöreskog & Sörbom, 1993).

2.1.3 Results: Summary of Key Findings

I only discuss the main insights from the descriptive, bivariate, and multivariate analysis in the following. First, fit indices detected misspecification of the factor dimensionality and unmodeled cross-loadings. As expected, all fit indices showed worse fit as the two factors' correlation in the population model decreased, but a one-factor analysis model was fit. With more and higher cross-loadings in the population model, unmodeled in the analysis model, fit indices showed worse fit. However, the expected pattern for unmodeled cross-loadings only occurred if the two factors did not correlate in the population model and could not correlate in the analysis model.

Second, as the last finding suggested, fit indices' sensitivity to misspecification depended on several characteristics. The most interesting finding was the following: When the factors of the two-factor analysis model were allowed to correlate and the proportion of unmodeled cross-loadings increased, fit indices showed better (and not worse) fit. Fit indices signaled better fit as the degree of misspecification increased. Although it seems counterintuitive, this finding is plausible. A two-factor analysis model with correlated factors (flexibly) accounts for substantial cross-loadings in the population model (unmodeled in the analysis model) through other model parameters. Most prominently, it accounts for substantial unmodeled cross-loadings through the factor correlation. The factor correlation increases as the factor with unmodeled cross-loadings becomes a blend of both factors. The higher the proportion of unmodeled cross-loadings, the stronger the correlation between both factors (i.e., the more one factor becomes a blend of both), and, ultimately, the better the fit. I found a strong

correlation (tau-b = .54) between the estimated factor correlation and the proportion of unmodeled cross-loadings, which corroborated the interpretation. Likewise, factor loadings became higher for items with unmodeled cross-loadings. Through the simulation design, an item's residual variance decreased if its factor loading increased or when a cross-loading was added to the population model.

Third, model, data, and estimation characteristics did not only interact with model misspecification. Fit indices were susceptible to characteristics even when controlling for model misspecification or looking at correctly specified models. Especially the type of estimator, loading magnitude, and factor correlation influenced fit indices. Fit indices were differently susceptible for correctly specified and misspecified models. They were even differently susceptible to various kinds of those models (i.e., those of the factor dimensionality and cross-loading simulation). My results indicated that established patterns of fit index susceptibility were more complex than previously assumed, as those patterns changed with different model, data, and estimation characteristics. Most interestingly, studies (Xia & Yang, 2019) have shown that fit indices (i.e., χ^2, CFI, and RMSEA) indicated better fit for misspecified models based on estimators for categorical data (i.e., DWLS) than for continuous data (i.e., ML). My simulation study revealed the same pattern as former studies (Xia & Yang, 2019). It even found the pattern for fit indices (χ^2, χ^2/df, CFI, and RMSEA) of models with different types of misspecification (i.e., factor dimensionality and cross-loadings). There was just one exception, which revealed the complexity of influences on fit indices. The pattern reversed with uncorrelated factors and unmodeled cross-loadings: Fit indices (i.e., χ^2, χ^2/df, CFI, and RMSEA) indicated worse fit for misspecified models based on estimators for categorical data (i.e., DWLS) than for continuous data (i.e., ML).

Fourth, investigating several known influences in tandem also revealed that some characteristics did not impact fit indices as substantially as previous research suggested. For instance, Xia and Yang (2018) showed that fit indices (i.e., χ^2, χ^2/df, CFI, and RMSEA) from misspecified models indicated better fit with asymmetric response distributions than with symmetric ones (estimated with DWLS and WLSMV that are estimators for categorical items). I replicated this finding in a multivariate regression analysis, though the interaction effect (DWLS/WLSMV × asymmetric response distribution) was relatively small compared to others. To illustrate, the CFI of a two-factor analysis model with unmodeled cross-loadings changed only by 0.001 points with the interaction term DWLS × asymmetric (reference: ML,

symmetric), whereas it changed by 0.032 points with the interaction term DWLS × correlated factors (reference: ML, uncorrelated factors).

2.1.4 Discussion: We Need Alternatives to Fixed Cutoffs!

These findings showed that fit indices depended strongly on several model, data, and estimation characteristics other than misspecification. There cannot be a single cutoff judging model fit across diverse scenarios. Applied researchers must switch to cutoffs tailored to the specific setting of interest. Based on the simulation study, I provided two tools to generate cutoffs tailored to diverse empirical settings.

First, I generated scenario-specific cutoff tables. The tables contain cutoffs from simulated, scenario-specific fit index distributions at a 5% Type I error rate (i.e., 5% of the time, cutoffs wrongly rejected correctly specified models). Similar to looking up critical values of *z*-scores or *t*-statistics, researchers select the cutoff that best matches their empirical setting from the scenario-specific cutoff tables. The first manuscript (Groskurth, Bluemke, & Lechner, 2022a) includes scenario-specific cutoff tables for fit indices of this simulation study.

Second, I used those scenario-specific cutoff tables and regressed the cutoffs, as the dependent variable, on all model, data, and estimation characteristics (as well as their interactions), as independent variables, separately for each fit index. Thus, I obtained a regression formula that contains weights (i.e., regression coefficients) for each characteristic considered in the simulation study. This regression formula can be used to derive a tailored cutoff. Researchers need to plug the characteristics of the empirical setting (e.g., ML estimator, six items, seven response options, symmetric response distribution, average standardized loading magnitude of .80, sample size of 500, one factor) into the regression formula. The weighted sum of the characteristics prescribed through the regression formula predicts the cutoff. The first manuscript (Groskurth, Bluemke, & Lechner, 2022a) includes regression formulae for fit indices of this simulation study. Chapter 5 contains an empirical example of generating tailored cutoffs from the table- and regression-based approaches.

The table- and regression-based approaches allow for scenario-specific cutoffs; however, those cutoffs are valid only if the setting of interest does not deviate strongly from the scenarios in the initial simulation. Further, it remains unknown which fit index (out of various ones commonly considered) is most helpful for fit-misfit decisions in the setting of interest. Thus, in the second manuscript, I explored a more general approach that generates tailored cutoffs for well-performing fit indices within the setting of interest.

2.2 Manuscript II: Tailored Cutoffs as Alternatives to Fixed Cutoffs

Groskurth, K., Bhaktha, N., & Lechner, C. M. (2022). *Making model judgments ROC(K)-solid: Tailored cutoffs for fit indices through simulation and ROC analysis in structural equation modeling.* Invited revision at *Psychological Methods*.

2.2.1 Motivation: Why a New Tailored Cutoff Approach?

Having already outlined some approaches to tailored cutoffs in the first manuscript (Groskurth, Bluemke, & Lechner, 2022a), I took an even closer look at different—more flexible and widely applicable—approaches to tailored cutoffs in the second manuscript (Groskurth, Bhaktha, & Lechner, 2022). Besides the table- and regression-based approaches outlined in the first manuscript (Groskurth, Bluemke, & Lechner, 2022a), other approaches to tailored cutoffs have been developed in recent years. The χ^2 distribution-based approach uses the χ^2-distributional features without and with misspecification at different sample sizes, degrees of freedom, and the number of items to generate cutoffs for fit indices including the χ^2 test statistic (Moshagen & Erdfelder, 2016; see also Jak et al., 2021; Jobst et al., 2021). Further, simulations can generate tailored cutoffs. Researchers obtain fit index distributions by fitting the model of interest to simulated data for which the population model is known (McNeish & Wolf, 2021, 2022; Millsap, 2007, 2013; Niemand & Mai, 2018; Mai et al., 2021; Pornprasertmanit, 2014; see also Schmalbach et al., 2019; Pornprasertmanit et al., 2021). Similarly, tailored cutoffs can result from bootstrapping—the non-parametric variant of the simulation-based approach outlined before. The non-parametric bootstrap approach obtains fit index distributions from a model of interest fit to transformed data. The data originates from the empirical data transformed as if the model has generated it (Bollen & Stine, 1992; Kim & Millsap, 2014; see also Yuan & Hayashi, 2003; Yuan et al., 2004, 2007).

All approaches are superior to fixed cutoffs. Some of them allow a quick derivation of tailored cutoffs (i.e., table-, regression-, and χ^2 distribution-based approaches). Others enable the generation of cutoffs for several fit indices across various settings (i.e., bootstrap and simulation-based approaches). However, none of the existing approaches allows for evaluating the performance of several fit indices (i.e., their ability to discriminate between correctly specified and misspecified models) within the setting of interest. Assessing the performance of fit indices, in general, helps to answer the question on which fit index one can (primarily) rely

on for fit-misfit decisions—an important question with which many researchers grapple (e.g., Jackson et al., 2019; Mai et al., 2021).

2.2.2 Approach: The Simulation-cum-ROC Approach

Thus, I developed the so-called simulation-cum-ROC approach. It builds on the simulation-based approach (e.g., McNeish & Wolf, 2021; Millsap, 2013; Pornprasertmanit, 2014) rather than the bootstrap approach as the simulation-based approach has a long tradition for generating cutoffs (e.g., Hu & Bentler, 1999). Thus, the simulation-based approach is presumably better known and easier to understand than the bootstrap approach. Features of the receiver operating characteristic (ROC) analysis (for a detailed description of ROC analysis, see Flach, 2016) herein advance the simulation-based approach. The simulation-cum-ROC approach allows to generate tailored cutoffs at balanced Type I and Type II error rates (i.e., false rejection and false acceptance probabilities) for all fit indices of interest. Importantly, it allows to evaluate the performance of fit indices and helps to pick the fit index that performs best in the setting of interest.

The simulation-cum-ROC approach works as follows (see also Figure 2.1). As an input step, researchers need to fit their model of interest to empirical data. This is important to obtain the model, data, and estimation characteristics of interest (e.g., magnitude of factor loadings, multivariate response distributions, type of estimator). Those characteristics are needed in the next step to tailor the simulation to the setting of interest.

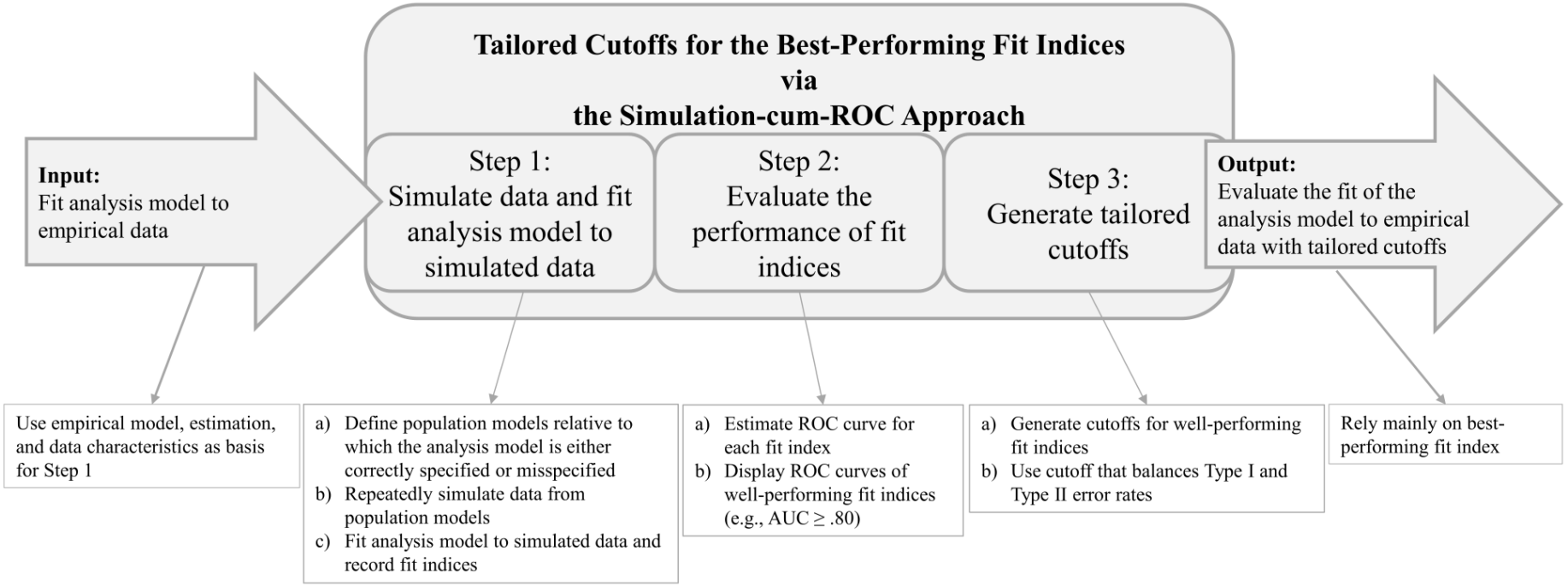
As a first step of the simulation-cum-ROC approach, researchers must repeatedly draw data from two different population models through a Monte Carlo simulation. Those models should align with the hypotheses of the specific research question. The null hypothesis, H_0, states that a population model structurally identical to the analysis model has generated the underlying data. The alternative hypothesis, H_1, states that a population model structurally different from the analysis model to a specific degree of intolerable misspecification has generated the underlying data (Neyman & Pearson, 1928, 1933; see also Biau et al., 2010; Moshagen & Erdfelder, 2016; Perezgonzalez, 2015). Thus, the population model of H_0 is structurally identical to the analysis model. Here, the analysis model is correctly specified. The population model of H_1 is structurally different from the analysis model (e.g., population model parameters remain unmodeled in the analysis model; Hu & Bentler, 1998). Here, the analysis model is misspecified (e.g., under-parameterized to the population model; Hu & Bentler, 1998). What constitutes a relevant H_1 population model and, thus, misspecification is open to the

researcher and their research question. Then, researchers fit the analysis model to data simulated from the H_0 and H_1 population models and record the fit index values. Importantly, they orientate the simulation upon the model, data, and estimation characteristics of interest (e.g., magnitude of factor loadings, multivariate response distributions, type of estimator) identified in the input step. For instance, the analysis model is the one of interest. The population models have, for instance, identical parameter values (e.g., factor loadings) to those of the empirical model (i.e., the model of interest fit to empirical data) identified in the input step. Likewise, the multivariate response distribution, sample size, and estimator used are identical to those in the input step.

In the second step of the simulation-cum-ROC approach, researchers must thoroughly analyze the resulting fit index distributions from correctly specified and misspecified models. ROC analysis visualizes the fit indices' ability to discriminate between those distributions at different cutoff points via the ROC curve. The area under the curve, AUC, quantifies what the ROC curve visualizes—the performance of each fit index. Generally, fit indices with AUC values closer to 1 can better discriminate between correctly specified and misspecified models. Thus, researchers should rely only on fit indices with high AUC values (e.g., AUC $\geq$ .80; Padgett & Morgan, 2021) when evaluating model fit. However, it can be informative to investigate distributions of low-performing fit indices as well. Fit indices quantify different aspects of the model–data relation (for an overview, see Schermelleh-Engel et al., 2003). For instance, the χ^2 test statistic (e.g., Bollen, 1989) quantifies the discrepancy between the model-implied and sample-based variance-covariance matrix. SRMR quantifies the average residuals between model-implied and sample-based covariance matrices (Bentler, 1995). As fit indices characterize the model–data relation differently, the shape and overlap of their distributions help to diagnose the model further (see Browne et al., 2002; Lai & Green, 2016; Moshagen & Auerswald, 2018). Fit index distributions are, thus, also printed in the third step.

In the third step of the simulation-cum-ROC approach, researchers choose optimal cutoffs. The optimal cutoff is the fit index value that best balances Type I error rates (i.e., incorrectly rejecting a correctly specified model) and Type II error rates (i.e., incorrectly accepting a misspecified model) for fit indices with high AUC values (e.g., AUC $\geq$ .80; Padgett & Morgan, 2021). Fit index distributions, the accuracy, Type I error rates, and Type II error rates are printed together with the optimal cutoffs. They reveal which error rates are to be expected when applying the cutoffs in the specific setting.

Figure 2.1: *The Simulation-cum-ROC Approach*



Note. Figure copied from Groskurth, Bhaktha, and Lechner (2022).

Finally, as an output step, researchers compare empirical fit index values to the tailored cutoffs generated via the simulation-cum-ROC approach to decide whether the model of interest fits the empirical data. Researchers should rely primarily on fit-or-misfit indications from the fit index with the highest AUC value (identified in the second step of the simulation-cum-ROC approach). The second manuscript (Groskurth, Bhaktha, & Lechner, 2022) includes empirical examples illustrating how tailored cutoffs can be generated via the simulation-cum-ROC approach. Chapter 5 compares tailored cutoffs generated via the simulation-cum-ROC approach to table- and regression-based cutoffs within an empirical example.

2.2.3 Discussion: Usability of the Simulation-cum-ROC Approach

Overall, the simulation-cum-ROC approach generates cutoffs at balanced Type I and Type II error rates tailored to the setting of interest. Thus, it constitutes a clear advancement over generic fixed cutoffs (e.g., Hu & Bentler, 1999). Enriching Monte Carlo simulations with ROC analysis allows for evaluating the performance of fit indices in discriminating between correctly specified and misspecified models. Thus, the simulation-cum-ROC approach provides guidance on which fit index to rely on in the setting of interest, which uniquely distinguishes it from other approaches to tailored cutoffs.

The simulation-cum-ROC approach exhibits great flexibility: Several approaches to tailored cutoffs predefine what constitutes correctly specified and misspecified models (e.g., McNeish & Wolf, 2021; Niemand & Mai, 2018). The simulation-cum-ROC approach requires applied researchers to define population models relative to which the analysis model is either correctly specified or misspecified. Thus, researchers must decide which model–data relation they deem acceptable and which they deem unacceptable. Many considerations might influence these decisions (Cudeck & Henly, 1991; MacCallum, 2003; MacCallum & Tucker, 1991): Should the model of interest exactly describe the underlying population? If so, the analysis model should be structurally identical to the H_0 population model. Cutoffs point to good fit if the analysis model is identical to the population model. Should the model of interest be generalized across different populations, ignoring minor factors that might be present only in the specific population? If so, the analysis model should be approximately like the H_0 population model (e.g., minor factors are apparent in the H_0 population model but not included in the analysis model). Cutoffs point to good fit if the analysis model captures the relevant complexities of the population model (e.g., major/general factors). At the same time, it stays plausible across different population models (e.g., ignoring minor factors that might be specific

to a certain population). Is there a strong theory for an alternative model based on which I want to reject the model of interest if that alternative model would have generated the data? If there is a strong theory for an alternative model, this is a relevant H_1 population model. Otherwise, researchers must find another plausible H_1 population model. Additionally, researchers can define not only single but several H_1 population models (but also H_0 population models when using approximately fitting analysis models) to include a wide variety of forms and sizes of misspecification in the simulation-cum-ROC approach (Pornprasertmanit et al., 2013; Pornprasertmanit, 2014).

Further, the simulation-cum-ROC approach allows researchers to evaluate which Type I and Type II error rates they are willing to accept. Several approaches obtain tailored cutoffs at predefined error rates—and provide no cutoff if it exceeds certain error rates (10%; McNeish and Wolf, 2021). The simulation-cum-ROC approach provides cutoffs at balanced Type I and Type II error rates. Although the AUC screens out low-performing fit indices (Padgett & Morgan, 2021, recommended AUC $\geq$ .80), error rates of cutoffs might exceed conventional levels of 10%. It is left to the researcher whether they decide to use such cutoffs. Researchers might be willing to use cutoffs exceeding conventional error probabilities when they are highly interested in the very concrete operationalization of their hypotheses (e.g., the alternative population model has a strong theoretical foundation).

The flexibility of the simulation-cum-ROC approach requires researchers to thoroughly think through their research question, hypotheses, and error probabilities based on their model evaluation of interest. To avoid subjectivity (or arbitrary decisions) influencing the results, researchers must justify their choices (e.g., population models, accepted error probabilities) when applying the simulation-cum-ROC approach. The simulation-cum-ROC approach requires researchers to make choices explicitly (that are otherwise implicitly made, e.g., McNeish & Wolf, 2021). As such, others can evaluate the appropriateness of this argumentation. To ease the computational part and, in turn, the application of the simulation-cum-ROC approach to tailored cutoffs, I appended R code to the manuscript and developed a shiny app (available under https://kg11.shinyapps.io/tailoredcutoffs/).

2.3 Manuscript III: Effect Size Measures to Quantify Measurement Non-Invariance

Groskurth, K., Bluemke, M., & Lechner, C. M. (2022b). *Measurement Invariance Violation Indices (MIVIs): Effect sizes for non-invariance of items and item sets.* Manuscript in preparation.

2.3.1 Motivation: How Detrimental Is Non-Invariance?

The third manuscript went beyond the binary fit-misfit decision of model evaluation and evaluated the size and impact of misfit itself. It focused on the unique context of measurement non-invariance. Here, the goal is to assess whether group-level statistics (such as average self-esteem) are comparable across groups (such as countries) without bias (i.e., non-invariance). Measurement invariance testing (e.g., Davidov et al., 2014; Millsap, 2011; Steenkamp & Baumgartner, 1998) commonly operates in a sequential hierarchy. Configural invariance tests investigate whether the item-to-factor structure is the same across groups. Metric invariance tests examine whether factor loadings are the same across groups. Scalar invariance tests additionally constrain the intercepts to be the same across groups. Finally, uniqueness invariance tests additionally constrain residual variances to be the same across groups.

Researchers should refrain from comparing observed statistics (such as variances or mean scores) across groups when non-invariance bias is large (i.e., strong or many parameter differences across groups). Small non-invariance bias (i.e., tiny or few parameter differences across groups) need not be detrimental to differences in cross-group statistics (Chen, 2008). Further, uniform non-invariance (i.e., all non-invariant parameters are higher in one group) biases differences in cross-group statistics stronger than mixed non-invariance (i.e., some non-invariant parameters are higher in one group, others are higher in another group; Chen, 2008)

Effect size measures are useful for quantifying non-invariance bias (e.g., Millsap & Olivera-Aguilar, 2012; Nye & Drasgow, 2011b; Oberski, 2014; Pornprasertmanit, 2022). However, existing effect size measures are relatively complicated to apply (as they require extra statistical packages; Dueber, 2019; Nye & Drasgow, 2011b; Oberski, 2014). Alternatively, they focus only on single items instead of the complete item set (Millsap & Olivera-Aguilar, 2012; Pornprasertmanit, 2022), making it impossible to investigate aggregation or compensation effects (i.e., due to uniform or mixed non-invariance).

2.3.2 Approach: The Measurement Invariance Violation Indices (MIVIs)

To address these shortcomings, I proposed the easy-to-apply Measurement Invariance Violation Indices (MIVIs) that are effect size measures for items and item sets. The core idea of MIVIs is to quantify the non-invariant differences in parameters (i.e., factor loadings, intercepts, or residual standard deviations) in units of the standard deviation of the latent variable pooled across groups. The parameter differences constitute the numerator. The pooled latent standard deviation forms the denominator. Pooled latent standard deviations are favorable compared to pooled observed standard deviations of items and item sets. Unlike observed standard deviations (of either items or item sets), latent standard deviations are the same for all items in an item set, independent of the number of items in an item set, and consist of true score variance only (i.e., the variance of the latent variable).

MIVIs rest on partially invariant multi-group confirmatory factor analysis models (Byrne et al., 1989). Full invariance requires that cross-group parameter constraints must hold for all items. Partial invariance implies that cross-group parameter constraints must only hold for some (at least two) items (Byrne et al., 1989; Steenkamp & Baumgartner, 1998; Steinmetz, 2013). The model has group-equivalent and group-specific factor loadings, intercepts, or unique variances.

I derived absolute and signed (i.e., directional) versions of MIVIs for items and item sets (summarized in Table 2.1). Absolute MIVIs contain the absolute value of parameter differences in the numerator. They are bounded at zero and can only have positive values. Signed MIVIs trace the mathematical signs of the parameter differences. Thus, signed MIVIs allow for the investigation of aggregation or compensation effects when either uniform or mixed non-invariance is present. Signed MIVIs are unbounded.

MIVIs at the item level only include parameter differences of single items. They are specifically helpful in scale development processes with no fixed item set. MIVIs at the level of item sets summarize across all parameter differences. Dividing those MIVIs by the number of items in a set gives an overall impression of how biased the item set is on average. Those MIVIs are especially relevant if non-invariance occurs in a fixed item set needed for further analysis (such as comparing observed variances or mean scores across groups).

Table 2.1: *Measurement Invariance Violation Indices (MIVIs)*

		Measurement Invariance Violation Indices (MIVIs)	
		Absolute	Signed
Item level	Loading	$MIVI\text{–}L_{Item\ j\|absolute} = \frac{\lvert\lambda_{2j} - \lambda_{1j}\rvert}{SD_{LV,\ pooled}}$	$MIVI\text{–}L_{Item\ j\|signed} = \frac{\lambda_{2j} - \lambda_{1j}}{SD_{LV,\ pooled}}$
	Intercept	$MIVI\text{–}I_{Item\ j\|absolute} = \frac{\lvert\tau_{2j} - \tau_{1j}\rvert}{SD_{LV,\ pooled}}$	$MIVI\text{–}I_{Item\ j\|signed} = \frac{\tau_{2j} - \tau_{1j}}{SD_{LV,\ pooled}}$
	Uniqueness	$MIVI\text{–}U_{Item\ j\|absolute} = \frac{\lvert\sqrt{\theta_{2j}} - \sqrt{\theta_{1j}}\rvert}{SD_{LV,\ pooled}}$	$MIVI\text{–}U_{Item\ j\|signed} = \frac{\sqrt{\theta_{2j}} - \sqrt{\theta_{1j}}}{SD_{LV,\ pooled}}$
Item set level	Loading	$MIVI\text{–}L_{Item\ set\|absolute} = \frac{\sum_{j=1}^{p} \lvert\lambda_{2j} - \lambda_{1j}\rvert}{SD_{LV,\ pooled}}/p$	$MIVI\text{–}L_{Item\ set\|signed}$ = *not applicable*
	Intercept	$MIVI\text{–}I_{Item\ set\|absolute} = \frac{\sum_{j=1}^{p} \lvert\tau_{2j} - \tau_{1j}\rvert}{SD_{LV,\ pooled}}/p$	$MIVI\text{–}I_{Item\ set\|signed} = \frac{\sum_{j=1}^{p}(\tau_{2j} - \tau_{1j})}{SD_{LV,\ pooled}}/p$
	Uniqueness	$MIVI\text{–}U_{Item\ set\|signed} = \frac{\sum_{j=1}^{p}\lvert\sqrt{\theta_{2j}} - \sqrt{\theta_{1j}}\rvert}{SD_{LV,\ pooled}}/p$	$MIVI\text{–}U_{Item\ set\|signed} = \frac{\sum_{j=1}^{p}(\sqrt{\theta_{2j}} - \sqrt{\theta_{1j}})}{SD_{LV,\ pooled}}/p$
		$SD_{LV,\ pooled} = \sqrt{\frac{(n_2-1)\Phi_2 + (n_1-1)\Phi_1}{n_1+n_2-2}}$	

Assumptions / preconditions

1. Multi-factor models: separate MIVIs must be estimated per factor
2. Multi-group measurement model must be correctly identified
3. Common factor model must hold
4. When estimating MIVIs, partial invariance model for the invariance level of interest must hold
5. Final partial invariance model should be identified by fixing the latent variable at 1 / latent mean at 0
6. Three types of MIVIs must not be compared (i.e., MIVI–L, MIVI–I, MIVI–U)

Note. I defined MIVIs for a non-invariant (unstandardized) loading λ_{kj}, intercept τ_{kj}, or unique standard deviation $\sqrt{\theta_{kj}}$ of the non-invariant item *j* (*j* = 1, 2, …, *p*) in group *k* (*k* = 1, 2). The equation for the standard deviation of the latent variable pooled across groups, $SD_{LV,\ pooled}$, was obtained from Cohen (1988). Here, Φ_k is the variance of the scores on the latent variable, and n_k is the sample size in group *k*. $MIVI_{Item\ set|signed}$ is only given for intercepts or unique standard deviations but not for factor loadings. Loadings multiply with the latent means and variances, whereas intercepts or unique variances are additive components when producing observed means and variances of the items (or item sets; for a formula-based definition, see Nye & Drasgow, 2011b). Thus, loadings ultimately connect to the means and variances of the latent variable; compensation (or aggregation) across groups is, thus, not immediately conceivable.

Although effect sizes, such as MIVI values, can be used continuously, some guidance on what constitutes small, medium, or large non-invariance can be helpful. Small non-invariance bias may allow researchers to retain items in an item set or to compare group-level statistics, despite non-invariance being present.

Cohen's (1988, 1992) effect size guidelines for his *d*, quantifying the independent mean difference relative to the pooled standard deviation, could serve as initial guidance, especially for $MIVI - I$ quantifying intercept differences. He suggested that $|d| = 0.2$ may indicate a substantial but small, $|d| = 0.5$ a medium, and $|d| = 0.8$ a large effect. However, these guidelines are extremely rough. Appropriate guidelines are subject to many aspects like the substantive research question, type of analysis, empirical setting, and parameter of interest (i.e., loading, intercept, or uniqueness) that may influence what is considered a critical value (e.g., Steinmetz, 2013).

2.3.3 Discussion: Usability of MIVIs

By quantifying non-invariance bias in a continuous and easy-to-apply manner, MIVIs overcome the binary fit-misfit logic in model testing, precisely in measurement invariance testing. With not yet fixed item sets (such as in the scale development process), MIVIs help to select group-fair items. With fixed item sets, MIVIs can evaluate the questionnaire quality in a new context. Further, MIVIs can quantify the bias due to non-invariance in the item set that resonates in observed statistics compared across groups (e.g., variances or mean scores). To ease the application of MIVIs, I provided Mplus and R codes in the manuscript (Groskurth, Bluemke, & Lechner, 2022b).

Even when MIVIs identify non-substantial or mixed non-invariance (i.e., non-invariance that compensates across groups), researchers must be aware that non-invariance is present in the item set. MIVIs are no legitimation to completely ignore non-invariance; instead, MIVIs quantify its size and, ultimately, its potential impact on observed statistics compared across groups (e.g., variances or mean scores). If non-invariance is present in either form, researchers should always reflect on the causes of the specific non-invariance. Researchers can investigate causes of non-invariance theoretically (e.g., Chen, 2008) but also empirically, such as via a multiple-indicators multiple-causes model (or more flexible variations of it, e.g., Bauer, 2017; Kolbe et al., 2022).

3 General Discussion

3.1 Summary

The overarching goal of this thesis was to close the gap between the practice of using fixed cutoffs for model evaluation and the methodological criticism of its one-size-fits-all usage by pursuing two strategies. As a first strategy, I thoroughly investigated how several model, data, and estimation characteristics influence fit indices through a large-scale simulation study in the first manuscript (Groskurth, Bluemke, & Lechner, 2022a). Fit index values differ not only depending on model misspecification (as they should) but also depending on the setting of interest (as they should not). No generic, binary heuristic (i.e., fixed cutoff) will adequately differentiate between fit index values of correctly specified and misspecified models across diverse settings. Thus, cutoffs have no external validity; they are only valid within the setting from which they originate. Cutoffs must be tailored to the setting of interest. As a second strategy, I developed novel approaches for model evaluation that do not rely on fixed cutoffs. Thus, I summarized existing approaches to tailored cutoffs in the second manuscript (Groskurth, Bhaktha, & Lechner, 2022). Then, I developed easy-to-apply approaches to cutoffs tailored to the setting of interest, particularly table- and regression-based approaches in the first manuscript (Groskurth, Bluemke, & Lechner, 2022a) and the simulation-cum-ROC approach in the second manuscript (Groskurth, Bhaktha, & Lechner, 2022). I also developed so-called Measurement Invariance Violation Indices in the third manuscript (Groskurth, Bluemke, & Lechner, 2022b) that allow for continuously quantifying misfit (i.e., non-invariance) in addition to binary fit-misfit decisions (made through cutoffs for fit indices).

After summarizing the main contributions of the three manuscripts, I next discuss what they add to the methodology and general practice of model evaluation. The first section considers the methodological perspective and is concerned with the accumulation of knowledge through the three manuscripts. The second section considers the applied perspective and is concerned with the understandability of explanations and the applicability of approaches to change the practice of model evaluation through the three manuscripts.

3.2 Methodological Perspective: Contribution to the Methodology of Model Evaluation

In this section, I discuss how all three manuscripts contributed to accumulating knowledge on problems of and alternatives to fixed cutoffs for fit indices. Specifically, the first manuscript (Groskurth, Bluemke, & Lechner, 2022a) contributed to the literature on the sensitivity and susceptibility of fit indices to diverse model, data, and estimation characteristics—merely analyzed in isolation before. Evaluating several characteristics in tandem helped to evaluate the relative importance of those influences. As such, some characteristics had a smaller impact on fit indices than previously assumed. For instance, the response distribution had not such a strong influence on DWLS/WLSMV-based fit indices as initially suggested by Xia and Yang (2018). Differently, other more underappreciated characteristics had a stronger impact on fit indices than previously assumed. Especially interesting was that several unmodeled cross-loadings resulted in good fit when factors correlated. The correlated two-factor model (flexibly) accounts for unmodeled cross-loadings through other parameters (i.e., factor correlation and factor loadings). Further, nearly every characteristic impacted fit indices to some extent—making the influences non-traceable across diverse settings.

The large-scale simulation study focused on confirmatory factor analysis (CFA). I know only one similarly large simulation study on another model class, exploratory structural equation modeling (ESEM; Garrido et al., 2016). Garrido et al. (2016) mainly included the characteristics I included (i.e., factor loading, number of items, number of factors, factor correlation, sample size, response categories, distribution, and estimators) but only focused on correctly specified models. Findings somewhat differed across model classes. For instance, Garrido et al. (2016) found a strong influence of the number of items per factor and sample size on all fit indices they investigated (i.e., CFI, TLI, RMSEA, SRMR) for correctly specified ESEM models. In my study (Groskurth, Bluemke, & Lechner, 2022a), the number of items only strongly impacted χ^2, and the sample size only strongly impacted SRMR for correctly specified CFA models. The different findings could hint at model-specific operations. Put differently, values of fit indices and the strength of influences on these values may differ by model type.[1] I do not know about any comparative large-scale simulation study for longitudinal

[1] Alternatively, different operationalizations of these characteristics may produce different results. Garrido et al. (2016), for instance, simulated smaller sample sizes with levels of 100, 300, and 1,000 than I did in Groskurth, Bluemke, and Lechner (2022a) with sample size levels of 200, 500, and 2,000.

or multilevel models. However, such a large simulation study on the sensitivity and susceptibility of fit indices in either longitudinal or multilevel models could help gain further insights into related model classes.

One could argue that additional research on the sensitivity and susceptibility of fit indices to different model classes is cumbersome. Either way, one can interpret values of fit indices only in the specific setting of interest; there is no generalization at all possible. Values of fit indices differ depending on the setting of interest, different cutoffs to classify "good" and "bad" values of fit indices must ultimately apply.

Further, the manuscripts contributed to the literature on so-called tailored cutoffs for fit indices. I summarized and compared the different approaches to tailored cutoffs (Groskurth, Bluemke, & Lechner, 2022a; Groskurth, Bhaktha, & Lechner, 2022). There had been no systematic review of all these approaches before these manuscripts. So far, I have only compared the approaches conceptually. I did not yet conduct a simulation study that systematically compared, for instance, the Type I and Type II error rates of cutoffs generated from the different approaches in different settings. Such a simulation study can further increase the understanding of the different approaches—as a relevant next step for future research building on my first two manuscripts (Groskurth, Bluemke, & Lechner, 2022a; Groskurth, Bhaktha, & Lechner, 2022).

I did not only compare the different approaches to tailored cutoffs but also developed approaches myself. Based on the large-scale simulation study of the first manuscript (Groskurth, Bluemke, & Lechner, 2022a), I built cutoff tables for specific scenarios valid for comparable empirical settings. I borrowed a feature from other statistics, such as *t*-tests, where selecting cutoffs from large-scale tables is common practice. Further, I used the regression-based approach invented by Nye and Drasgow (2011a) and derived regression formulae for generating tailored cutoffs. My simulation included more characteristics than Nye and Drasgows' (e.g., they focused only on the DWLS estimator), so researchers may use my regression formulae for a broader set of empirical settings. Crucially, empirical settings must be like the scenarios in the simulation, either from Groskurth, Bluemke, and Lechner (2022a) or Nye and Drasgow (2011a). Thus, regression formulae or cutoff tables are still limited, so I developed the more flexible simulation-cum-ROC approach in the second manuscript (Groskurth, Bhaktha, & Lechner, 2022) that is applicable to all settings of interest. The simulation-cum-ROC approach follows a simulation-based approach to tailored cutoffs (like many others, such as McNeish & Wolf, 2021, or Millsap, 2013). Unlike others, it borrows ROC

analysis from signal detection theory (Wixted, 2020), which allows not only for generating tailored cutoffs with balanced Type I and Type II error rates but also evaluating the performance of fit indices in the setting of interest.

The second manuscript (Groskurth, Bhaktha, & Lechner, 2022) additionally outlined how researchers can integrate different and more advanced definitions of H_0 and H_1 population models into the simulation-cum-ROC approach (or any other simulation-based approach to tailored cutoffs). For instance, researchers can generate cutoffs based on an analysis model approximately identical to the H_0 population model. Further, they can consider multiple forms and sizes of misspecification by including multiple H_1 population models (and H_0 population models alike when considering approximate fit). Up to now, I have only described the advanced definitions of H_0 and H_1 population models, which originate primarily from Millsap (2013) and Pornprasertmanit (2014). No standalone guideline study comparing, implementing, and sharing computational code exists for the different definitions of the H_0 and H_1 population models.

The third manuscript (Groskurth, Bluemke, & Lechner, 2022b) contributed to the literature on effect sizes for measurement non-invariance. It provides an approach that quickly quantifies measurement non-invariance in loadings, intercepts, and residual variances, the so-called Measurement Invariance Violation Indices (MIVIs). I built on just a handful of effect size measures for measurement non-invariance in the context of the classical test theory (Millsap & Olivera-Aguilar, 2012; Nye & Drasgow, 2011b; Oberski, 2014; Pornprasertmanit, 2022). So far, there has not been any systematic review of those effect size measures, which is an essential next step for the research community.

I have declared MIVIs as continuous effect sizes for non-invariance. However, researchers cannot only use effect sizes continuously. Categorical heuristics help to interpret what makes a small, medium, and large effect. Heuristics can be either derived by simulations (e.g., Nye et al., 2019) or gathered by empirical data (e.g., Gignac & Szodorai, 2016). So far, I have lent myself to Cohen's effect size guidelines (1992) for his *d*, which are certainly too generic. MIVIs need specific guidelines on what constitutes a small, medium, and large effect in its very own effect size logic—researchers must be able to differentiate sample fluctuations from systematic small, medium, or large non-invariance. However, the interpretation of MIVIs depends on the research question, the non-invariant parameter (i.e., factor loadings, intercepts, or residual variances), and the empirical setting. Indeed, effect sizes need a tailored approach to appropriate guidelines.

3.3 Applied Perspective: Contribution to the General Practice of Model Evaluation

So far, I have reviewed how the three manuscripts contributed to the accumulation of knowledge for the methodology of model evaluation. But to make a real impact on the practice of model evaluation, the manuscripts must be understandable and useful for applied researchers. Here, I discuss how the three manuscripts contributed to the gradual improvement of model evaluation practices.

The first manuscript (Groskurth, Bluemke, & Lechner, 2022a) showed that fit indices varied with model, data, and estimation characteristics in strong and unpredictable ways. Thus, fit index values are not easily explainable by specific characteristics, nor are influences of characteristics directly traceable. To make all the influences as easily digestible as possible, I generated an overview table showing all relevant (i.e., strong) effects. The intractability of specific influences on fit indices (especially in empirical settings) should not imply researchers should abandon those fit indices in general. Instead, they should use cutoffs more conscientiously—not in a one-size-fits-all but scenario-specific logic.

Therefore, I developed easy-to-apply alternatives to fixed cutoffs: Cutoff tables, where one can look up different cutoffs for different scenarios, and regression formulae, where researchers can plug in characteristics of interest to calculate tailored cutoffs (see Groskurth, Bluemke, & Lechner, 2022a)—as long as the characteristics of the setting of interest match those of the given, simulated scenarios. I anecdotally demonstrated their usage in the first manuscript (Groskurth, Bluemke, & Lechner, 2022a). In Chapter 5, I demonstrated their usage through an empirical example.

Further, I developed a highly flexible (though more involved) approach to tailored cutoffs: the simulation-cum-ROC approach (Groskurth, Bhaktha, & Lechner, 2022). Although researchers can apply this approach without any knowledge of conducting simulations, it certainly helps to have some pre-knowledge—or even more in-depth knowledge when using the approach's full flexibility.

To make applying the simulation-cum-ROC approach as easy as possible, I shared the R code of my examples (which researchers can adjust according to their needs). To ease the application even more, I developed a shiny app: Researchers need to plug in their model, data, and estimation characteristics of interest. With this information, the app generates cutoffs via the simulation-cum-ROC approach. Researchers do not need to execute R or any statistical

software. An important step for future research is to increase the flexibility of the shiny app even more. It should include not only single but several different definitions of population models (see Pornprasertmanit, 2014). Further, the shiny app should not only include balanced but different weightings of Type I and Type II error rates.

Another important step for future research is to provide a detailed tutorial on deriving cutoffs via the simulation-cum-ROC approach across diverse settings (apart from the examples in the manuscript). The tutorial could also extend the simulation-cum-ROC approach to different model classes (such as multi-group confirmatory factor analysis models). Additionally, its extension to research questions other than global fit evaluation (such as nested fit evaluation within measurement invariance testing, e.g., Pornprasertmanit et al., 2013) is an important future step. Including all flexible options of the simulation-cum-ROC approach into the shiny app and writing a hands-on tutorial on the simulation-cum-ROC approach will undoubtedly foster the applicability of the approach and, in turn, the use of tailored cutoffs.

Further, I have shown that model evaluation does not need to stop when the model of interest fails (i.e., is rejected): In the specific context of (partial) measurement non-invariance, I developed the so-called Measurement Invariance Violation Indices (MIVIs) to quantify the bias of non-invariant loadings, intercepts, or residual variances (Groskurth, Bluemke, & Lechner, 2022b). I have provided simulated examples in the third manuscript (Groskurth, Bluemke, & Lechner, 2022b) and an empirical example in Chapter 5 that illustrates the application of MIVIs for intercept and uniqueness non-invariance, respectively. Further research can be built on this, illustrating the use of MIVIs to quantify non-invariance in loadings.

The actual implementation of MIVIs is easy: Researchers must add a few lines of code in statistical programs such as R or Mplus. Thus, I do not believe it is essential to implement MIVIs in an R package, but this could be a next step in the distant future. More important, as outlined previously, tailored guidelines for interpreting MIVIs as small, medium, or large effects are desperately needed.

3.4 Future Directions

3.4.1 For Methodological Research

The preceding reviews of the manuscripts' contributions to the methodology and general practice of model evaluation have already revealed some directions for future research based on the three manuscripts' findings. Here, I take a broader look at much-needed future directions for research in the general field of model evaluation. These broad directions point to a more in-depth integration of approximate fit and the use of effect size measures for model evaluation.

The famous phrase "All models are wrong, but some are useful," dating back to Box (1978, p. 202), presumably guides many researchers through model evaluation. The phrase essentially taps into the issue of approximate model fit (Cudeck & Henly, 1991; MacCallum, 2003; MacCallum & Tucker, 1991). It means that researchers are satisfied with selecting a good (enough) model approximation to the data. Phrased differently, the analysis model should capture the relevant complexities of the population model (e.g., major/general factors). At the same time, it should generalize across population models (e.g., ignoring minor factors that might be specific to a certain population). It also implies that researchers do not want a model that exactly describes a certain population (including all minor common factors specific to that population). In most cases (MacCallum, 2003), the goal of model evaluation is to find a model that does not fit the data exactly but only approximately (i.e., considering minor model error acceptable such as dismissing minor factors).

Contrary to this goal of model evaluation is a strict reliance on tests of exact fit (i.e., χ^2 test statistic) or the generation of cutoffs for fit indices from exactly fitting models (e.g., Groskurth, Bhaktha, & Lechner, 2022; Hu & Bentler, 1999; Niemand & Mai, 2018). To derive cutoffs for approximately fitting models, researchers must first define approximate fit: How much misfit am I willing to accept? Only a few approaches concretely provide ways to define approximate fit (e.g., Millsap, 2013; Pornprasertmanit, 2014; Yuan et al., 2007). There has not yet been any extensive review (except for a very small section in the second manuscript, Groskurth, Bhaktha, & Lechner, 2022) on definitions of approximate fit, existing approaches to approximate fit, and guidelines on how to operationalize approximate fit in specific settings.

Related to the question of approximate fit is the impact of misfit itself: How strongly does certain misfit impact the estimation of parameters? Put differently: How detrimental is certain misfit? Answering this question helps to understand how much misfit researchers can accept. To answer this question, researchers can investigate, for instance, how strongly misfit

biases parameter estimates and summary statistics (such as variances and means) through simulation studies. This may show that certain extreme positions are explicitly legitimate: Robitzsch and Lüdtke (2022) argued against the local optimization of models, which they call model-based inference, and in favor of design-based inference. Design-based inference means, for example, that researchers assume unit loadings, even if the model fits worse than a model with different loadings—simply because the theory dictates it, and the assumption of unit loadings is relevant for further analysis. This is an appropriate technique if such misfit does not severely bias parameter estimates and summary statistics (such as variances and means).

3.4.2 For Dissemination and Wide-Spread Implementation of Methodological Advancements

Similarly (or even more) important than identifying fields of action for research is to foster the dissemination of methodological advancements, especially through hands-on tutorials, requirements raised by editors and reviewers, and their inclusion in method curricula.

In general, it seems that tutorials, especially in the context of new methods for model evaluation, are missing. It is, for instance, long known that fixed cutoffs for fit indices are invalid in many empirical settings (Marsh et al., 2004). Alternatives have been suggested ten years after that (e.g., Millsap, 2013) but have not been regularly picked up. Illustrative, widely spread tutorials could have helped to disseminate the idea of tailored cutoffs even earlier. Svetina et al. (2020) is a good example of a hands-on tutorial in the context of measurement invariance testing with categorical items.

However, not only the idea of tailored cutoffs (such as any other methodological innovation) must be promoted in a more accessible way to applied researchers, but also the journals' acceptance of fixed cutoffs for model evaluation must undoubtedly change. Applying fixed cutoffs seems legitimate, as journals continue to publish studies evaluating models with fixed cutoffs. Suppose journals would require (or at least recommend) using more valid approaches for model evaluation than fixed cutoffs. In that case, more valid alternatives will undoubtedly replace fixed cutoffs soon.

Additionally, lecturers should teach approaches to tailored cutoffs in structural equation modeling classes. Cursory glances at textbooks and lecture slides suggest that, all too often, students today are taught to apply conventional fixed cutoffs for canonical fit indices rather mechanistically. They seem to never learn about the origins of cutoffs and the potential problems of fit indices. Raising awareness of these issues and highlighting potential

alternatives would be vital to improving model evaluation. The well-known problems of fixed cutoffs must be shared across students so that they "grow up" with more valid alternatives already. Learning and getting to know cutting-edge research from the beginning of an academic career might help disseminate new methods and foster critical examination of methods in general.

3.5 Advice for Applied Researchers

Before concluding, I want to add general advice to applied researchers on choosing appropriate cutoffs for fit indices. Structural equation modeling techniques, particularly confirmatory factor analysis, are primarily tools to investigate substantive cause-effect relationships of variables. Applied research focuses on specific relationships, such as how self-esteem affects life satisfaction (Cao & Liang, 2020) or how religious belief affects death anxiety (Jong & Halberstadt, 2016). Selecting a well-fitting model to represent constructs of interest is not the focus but a necessary pre-step to investigate the research question of interest.

Although this research investigates global (mis)fit through fit indices, researchers should examine a model of interest from different angles. Investigating local (mis)fit is additionally relevant (i.e., how well each part of the model fits the data). Researchers are well acquainted with the modification index and expected parameter change. The modification index shows how strongly χ^2 improves when freely estimating an initially fixed parameter (Satorra, 1989; Sörbom, 1989; see also Whittaker, 2012). The expected parameter change indicates the size of a fixed parameter if it will be freely estimated (Saris et al., 1987; Chou & Bentler, 1993; see also Whittaker, 2012). Both work best in combination and in addition to the power of the modification index (Saris et al. 2009). Investigating standardized residuals between model-implied and sample-based covariances and—even better—the model's plausibility of conditional (in)dependence assumptions between variables help to localize where the model does or does not fit (Maydeu-Olivares & Shi, 2017; Thoemmes et al., 2018). Researchers might also inspect the plausibility of parameters (e.g., their direction and size) and how well the model was replicated in previous investigations (if applicable) to decide whether they retain or disregard the model of interest. Thus, evaluating the global fit via fit indices is necessary but insufficient for the fit-misfit decision of the model of interest.

This thesis showed that researchers must not completely abandon cutoffs for fit indices. Instead, they should use cutoffs for fit indices more conscientiously. When simulation scenarios from which cutoffs originated match empirical settings, researchers can use the

cutoffs suggested by Hu and Bentler (1999) or any other previous simulation study. Likewise, simple approaches to tailored cutoffs, such as table-based (e.g., Groskurth, Bluemke, & Lechner, 2022a) or regression-based ones (e.g., Groskurth, Bluemke, & Lechner, 2022a; Nye & Drasgow, 2011a), allow obtaining valid cutoffs. However, the empirical setting does not always match any already simulated scenario. Only in that case must researchers use highly flexible tools such as the simulation-cum-ROC approach (Groskurth, Bhaktha, & Lechner, 2022) to obtain valid cutoffs. Additionally, the simulation-cum-ROC approach is helpful when researchers need guidance on which fit index to rely on in fit-misfit decisions in a specific setting.

3.6 Conclusion

This thesis outlined problems surrounding the one-size-fits-all usage of fixed cutoffs for evaluating structural equation models, particularly confirmatory factor analysis models. To address these problems, it provided new, more valid perspectives on model evaluation. It would be naïve to assume that this research would magically change the standard practice of model evaluation via fixed cutoffs for fit indices that is so firmly entrenched. There is still a lot to do until the status quo of model evaluation changes, such as the promotion of approximate fit, the evaluation of the bias induced by misfit, and changing method curricula. Nonetheless, I hope this research is at least a step toward more valid model evaluation.

4 References

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

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling, 12*(1), 41–75. https://doi.org/10.1207/s15328007sem1201_3

Beierl, E. T., Bühner, M., & Heene, M. (2018). Is that measure really one-dimensional? *Methodology, 14*(4), 188–196. https://doi.org/10.1027/1614-2241/a000158

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M. (1995). *EQS structural equations program manual.* Multivariate Software.

Berry, J. W., Poortinga, Y. H., Segall, M. H., & Dasen, P. R. (2002). *Cross-cultural psychology: Research and applications* (2nd ed.) Cambridge University Press.

Biau, D. J., Jolles, B. M., & Porcher, R. (2010). P value and the theory of hypothesis testing: an explanation for new researchers. *Clinical Orthopaedics and Related Research, 468*(3), 885–892. https://doi.org/10.1007/s11999-009-1164-4

Bollen, K. A. (1989). *Structural equations with latent variables.* Wiley.

Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research, 21*(2), 205–229. https://doi.org/10.1177%2F0049124192021002004

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Box, G. E. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). Academic Press. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

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. https://doi.org/10.1037/1082-989X.7.4.403

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

Cao, Q., & Liang, Y. (2020). Perceived social support and life satisfaction in drug addicts: Self-esteem and loneliness as mediators. *Journal of Health Psychology, 25*(7), 976–985. https://doi.org/10.1177/13591053177406

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*(5), 1005–1018. https://doi.org/10.1037/a0013193

Chou, C. P., & Bentler, P. M. (1993). Invariant standardized estimated parameter change for model modification in covariance structure analysis. *Multivariate Behavioral Research, 28*(1), 97–110. https://doi.org/10.1207/s15327906mbr2801_6

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

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. https://doi.org/10.1037/0033-2909.109.3.512

Cuijpers, P., Pineda, B. S., Quero, S., Karyotaki, E., Struijs, S. Y., Figueroa, C. A., ... & Muñoz, R. F. (2021). Psychological interventions to prevent the onset of depressive disorders: A meta-analysis of randomized controlled trials. *Clinical Psychology Review, 83*, Article 101955. https://doi.org/10.1016/j.cpr.2020.101955

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*(1), 55–75. https://doi.org/10.1146/annurev-soc-071913-043137

Della Porta, D., & Keating, M. (2008). How many approaches in the social sciences? An epistemological introduction. In M. Keating & D. Della Porta (Eds.), *Approaches and methodologies in the social sciences: A pluralist perspective* (pp. 19–39). Cambridge University Press.

DiStefano, C., McDaniel, H. L., Zhang, L., Shi, D., & Jiang, Z. (2019). Fitting large factor analysis models with ordinal data. *Educational and Psychological Measurement, 79*(3), 417–436. https://doi.org/10.1177/0013164418818242

Dueber, D. (2019). *Package 'dmacs'*. https://cran.r-project.org/web/packages/dmacs/dmacs.pdf

Durkheim, E. (1951). *Suicide, a study in sociology* (J. A. Spaulding & G. Simpson, Trans.). The Free Press. (Original work published in 1897)

Flach, P. A. (2016). ROC analysis. In C. Sammut, & G. I. Webb (Eds.), *Encyclopedia of machine learning and data mining* (pp. 1–8). Springer. https://doi.org/10.1007/978-1-4899-7502-7_739-1

Flora, D. B. (2020). Thinking about effect sizes: From the replication crisis to a cumulative psychological science. *Canadian Psychology/Psychologie Canadienne, 61*(4), 318–330. https://doi.org/10.1037/cap0000218

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychological Methods, 21*(1), 93–111. https://doi.org/10.1037/met0000064

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Gomer, B., Jiang, G., & Yuan, K.-H. (2019). New effect size measures for structural equation modeling. *Structural Equation Modeling, 26*(3), 371–389. https://doi.org/10.1080/10705511.2018.1545231

Groskurth, K., Bhaktha, N., & Lechner, C. M. (2022). *Making model judgments ROC(K)-solid: Tailored cutoffs for fit indices through simulation and ROC analysis in structural equation modeling.*

Groskurth, K., Bluemke, M., & Lechner, C. M. (2022a). *Why we need to abandon fixed cutoffs for goodness-of-fit indices: A thorough simulation and possible solutions.*

Groskurth, K., Bluemke, M., & Lechner, C. M. (2022b). *Measurement Invariance Violation Indices (MIVIs): Effect sizes for non-invariance of items and item sets.*

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. https://doi.org/10.1177/0013164410384856

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. https://doi.org/10.1037/a0024917

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

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

Jackson, D. L., Gillaspy Jr, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: an overview and some recommendations. *Psychological Methods, 14*(1), 6–23. https://doi.org/10.1037/a0014694

Jak, S., Jorgensen, T. D., Verdam, M. G., Oort, F. J., & Elffers, L. (2021). Analytical power calculations for structural equation modeling: A tutorial and Shiny app. *Behavior Research Methods, 53*(4), 1385–1406. https://doi.org/10.3758/s13428-020-01479-0

Jobst, L. J., Bader, M., & Moshagen, M. (2021). A tutorial on assessing statistical power and determining sample size for structural equation models. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000423

Jong, J., & Halberstadt, J. (2016). *Death anxiety and religious belief: An existential psychology of religion.* Bloomsbury.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Scientific Software International.

Kaplan, D. (1991). On the modification and predictive validity of covariance structure models. *Quality and Quantity, 25*(3), 307–314. https://doi.org/10.1007/BF00167535

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. https://doi.org/10.1207/S15328007SEM1003_1

Kim, H., & Millsap, R. (2014). Using the Bollen-Stine bootstrapping method for evaluating approximate fit indices. *Multivariate Behavioral Research, 49*(6), 581–596. https://doi.org/10.1080/00273171.2014.947352

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

Kolbe, L., Molenaar, D., Jak, S., & Jorgensen, T. D. (2022). Assessing measurement invariance with moderated nonlinear factor analysis using the R package OpenMx. *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000501

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research, 51*(2–3), 220–239. https://doi.org/10.1080/00273171.2015.1134306

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

MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research, 38*(1), 113–139. https://doi.org/10.1207/S15327906MBR3801_5

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

Mai, R., Niemand, T., & Kraus, S. (2021). A tailored-fit model evaluation strategy for better decisions about structural equation models. *Technological Forecasting and Social Change, 173*, Article 121142. https://doi.org/10.1016/j.techfore.2021.121142

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. https://doi.org/10.1016/j.paid.2006.09.023

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. https://doi.org/10.1207/s15328007sem1103_2

Maydeu-Olivares, A., & Shi, D. (2017). Effect sizes of model misfit in structural equation models: Standardized residual covariances and residual correlations. *Methodology, 13*(Suppl 1), 23–30. https://doi.org/10.1027/1614-2241/a000129

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

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000425

McNeish, D., & Wolf, M. G. (2022). Dynamic fit index cutoffs for one-factor models. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-022-01847-y

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. https://doi.org/10.1080/00223891.2017.1281286

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95*(4), 728–743. https://doi.org/10.1037/a0018966

Millsap, R. E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences, 42*(5), 875–881. https://doi.org/10.1016/j.paid.2006.09.021

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

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

Millsap, R. E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. Hoyle, D. Kaplan, G. A. Marcoulides, & S. West (Eds.), *Handbook of Structural Equation Modeling* (pp. 380–392). Guilford Press.

Mooney, C. Z. (1997). *Monte Carlo simulation.* Sage.

Moshagen, M., & Auerswald, M. (2018). On congruence and incongruence of measures of fit in structural equation modeling. *Psychological Methods, 23*(2), 318–336. https://doi.org/10.1037/met0000122

Moshagen, M., & Erdfelder, E. (2016). A new strategy for testing structural equation models. *Structural Equation Modeling, 23*(1), 54–60. https://doi.org/10.1080/10705511.2014.950896

Musil, C. M., Jones, S. L., & Warner, C. D. (1998). Structural equation modeling and its relationship to multiple regression and factor analysis. *Research in Nursing & Health, 21*(3), 271–281. https://doi.org/10.1002/(SICI)1098-240X(199806)21:3<271::AID-NUR10>3.0.CO;2-G

Muthén, L.K., & Muthén, B.O. (1998-2017). *Mplus user's guide* (8th ed). Muthén & Muthén.

Newton, I. (1999). *The principia: Mathematical principles of natural philosophy* (B. Cohen & A. Whitman, Trans.). University of California Press. (Original work published in 1687)

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika, 20A*(1), 175–240. https://doi.org/10.2307/2331945

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231*(694–706), 289–337. https://www.jstor.org/stable/91247

Niemand, T., & Mai, R. (2018). Flexible cutoff values for fit indices in the evaluation of structural equation models. *Journal of the Academy of Marketing Science, 46*, 1148–1172. https://doi.org/10.1007/s11747-018-0602-9

Nießen, D., Partsch, M., Groskurth, K. (2020). Data for: An English-language adaptation of the Risk Proneness Short Scale (R-1) (Version: 1.0.0). https://doi.org/10.7802/2080

Nye, C. D., & Drasgow, F. (2011a). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods, 14*(3), 548–570. https://doi.org/10.1177/1094428110368562

Nye, C. D., & Drasgow, F. (2011b). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology, 96*(5), 966–980. https://doi.org/10.1037/a0022955

Nye, C. D., Bradburn, J., Olenick, J., Bialko, C., & Drasgow, F. (2019). How big are my effects? Examining the magnitude of effect sizes in studies of measurement equivalence. *Organizational Research Methods, 22*(3), 678–709. https://doi.org/10.1177/1094428118761122

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis, 22*(1), 45–60. https://doi.org/10.1093/pan/mpt014

Oberski, D. L., Vermunt, J. K., & Moors, G. B. (2015). Evaluating measurement invariance in categorical data latent variable models with the EPC-interest. *Political Analysis, 23*(4), 550–563. https://doi.org/10.1093/pan/mpv020

Padgett, R. N., & Morgan, G. B. (2021). Multilevel CFA with ordered categorical data: A simulation study comparing fit indices across robust estimation methods. *Structural Equation Modeling*, *28*(1), 51–68. https://doi.org/10.1080/10705511.2020.1759426

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology, 6,* Article 223. https://doi.org/10.3389/fpsyg.2015.00223

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879 -903. https://doi.org/10.1037/0021-9010.88.5.879

Popper, K. (2002). *The logic of scientific discovery.* Routledge Classics.

Pornprasertmanit, S. (2014). *The unified approach for model evaluation in structural equation modeling* [Unpublished doctoral dissertation]. University of Kansas. http://hdl.handle.net/1808/16828

Pornprasertmanit, S. (2022). *A note on effect size for measurement invariance.* https://cran.r-project.org/web/packages/semTools/vignettes/partialInvariance.pdf

Pornprasertmanit, S., Miller, P., Schoemann, A., & Jorgensen, T. D. (2021). *simsem: SIMulated Structural Equation Modeling.* R package version 0.5-16. https://CRAN.R-project.org/package=simsem

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin, 132*(1), 1–25. https://doi.org/10.1037/0033-2909.132.1.1

Robitzsch, A., & Lüdtke, O. (2022). Some thoughts on analytical choices in the scaling model for test scores in international large-scale assessment studies. *Measurement Instruments for the Social Sciences, 4,* Article 9. https://doi.org/10.1186/s42409-022-00039-w

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton University Press.

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

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16*(4), 561–582. https://doi.org/10.1080/10705510903203433

Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. *Sociological Methodology, 17*, 105–129. https://doi.org/10.2307/271030

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika, 54*(1), 131–151. https://doi.org/10.1007/BF02294453

Savalei, V., & Dunn, E. (2015). Is the call to abandon p-values the red herring of the replicability crisis? *Frontiers in Psychology, 6*, Article 245. https://doi.org/10.3389/fpsyg.2015.00245

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23–74.

Schmalbach, B., Irmer, J. P., & Schultze, M. (2019). *ezCutoffs: Fit Measure Cutoffs in SEM*. R package version 1.0.1. https://CRAN.R-project.org/package=ezCutoffs

Shi, D., & Maydeu-Olivares, A. (2020). The effect of estimation methods on SEM fit indices. *Educational and Psychological Measurement, 80*(3), 421–445. https://doi.org/10.1177/0013164419885164

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. https://doi.org/10.1177/0013164418783530

Shi, D., Maydeu-Olivares, A., & DiStefano, C. (2018). The relationship between the standardized root mean square residual and model misspecification in factor analysis models. *Multivariate Behavioral Research, 53*(5), 676–694. https://doi.org/10.1080/00273171.2018.1476221

Sörbom, D. (1989). Model modification. *Psychometrika, 54*(3), 371–384. https://doi.org/10.1007/BF02294623

Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78–90. https://doi.org/10.1086/209528

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*(2), 173–180. https://doi.org/10.1207/s15327906mbr2502_4

Steinmetz, H. (2013). Analyzing observed composite differences across groups: Is partial measurement invariance enough? *Methodology*, *9*(1), 1–12. https://doi.org/10.1027/1614-2241/a000049

Supple, A. J., Su, J., Plunkett, S. W., Peterson, G. W., & Bush, K. R. (2013). Factor structure of the Rosenberg Self-Esteem Scale. *Journal of Cross-Cultural Psychology, 44*(5), 748–764. https://doi.org/10.1177/0022022112468942

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: an illustration using Mplus and the lavaan/semtools packages. *Structural Equation Modeling, 27*(1), 111–130. https://doi.org/10.1080/10705511.2019.1602776

Thoemmes, F., Rosseel, Y., & Textor, J. (2018). Local fit evaluation of structural equation models using graphical criteria. *Psychological Methods, 23*(1), 27–41. https://doi.org/10.1037/met0000147

Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *The Journal of Experimental Education, 80*(1), 26–44. https://doi.org/10.1080/00220973.2010.531299

Widaman, K. F., & Revelle, W. (2022). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods.* Advance online publication. https://doi.org/10.3758/s13428-022-01849-w

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology, 46*(2), 201–233. https://doi.org/10.1037/xlm0000732

Xia, Y., & Yang, Y. (2018). The influence of number of categories and threshold values on fit indices in structural equation modeling with ordered categorical data. *Multivariate Behavioral Research, 53*(5), 731–755. https://doi.org/10.1080/00273171.2018.1480346

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods, 51*(1), 409–428. https://doi.org/10.3758/s13428-018-1055-2

Yuan, K.-H., & Hayashi, K. (2003). Bootstrap approach to inference and power analysis based on three test statistics for covariance structure models. *British Journal of Mathematical and Statistical Psychology*, *56*(1), 93–110. https://doi.org/10.1348/000711003321645368

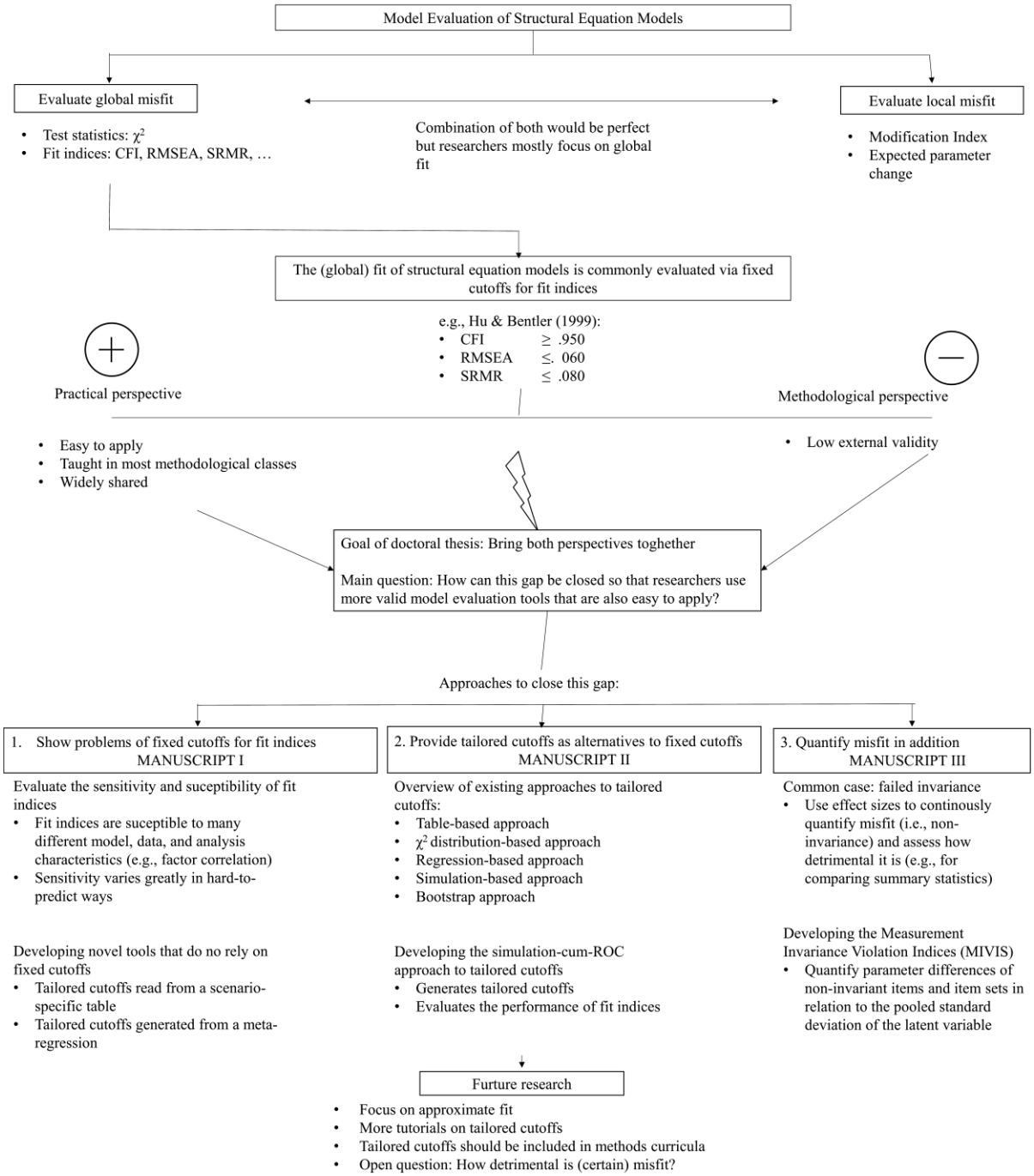
Yuan, K.-H., Bentler, P. M., & Chan, W. (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika*, *69*(3), 421–436. https://doi.org/10.1007/BF02295644

Yuan, K.-H., Hayashi, K., & Yanagihara, H. (2007). A class of population covariance matrices in the bootstrap approach to covariance structure analysis. *Multivariate Behavioral Research, 42*(2), 261–281. https://doi.org/10.1080/00273170701360662

5 Appendix

Integrative Concept of This Thesis

Figure 5.1: *Integrative Concept*



Examples of All Approaches Developed in This Thesis

To illustrate all approaches developed in this doctoral thesis, I use the example of the Rosenberg Self-Esteem Scale (Rosenberg, 1965). Rosenberg (1965) constructed ten items that are supposed to measure global self-esteem, with half of the items referring to positive feelings about the self and the other half to negative ones. Although constructed as a one-factor scale, several studies found a two-factor structure (or more complex structures, as outlined in Supple et al., 2013). For this example, I used publicly available data (Nießen et al., 2020) of a quota sample aged 18 to 69 from the United Kingdom (UK; $N = 468$). I included the R code for this example, which also contains details on the packages I used, in an OSF online repository (https://osf.io/vwjmq/).

As shown in Figure 5.2, I fit the two-factor model to the empirical data with MLR and recorded the empirical values of commonly used fit indices (i.e., $\chi^2(34) = 119.05$, $p < .001$, CFI = .947, RMSEA = .073, SRMR = .051). The classical way to evaluate whether fit indices point to good or bad fit of the model is to compare them with fixed cutoffs (e.g., Hu & Bentler, 1999; CFI around ≥ .95, RMSEA around ≤ .06, SRMR around ≤ .08). In this example of the two-factor Rosenberg Self-Esteem Scale model, only RMSEA exceeded its cutoff and, thus, pointed to bad fit. CFI was around its cutoff and, thus, acceptable. SRMR pointed to good fit. When relying on fixed cutoffs, I would conclude that the two-factor model fits the data. As outlined in this thesis, the values of fit indices depend on the setting of interest. Once-proposed cutoffs are not generalizable to empirical settings different from the initial simulated scenarios. Hu and Bentler (1999) did not base their cutoffs on a two-factor model with ten items but on a three-factor model with 15 items. Thus, their cutoffs were invalid for the present setting.

I developed a so-called table-based approach to tailored cutoffs in the first manuscript (Groskurth, Bluemke, & Lechner, 2022a). I obtained cutoffs (at a 5% Type I error rate) for several scenarios from a large-scale simulation study. I saved them in a large scenario-specific table (included in Additional File 4 of the first manuscript). I can read out the cutoff that best matches the empirical setting. According to the table-based cutoffs that best matched this empirical setting (CFI ≥ .992, RMSEA ≤ .027, SRMR ≤ .030), I rejected the two-factor model, as empirical fit index values indicated bad fit when compared against table-based cutoffs. Overall, the cutoffs' simulated scenarios were like the empirical setting investigated here. However, there were still some differences, as outlined in Figure 5.2: For instance, I investigated ten items in this empirical setting, but the closest simulated scenario contained 12

items. The simulation considered only a factor correlation of r = .30, whereas the current model's factors correlated at r = .71.

In the first manuscript (Groskurth, Bluemke, & Lechner, 2022a), I also developed regression formulae that allow the generation of tailored cutoffs for empirical settings not exactly equal to the simulated scenarios. The scenario-specific cutoff tables were the basis for the regression formulae. Thus, both originate from the same simulation study—the regression formulae just better accommodate settings different from the simulated scenarios through extrapolation. To obtain cutoffs, I plugged the model, data, and estimation characteristics from this empirical setting (as outlined in Figure 5.2) in the regression formulae (i.e., the R code from Additional File 5 of the first manuscript). The cutoffs (i.e., CFI = 1, RMSEA ≤ .033, SRMR ≤ .034) were relatively close to those from the table-based approach (i.e., CFI ≥ .992, RMSEA ≤ .027, SRMR ≤ .030). However, neither table-based nor regression-based cutoffs were simulated explicitly for the setting of interest. Neither were Type I error rates considered in addition to Type II error rates, nor were researchers given any guidance as to which fit indices were most reliable in the setting of interest.

To obtain information about the performance of fit indices and explicitly simulate cutoffs at balanced Type I and Type II error rates for the setting of interest, I needed to generate tailored cutoffs via the simulation-cum-ROC approach outlined in the second manuscript (Groskurth, Bhaktha, & Lechner, 2022). Here, I used the same example of the Rosenberg Self-Esteem Scale model as in the second manuscript. I also generated tailored cutoffs in the same way (i.e., with a two-factor model as the H_0 population model and a bi-factor model as the H_1 population model). The second manuscript contains all the details. Here, tailored cutoffs generated via the simulation-cum-ROC approach ($\chi^2(34) \leq 70.82$, CFI ≥ .984, RMSEA ≤ .047, SRMR ≤ .031) were very similar to those of the regression-based approach (CFI = 1, RMSEA ≤ .033, SRMR ≤ .034). They were also like those of the table-based approach (CFI ≥ .992, RMSEA ≤ .027, SRMR ≤ .030). This was certainly due to the similarity of this empirical setting to the simulated scenarios where table-based but also regression-based cutoffs originated from. If the settings were more dissimilar, table-based but also regression-based cutoffs would be more invalid and, thus, dissimilar to those of the simulation-cum-ROC approach.

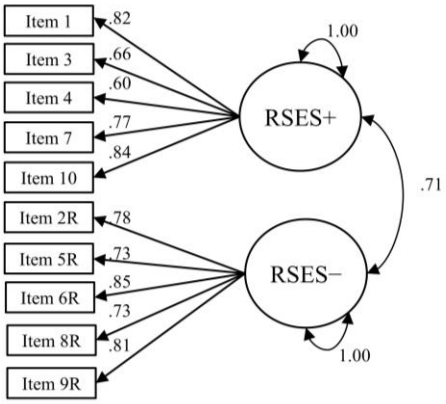
The simulation-cum-ROC approach suggested that all fit indices performed well for the setting of interest. The empirical values of fit indices compared to the cutoffs of the simulation-cum-ROC approach (in Figure 5.2) suggested that the two-factor model of the Rosenberg Self-Esteem Scale fit poorly. The simulation-cum-ROC approach generates the most tailored and,

thus, most valid cutoffs among the approaches presented in Figure 5.2. Unlike the fixed cutoffs that classified the model as correctly specified according to most fit indices (i.e., only RMSEA points to misspecification), tailored cutoffs classified the model as misspecified according to all fit indices in this example. The simulation-cum-ROC approach also revealed that all fit indices could equally discriminate between correctly and misspecified models as defined here. Thus, I did not need to prioritize any fit index in this setting.

Figure 5.2: *Testing the Two-Factor Model of the Rosenberg Self-Esteem Scale*

type of estimator: **MLR**
number of items: **10** (table-based: 12)
number of response options: **4** (table-based: 3 or 5)
distribution of response options: **asymmetric**
magnitude of standardized factor loadings: **Ø .76** (table-based: .80)
sample size: **468** (table-based: 500)
number of factors: **2**
factor correlation: **.71** (table-based, regression-based: .30)

Analysis model**: two-factor**
Population model relative to which the analysis model
… is correctly specified: **two-factor**
… is misspecified: **bi-factor**



Least suited for empirical setting of interest

Best suited for empirical setting of interest

Fit indices	χ^2	*df*	CFI	RMSEA	SRMR
Fixed cutoffs (e.g., Hu & Bentler, 1999)	non-significant	-	around ≥ .950	around ≤ .060	around ≤ .080
Table-based cutoffs	-	-	≥ .992	≤ .027	≤ .030
Regression-based cutoffs	-	-	= 1	≤ .033	≤ .034
Simulation-cum-ROC cutoffs[a]	**≤ 70.82**	34	**≥ .984**	**≤ .047**	**≤ .031**
Empirical values	**119.05*****	34	**.947**	**.073**	**.051**

Note. Standardized loadings. [a]AUC = 1 for all fit indices, Type I and Type II error rates = 0% for all cutoffs of the simulation-cum-ROC approach. *** $p < .001$. I identified the model by fixing the latent variables at 1 and latent means at 0.

So far, I have evaluated whether the model of interest fits the data via binary fit-misfit decisions (using cutoffs for fit indices). In the following, I move over to another context, the context of measurement invariance testing, where it is especially important to continuously quantify misfit (i.e., non-invariance), if present, in addition to binary fit-misfit decisions. Measurement invariance tests (e.g., Davidov et al., 2014; Millsap, 2011; Steenkamp & Baumgartner, 1998) evaluate the comparability of cross-group statistics. Quantifying non-invariance, if present, helps to investigate how biased cross-group comparisons are.

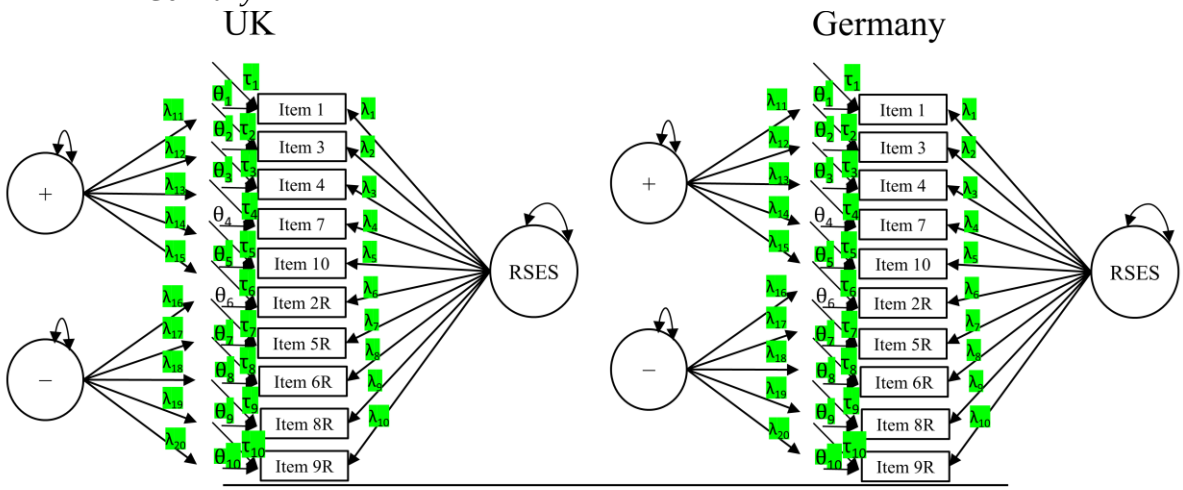
To conduct measurement invariance tests, I needed to find a well-fitting model. I modified the previously evaluated two-factor model and found evidence for a bi-factor model for the Rosenberg Self-Esteem Scale in the UK data. All details on this model modification (including the generation of tailored cutoffs via the simulation-cum-ROC approach) were included in the R code, stored in an OSF online repository (https://osf.io/vwjmq/). Suppose my goal was to evaluate the comparability of the bi-factor Rosenberg Self-Esteem Scale model across the UK and another country included in the data, Germany (N = 474). As outlined in Figure 5.3, I found evidence for configural invariance (i.e., equal item-to-factor structure), metric invariance (i.e., equal loadings), and scalar invariance (i.e., equal intercepts). The fit did not deteriorate strongly through the additional constraints on loadings and intercepts, according to CFI and RMSEA. The uniqueness model (i.e., equal loadings, intercepts, and residual variances) fit worse than the scalar model. When freely estimating two residual variances, the final partial uniqueness model fit as well as the scalar model. I accepted the partial uniqueness invariance, implying that latent (co)variances and observed means were comparable across the UK and Germany without bias (e.g., Chen, 2008; Steenkamp & Baumgartner, 1998; Steinmetz, 2013). The partial uniqueness model did not allow the comparison of observed variances as the two non-invariant residual variances could bias such comparisons (e.g., Steenkamp & Baumgartner, 1998). The non-invariant residual variances pertained to Item 7 (“I feel that I’m a person of worth, at least on an equal plane with others”) and Item 2R (“At times I think I am no good at all”). The R code contains further details on arriving at the partial uniqueness model.

An important question remained: How large is the non-invariance of the two residual variances identified through the partial uniqueness model? Measurement Invariance Violation Indices (MIVIs) developed in the third manuscript (Groskurth, Bluemke, & Lechner, 2022b) help to answer that question. I estimated MIVIs based on the pooled standard deviation of the general factor (loading on all ten items). MIVIs estimated at the item level showed that non-invariance due to Item 7 ($MIVI\text{–}U_{Item\ 7|absolute}$= 0.14) and Item 2R ($MIVI\text{–}U_{Item\ 2R|absolute}$=

0.13) was equally large. The signed MIVIs revealed a negative value for Item 7 and a positive for Item 2R. Thus, the non-invariance pattern was mixed; non-invariant residual variances were smaller in the UK than in Germany for Item 7 and larger in the UK than in Germany for Item 2R. Adding up all absolute non-invariant parameters differences within the item set (i.e., $MIVI\text{-}U_{Item\ set|absolute}$) amounted to an average bias of 0.03 pooled latent standard deviations. When non-invariant parameter differences were allowed to cancel out in the item set (i.e., $MIVI\text{-}U_{Item\ set|signed}$), they compensated for each other, amounting to an average bias of -0.00 pooled latent standard deviations. The confidence interval of $MIVI\text{-}U_{Item\ set|signed}$ included zero (i.e., $CI_{95\%}[-0.01; 0.01]$). Thus, although non-invariance was present, the non-invariant parameter differences in residual variances would not impact cross-country differences of observed scale score statistics.

This example showed that MIVIs provided more in-depth information on a model's cross-group comparability. Measurement invariance tests must not stop after the binary fit-misfit decision. Instead, one can quantify non-invariance, if present, with the aid of MIVIs to evaluate the strength of the model's incomparability—relevant to assessing the non-invariance bias in further analysis (e.g., comparisons of summary statistics).

Figure 5.3: *Investigating Measurement Invariance for the Rosenberg Self-Esteem Across the UK and Germany*



Invariance level	χ^2	*df*	CFI	RMSEA	SRMR
Configural	78.52**	50	.991	.035	.019
Metric	108.98**	67	.987	.036	.048
Scalar	121.93***	74	.986	.037	.048
Uniqueness	166.16***	84	.975	.046	.048
Partial uniqueness	133.62***	82	.984	.037	.046

		Absolute	Signed
Item level	Item 7	$MIVI\text{–}U_{Item\ 7\|absolute}$ = 0.14 $CI_{95\%}$[0.06; 0.22]	$MIVI\text{–}U_{Item\ 7\|signed}$ = −0.14 $CI_{95\%}$[−0.22; −0.06]
	Item 2R	$MIVI\text{–}U_{Item\ 2R\|absolute}$ = 0.13 $CI_{95\%}$[0.05; 0.21]	$MIVI\text{–}U_{Item\ 2R\|signed}$ = 0.13 $CI_{95\%}$[0.05; 0.21]
Item set level		$MIVI\text{–}U_{Item\ set\|absolute}$ = 0.03 $CI_{95\%}$[0.02; 0.04]	$MIVI\text{–}U_{Item\ set\|signed}$ = −0.00 $CI_{95\%}$[−0.01; 0.01]

Note. The final partial uniqueness model with green layered parameters fixed across groups and non-layered parameters freely estimated across groups is displayed. Bootstrap confidence intervals are smaller than they should be due to an internal error in the R package lavaan (version 0.6.12; Rosseel, 2012) that does not allow including the standard error of the latent variable's variance. I estimated MIVIs based on the pooled standard deviation of the general RSES (= Rosenberg Self Esteem Scale) factor and used the total number of items (i.e., ten) when estimating $MIVI\text{–}U_{Item\ set}$. ** $p < .010$, *** $p < .001$. I identified the model by fixing the latent variables at 1 and latent means at 0.

Statement of Originality

Eidesstattliche Versicherung gemäß § 9 Absatz 1 Buchstabe e) der Promotionsordnung der Universität Mannheim zur Erlangung des Doktorgrades der Sozialwissenschaften:

1. Bei der eingereichten Dissertation mit dem Titel *Why one size does not fit all: Evaluating the validity of fixed cutoffs for model fit indices and developing new alternatives* handelt es sich um mein eigenständig erstelltes eigenes Werk.
2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtliche Zitate aus anderen Werken als solche kenntlich gemacht.
3. Die Arbeit oder Teile davon habe ich bisher nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.
4. Die Richtigkeit der vorstehenden Erklärung bestätige ich.
5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt.

Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Copies of Manuscripts

MANUSCRIPT I

Groskurth, K., Bluemke, M., & Lechner, C. M. (2022a). *Why we need to abandon fixed cutoffs for goodness-of-fit indices: A thorough simulation and possible solutions.* Invited revision at *Behavior Research Methods*.

MANUSCRIPT II

Groskurth, K., Bhaktha, N., & Lechner, C. M. (2022). *Making model judgments ROC(K)-solid: Tailored cutoffs for fit indices through simulation and ROC analysis in structural equation modeling.* Invited revision at *Psychological Methods*.

MANUSCRIPT III

Groskurth, K., Bluemke, M., & Lechner, C. M. (2022b). *Measurement Invariance Violation Indices (MIVIs): Effect sizes for non-invariance of items and item sets.* Manuscript in preparation.

Why we need to abandon fixed cutoffs for goodness-of-fit indices: A thorough simulation and possible solutions

Katharina Groskurth[1,2], Matthias Bluemke[1], and Clemens M. Lechner[1]

[1] GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany

[2] Graduate School of Economic and Social Sciences, University of Mannheim, Germany

Abstract

To evaluate model fit in confirmatory factor analysis, researchers compare goodness-of-fit indices (GOFs) against fixed cutoff values (e.g., CFI > .95) derived from simulation studies. Methodologists have cautioned that cutoffs for GOFs are only valid for settings similar to those covered in the simulation scenarios from which cutoffs originated—which contrasts with their widespread use across various empirical settings. This research comprehensively addresses the limited generalizability of fixed cutoffs for commonly used GOFs (i.e., χ^2, χ^2/df, CFI, RMSEA, SRMR) by following two paths. First, we conducted the most thorough simulation study to date on the sensitivity of GOFs to model misspecification (i.e., misspecified factor dimensionality and unmodeled cross-loadings) and their susceptibility to further data and analysis characteristics (i.e., estimator, number of indicators, number and distribution of response options, loading magnitude, sample size, and factor correlation). We integrated all characteristics in our simulation study that had been identified as influential in previous studies. Our simulation enabled us to replicate well-known issues with GOFs but also to uncover several previously unknown or at least underappreciated issues. Especially the factor correlation moderated several effects on GOFs. Second, we discussed several strategies for assessing model fit that take the context dependency of GOFs into account. We argued that tailored cutoffs are the way forward. Based on the large-scale simulation study, we generated large tables with scenario-specific cutoffs and regression formulae to predict cutoffs tailored to several empirical settings.

Are Goodness-of-Fit Indices Any Good? Why We Cannot Be so Sure

In social and behavioral science research, researchers commonly employ goodness-of-fit indices (GOFs) to evaluate the fit of latent variable models such as confirmatory factor analysis (CFA) models. The most widely used GOFs (e.g., Jackson et al., 2009) are the chi-square test statistic divided by the model degrees of freedom (χ^2/df), the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR). In addition, researchers often rely on the traditional chi-square test statistic of exact model fit (χ^2). Although strictly speaking, it is not a GOF but a formal test, researchers use χ^2 in much the same way as they use GOFs (see also Jöreskog & Sörbom, 1993), which is why we henceforth subsume it under the same rubric.

Cutoffs for GOFs, on which binary decisions about accepting or rejecting a model rest, were derived from simulation studies. Simulation studies represent highly controlled settings in which—different from the analysis of real data—researchers know and have control over the population (or data-generating) model. Researchers specify a population model, simulate data based on that model, and introduce model misspecification of known strength in the analysis model. Then, they examine how GOFs respond to such misspecification. Based on the distribution of GOFs, researchers derive cutoffs for these GOFs so that a critical level of misspecification leads to model rejection. What constitutes a "critical" level of misspecification, and hence a reasonable cutoff, is a somewhat arbitrary decision. In the past two decades, the cutoffs suggested by Hu and Bentler (1999) have been the most prominent and widely used ones, with their article boasting more than 95,000 citations in GoogleScholar at the time of this writing. According to these authors, CFI $\geq$ .950, RMSEA $\leq$ .060, and SRMR $\leq$ .080 point to good model fit. Reußner (2019) and Rutkowski and Svetina (2014) have proposed similar cutoffs. Though based on statistical principles rather than derived from simulation studies, the observed χ^2 value should not exceed a critical χ^2 value to indicate a well-fitting model; the critical χ^2 value varies with the model degrees of freedom (Bollen, 1989; see Moshagen & Erdfelder, 2016, for additional suggestions on critical values). Ullman (2014) suggested that a ratio of χ^2/df below 2 indicates a well-fitting model.

However, there are severe problems with relying on fixed cutoffs for GOFs in model evaluation (e.g., McNeish & Wolf, 2021). The key underlying issue is that simulation studies can only cover a limited set of scenarios. These scenarios are far from covering all possible

combinations of data and analysis characteristics that researchers will encounter in applied settings. If GOFs only react to model misspecification predictably and uniformly, the constraints of simulation studies will not pose a major problem. If, by contrast, GOFs react not only to misspecification but also to other characteristics of the data and analysis, their validity for judging the model fit may be severely compromised. We henceforth refer to the undesirable dependence of GOFs on data and analysis characteristics as *susceptibility* and contrast it with their desirable *sensitivity* to misspecification. Such susceptibility to data and analysis characteristics of GOFs is no purely hypothetical concern. Although GOFs are designed to quantify the degree of model misspecification and should ideally not react to any data or analysis characteristic, they apparently do so as identified in several studies (for an overview, see Niemand & Mai, 2018).

It follows that established cutoffs for GOFs are sufficiently certain to be valid only in empirical settings (i.e., combinations of data and analysis characteristics) that closely resemble the scenarios covered by the simulations from which these cutoffs were derived. The range of scenarios covered by existing simulations is dwarfed by the diversity and complexity of empirical settings encountered in applied research. Consequently, cutoffs for GOFs may lack external validity, and blindly applying the same set of cutoffs to many different empirical settings can mislead researchers into erroneous conclusions about model fit and substantive questions.

Unfortunately, current reporting practice shows that researchers apply the same cutoffs in the presence of data or analysis characteristics that can differ markedly from the ones in the simulation studies (for an overview, see Jackson et al., 2009; McNeish & Wolf, 2021). It appears that concerns regarding overgeneralizations of cutoffs (e.g., Heene et al., 2011; Markland, 2007; Marsh et al., 2004; McNeish & Wolf, 2021; Niemand & Mai, 2018; Nye & Drasgow, 2011) have gone largely unheeded. The widespread, or in fact near-universal, practice of relying on (fixed) cutoffs for GOFs in model evaluation is alarming, given the ongoing uncertainty about the applicability of fixed cutoffs for GOFs to scenarios hitherto uncharted by simulation studies.

Just *how* problematic is the practice of using fixed cutoffs for GOFs? And *what* can be used as an alternative to fixed cutoffs? In the following, we review extant research on the susceptibilities of GOFs to data and analysis characteristics. Following our review of the susceptibilities, we present a thorough simulation study that integrates, replicates, and extends previous simulation studies and represents the most thorough simulation on the sensitivity and

susceptibility of GOFs obtained from CFA models to date. We discuss several time-honored and promising emerging alternatives for model fit evaluation that do not rely on fixed cutoffs. We argue that cutoffs need to be tailored to the empirical setting. Based on the large-scale simulation study, we generated user-friendly tables with scenario-specific cutoffs and developed regression formulae to predict cutoffs for several empirical settings.

Susceptibilities of GOFs to Data and Analysis Characteristics: A Review of Previous Findings

GOFs are intended to enable evaluations of model fit, specifically, to help detect non-negligible model misspecification.[1] However, previous investigations showed that GOFs are susceptible to various data and analysis characteristics other than model misspecification (e.g., Beauducel & Herzberg, 2006).[2] These are the sample size (e.g., DiStefano et al., 2019), type of estimator (e.g., Xia & Yang, 2019), the number of indicators[3] (e.g., Kenny & McCoach, 2003), number and distribution of response options (e.g., Xia & Yang, 2018), the magnitude of factor loadings (e.g., Heene et al., 2011), and the factor correlation (e.g., Beauducel & Wittmann, 2005).

[1] Researchers often assume that GOFs can detect all types of misspecification. As Hayduk (2014) demonstrated, χ^2, which is incorporated in χ^2/df, CFI, and RMSEA, cannot detect any misspecification in certain constellations of population and analysis models. The analysis model may appear to fit perfectly, although a different population model has generated the data. We hereby acknowledge the existence of close-fitting models that are seriously misspecified.

[2] We term the influences of data and analysis characteristics on GOFs as "problems" or "susceptibilities," even though many of the problems are natural (and sometimes even desirable) consequences of the statistical properties of GOFs. Especially the dependence of χ^2 on sample size is readily comprehensible. As a strict and formal test, rather than a GOF, χ^2 depends on the model degrees of freedom. Per definition, the power of χ^2 to detect model misspecification increases as the sample size grows (e.g., Moshagen & Erdfelder, 2016). From the perspective of applied researchers, it would be desirable for GOFs to quantify the degree of model misspecification across many data and analysis characteristics, irrespective of other considerations such as sample size or other empirical features. That is, GOFs should ideally reflect model misspecification only—any other influences on GOFs are undesirable (e.g., Schermelleh-Engel et al., 2003). Therefore, we label any influences on GOFs other than misspecification as problematic from an applied researcher's perspective.

[3] Adding indicators to the model is one way to vary the model complexity. Beauducel and Herzberg (2006) and Fan and Sivo (2007), for instance, varied the model complexity by changing the number of indicators *and* the number of factors. Moshagen (2012) and Shi, DiStefano, et al. (2018) showed that the number of indicators rather than the number of factors drive the effects of model complexity on model fit.

The impact of those characteristics differed for correctly specified and misspecified models. For correctly specified models, GOFs (i.e., χ^2, χ^2/df, CFI, RMSEA, and SRMR) typically signaled better model fit with increasing sample size (e.g., Beauducel & Herzberg, 2006; Chen et al., 2008; DiStefano et al., 2019; Kenny et al., 2015; Sharma et al., 2005; Shi et al., 2019). Likewise, GOFs (i.e., CFI and SRMR) of correctly specified models pointed to better fit with a higher magnitude of factor loadings (and a lower magnitude of residual variances, Beierl et al., 2018; Heene et al., 2011; Shi et al., 2019). GOFs (i.e., CFI, RMSEA, and SRMR) also signaled better model fit with a symmetric instead of an asymmetric response distribution (Reußner, 2019). The influence of the number of indicators on GOFs of correctly specified models interacted with the sample size: At small sample sizes (e.g., $N = 100$), GOFs (i.e., χ^2/df, CFI, and RMSEA) indicated worse model fit when indicators of similar psychometric quality were added (Kenny & McCoach, 2003; see also Sharma et al., 2005; Shi et al., 2019). At large sample sizes (N = 1,000), GOFs (i.e., χ^2/df and RMSEA) pointed to better model fit as the number of indicators increased (only CFI was no longer affected; Kenny & McCoach, 2003). Per statistical definition, χ^2 increases when adding indicators without further restrictions to the model (Bollen, 1989). Only the magnitude of factor covariance/correlation in correctly specified multidimensional models seemed to be something GOFs (i.e., χ^2, CFI, RMSEA, and SRMR) are impervious to (Beauducel & Herzberg, 2006; Beierl et al., 2018).

For misspecified models, studies found that GOFs (i.e., χ^2, χ^2/df, CFI, and SRMR[1]) typically signaled worse model fit with an increasing number of indicators (only RMSEA was affected vice versa, e.g., DiStefano et al., 2019; Kenny & McCoach, 2003; Savalei, 2012; Shi & Maydeu-Olivares, 2020; Shi et al., 2019). Likewise, GOFs (i.e., χ^2, RMSEA, and SRMR) of misspecified models showed worse model fit with a higher magnitude of factor loadings (and a lower magnitude of residual variances)—only CFI was affected inconsistently across studies (Beierl et al., 2018; Hancock & Mueller, 2011; Heene et al., 2011; McNeish et al., 2018; Shi et al., 2019; Shi & Maydeu-Olivares, 2020; Shi, Maydeu-Olivares, & DiStefano, 2018; cf. Moshagen & Auerswald, 2018, who kept the degree of misspecification and residual error variances constant). GOFs of misspecified models also suggested worse model fit with a symmetric instead of an asymmetric response distribution (Reußner, 2019; Xia & Yang,

[1] Shi, Maydeu-Olivares, and DiStefano (2018) only found the effect of model size on SRMR for models with unmodeled cross-loadings but not misspecified factor dimensionality.

2018).[1] Similarly, GOFs (i.e., χ^2 and SRMR) of models with uncorrelated factors pointed to worse fit than with correlated factors for certain misspecification (i.e., with unmodeled cross-loadings that all have the same sign; Beauducel & Wittmann, 2005). The influence of the sample size on GOFs of misspecified models was mixed: χ^2, χ^2/df, and RMSEA indicated worse model fit with increasing sample size, whereas CFI and SRMR suggested better model fit (e.g., Beauducel & Wittmann, 2005; DiStefano et al., 2019; Nye & Drasgow, 2011).

GOFs also depended directly on the estimator used. Researchers mainly apply maximum likelihood (ML; Bollen, 1989) or its robust cousin MLR that corrects the χ^2 test statistic and standard errors of ML-estimated parameters for non-normality (L. K. Muthén & B. O. Muthén, 1998-2017; Yuan & Bentler, 2000). Both imply parameter estimation based on unstandardized covariances or Pearson correlations. Diagonally weighted least squares (DWLS) based on polychoric correlations or its mean and variance adjusted (WLSMV) χ^2 test statistic and standard errors are less commonly applied (B. Muthén, 1984; B. Muthén et al., 1997). However, they are gaining relevance as more and more researchers note their utility for ordered-categorical data, such as data from rating scales (for an overview of the estimation procedures, see Li, 2016). Generally, the DWLS-/WLSMV-based GOFs (i.e., χ^2, CFI, and RMSEA) pointed to better model fit than the ML-based ones (Beauducel & Herzberg, 2006; Nye & Drasgow, 2011; Xia & Yang, 2019)—for correctly specified and misspecified models.[2] The effect was only reversed for SRMR; it indicated worse fit with DWLS than ML for correctly specified models (Beauducel & Herzberg, 2006). The type of estimator also influenced other effects: DWLS/WLSMV-based GOFs (i.e., χ^2, χ^2/df, CFI, and RMSEA) generally suggested worse fit with a higher (compared to a lower) number of response options—for correctly specified and misspecified models (Beauducel & Herzberg, 2006; Xia & Yang, 2018).

[1] In particular, Reußner (2019) found that CFI, RMSEA, and SRMR were susceptible to the type of the response distribution when using estimators that assume multivariate normal and continous data (i.e., maximum likelihood, ML). For estimators that make no assumption about the underlying response distribution (i.e., diagonally weighted least squares, DWLS), Xia and Yang (2018) mathematically derived that the number and distribution of response options directly influence GOFs (i.e., χ^2, χ^2/df, CFI, and RMSEA). Both characteristics determine the precision of polychoric correlations that features in the fit function of DWLS and the mean and variance adjustment (WLSMV) of the χ^2 test statistic, which transfers to χ^2/df, CFI, and RMSEA.

[2] Savalei (2020) proposed an analytical correction to DWLS-/WLSMV-based GOFs to make them appear like ML-based ones, which has not yet been implemented in major statistical programs like R or Mplus.

Thus, there are already several indications that GOFs are strongly susceptible to extraneous influences (other than misspecification). However, no prior simulation study has investigated all aforementioned influences on GOFs in tandem. Instead, the focus was mainly on one or two of those influences. For instance, research has repeatedly focused on the effects of different magnitudes of factor loadings on GOFs (e.g., Beierl et al., 2018; Heene et al., 2011; Shi et al., 2019). In turn, research often investigated the effects of the number of response options and type of estimator on GOFs in tandem (e.g., Xia & Yang, 2019).

As no prior simulation study has integrated all aforementioned influences on GOFs, it remains unclear how susceptible GOFs are to the joint influences of these characteristics. This includes not only the presence of multiple main effects but also how interaction terms (e.g., sample size × number of response options) attenuate or aggravate any known biases of GOFs. It remains unclear which effects on GOFs replicate when several characteristics are assessed jointly that have been identified as influential previously. The integration of several characteristics might reveal new influential patterns (i.e., relevant interaction effects) that could further extend the literature on the susceptibility of GOFs. Although a comprehensive study soon reaches a high level of complexity, such replication-extension studies are uniquely important for the cumulative scientific process (Bonett, 2012). They aid in exposing misleading findings from prior studies through replication checks, generalizing effect sizes across simulation scenarios, and assessing moderator effects through interaction terms.

The Present Simulation

Aims of the Simulation

In our Monte Carlo simulation study (for more details on Monte Carlo simulations, see Boomsma, 2013), we aim to replicate and extend previous findings from simulation studies. In particular, we look closely at the joint impact of a wide range of data and analysis characteristics on GOFs. We focus on CFA models here. CFA models (i.e., measurement models) are uniquely important within the latent variable modeling framework, as they are the basis for many structural models.

Design of the Simulation

To ensure external validity, we designed our simulation to cover realistic scenarios typically encountered in behavioral and social science research. Each scenario comprised a population

model with various characteristics. It included 1,000 randomly drawn samples of varying sizes based on that population model. Additionally, it incorporated a correctly specified or misspecified analysis model that we fit to each randomly sampled data through different estimators.

We focused on different combinations of population and analysis models to cover a breadth of models that may occur in real-world settings. In the first combination, the population model was either a one-factor or correlated two-factor model. Factor correlations were r = .70, .50, or .30 to include different magnitudes of misspecification (i.e., a difference of .30, .50, or .70, when viewed from two perfectly correlating factors r = 1, which essentially subsume to a single factor). We fit a one-factor analysis model to data generated from both population models (i.e., one-factor and correlated two-factor). Thus, each analysis model was either correctly specified or misspecified regarding factor dimensionality. Figure 1 shows exemplary population and analysis models for the factor dimensionality scenarios.

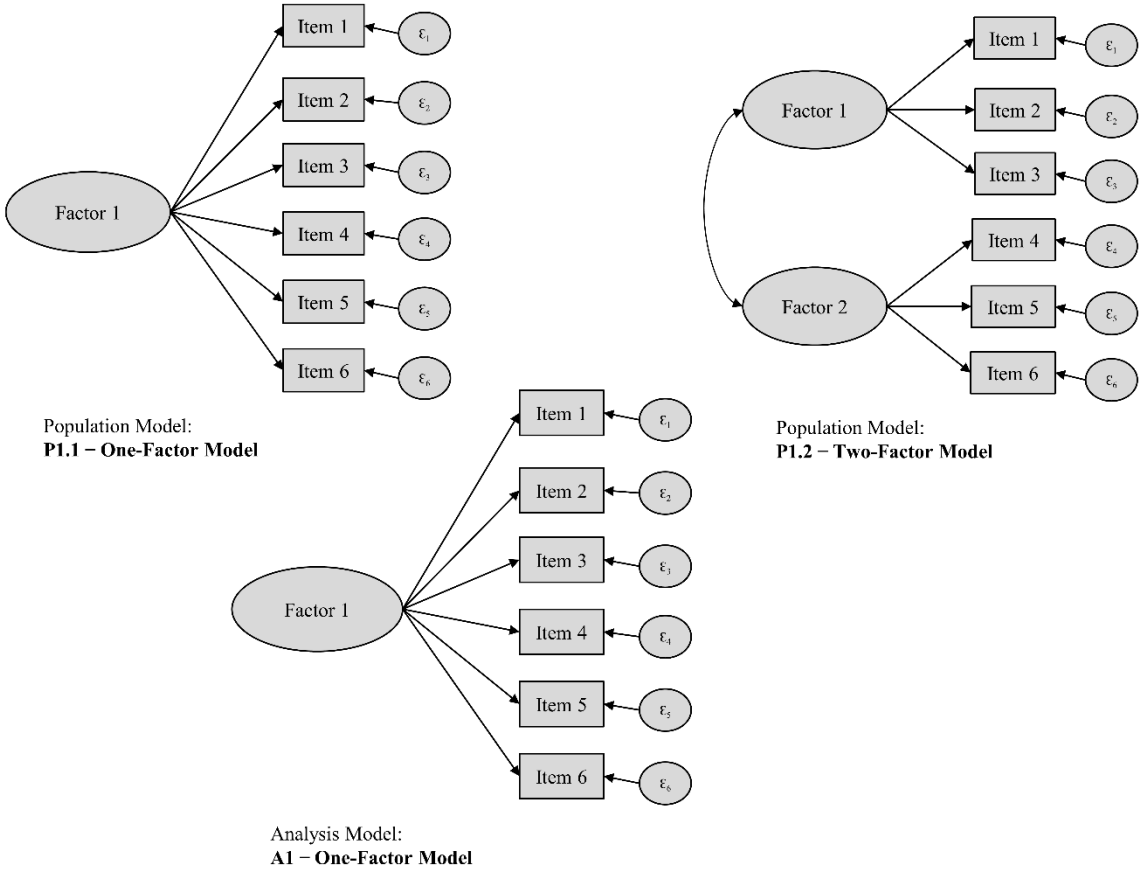
In the second combination, the population model was a two-factor model, either without or with cross-loadings. To include different magnitudes and proportions of misspecification, 17% or 33% of all items had cross-loadings with a standardized size of .20 or .30. Cross-loadings were only present on one of the two factors. We fit a two-factor analysis model without any cross-loadings to data generated from both types of population models (i.e., without and with cross-loadings). Thus, each analysis model was either correctly specified or misspecified regarding cross-loadings. Figure 2 shows exemplary population and analysis models for the cross-loading scenarios.

In both combinations, we varied six data and analysis characteristics in total: type of estimator, number of indicators, number of response options, distribution of response options, loading magnitude, and sample size.[1] With either correctly specified or misspecified models regarding cross-loadings, we also varied the factor correlation (i.e., factors were either correlated or uncorrelated). Depending on the factor correlation, the two factors of the

[1] To obtain ordered categorical indicators and determine the shapes of the resulting response distribution (i.e., symmetric or asymmetric), we cut the initially continuous data by setting different thresholds. To simulate a symmetric distribution of responses, we set thresholds to produce three (thresholds/z-values: −0.75, +0.75; with corresponding frequency percentages: 23%, 54%, 23%), five (thresholds/z-values: −1.20, −0.40, +0.40, +1.20; percentages: 12%, 23%, 31%, 23%, 12%), or seven equidistant response options (thresholds/z-values: −1.25, −0.75, −0.25, +0.25, +0.75, +1.25; percentages: 11%, 12%, 18%, 20%, 18%, 12%, 11%). To simulate an asymmetric response distribution, we shifted these response options to thresholds/z-values of +0.00, +1.04 (percentage: 50%, 35%, 15%) for the scenario with three response options; −0.39, +0.31, +0.74, +1.28 (percentage: 35%, 27%, 15%, 13%, 10%) for five response options; and −0.52, +0.00, +0.35, +0.64, +0.99, +1.40 (percentage: 20%, 20%, 15%, 15%, 10%, 10%, 10%) for seven response options.

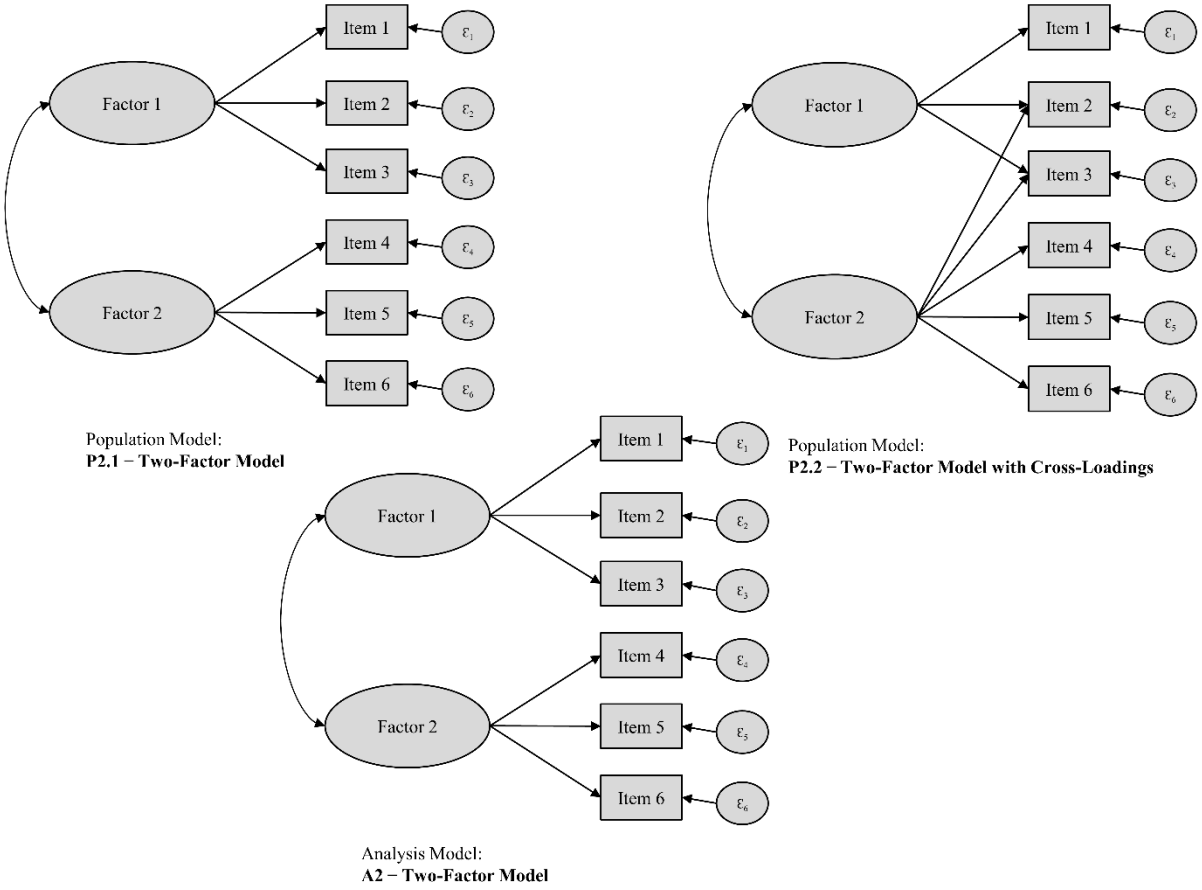
population and analysis models were either allowed to correlate or forced to be uncorrelated. With a misspecified factor dimensionality, the factor correlation confounds with the misspecification; we, thus, cannot include the factor correlation as an independent characteristic. Table 1 summarizes the different scenarios analyzed in this study—that were orientated upon typical settings encountered in empirical research.

Figure 1: *Exemplary Population and Analysis Models for the Factor-Dimensionality Scenarios*



Note. We chose to illustrate a model with six indicators for exemplary purposes here.

Figure 2: *Exemplary Population and Analysis Models for the Cross-Loading Scenarios*



Note. We chose to illustrate a model with six indicators, correlated factors, and two cross-loadings (i.e., cross-loadings exist for 33% of all six indicators) for exemplary purposes here.

Table 1: *Simulation Scenarios*

Realization					
	For all Population Models: Factor Variances = 1 Residual Variances = $1 - (Var(F1) \times \lambda_{F1}^2 + Var(F2) \times \lambda_{F2}^2 + 2 \times \lambda_{F1} \times \lambda_{F2} \times Cov(F1, F2))$ Replications = 1,000				
Characteristic	(1) Factor Dimensionality		(2) Cross-Loadings		Literature on typical settings used for operationalization
Population Model	One-Factor Model	Two-Factor Model	Two-Factor Model	Two-Factor Model w/ Cross-Loadings	
Analysis Model	One-Factor Model	One-Factor Model	Two-Factor Model	Two-Factor Model w/o Cross-Loadings	
Specification	Correct	Misspecified	Correct	Misspecified	
Magnitude of Misspecification	.00	.30, .50, .70	.00	.20, .30	
Proportion of Misspecification	0.00 (= 0%)	1.00 (= 100%)	0.00 (= 0%)	0.17, 0.33 (= 17%, 33%)	
Estimator	ML, MLR[c], DWLS, WLSMV	ML, MLR[c], DWLS, WLSMV	ML, MLR[c], DWLS, WLSMV	ML, MLR[c], DWLS, WLSMV	
Number of Indicators	6, 12	6, 12	6, 12	6, 12	Rammstedt & Beierlein (2014)
Response Options	3, 5, 7	3, 5, 7	3, 5, 7	3, 5, 7	Clark & Watson (2019); Simms et al. (2019)
Distribution of Responses[a]	Symmetric (skew=0.00), Asymmetric (skew=0.65)	Symmetric (skew=0.00), Asymmetric (skew=0.65)	Symmetric (skew=0.00), Asymmetric (skew=0.65)	Symmetric (skew=0.00), Asymmetric (skew=0.65)	Blanca et al. (2013)
Loading Magnitude	.40, .60, .80	.40, .60, .80	.40, .60, .80	.40, .60, .80	Soto & John (2017)
Sample Size	200, 500, 2,000	200, 500, 2,000	200, 500, 2,000	200, 500, 2,000	Bilsky et al. (2011); Comrey & Lee (1992); Nießen et al. (2019)
Factor Correlation	–	–	.00, .30 (factors not allowed/allowed to correlate)	.00, .30 (factors not allowed/allowed to correlate)	Groskurth et al. (2021); Kim et al. (2021); Lee & Cagle (2017); Soto & John (2017)
Total Number of Scenarios	432 (n = 432,000)	1,296 (n = 648,000)	864 (n = 864,000)	3,456 (n = 3,456,000)	
	1,728 (n = 1,728,000)		4,320 (n = 4,320,000)		
	6,048 (N = 6,048,000)				
Resampled Data[b]	7%				
Non-Convergence	2% (Final N = 5,956,844)				

Note. F1 = first factor; F2 = second factor; Var = variance; λ = factor loading; Cov = covariance; N = total number of data sets, n = subset of data sets. [a]For all scenarios: Excess kurtosis ≈ −0.80. [b]We had to re-simulate data whenever cell frequencies for any response option of any indicator resulted in fewer than five data points because DWLS/WLSMV can only estimate thresholds for response options that do contain observations. [c]We analyzed the commonly used Yuan and Bentler (2000) correction of the χ^2 test statistic here.

In our simulation study, all factors in the population models were latent variables with unit variance. Residual variances of the observed indicators depended on the population model parameters to obtain standardized observed indicators. Analysis models achieved identification by fixing the loading of the first indicator of each factor to unity. Unlike ML estimation, DWLS (and, accordingly, also WLSMV) include thresholds in the model parameterization that define the utilization of response options depending on the standing of the latent variable. To identify the DWLS/WLSMV-based analysis model, we followed the procedure Millsap (2011) outlined in line with the theta parameterization. Unlike delta parameterization, which fixes the residual variances of the intermediate continuous latent response variables to one, theta parameterization scales their distribution by fixing their variances to one.

We considered the following GOFs: χ^2 (Bollen, 1989) χ^2/df, CFI (Bentler, 1990; see also Widaman & Thompson, 2003), RMSEA (Steiger, 1990; see also Chen, 2007), and SRMR (Bentler, 1995; Hu & Bentler, 1999). Generally, GOF values closer to zero point to bad fit, except for CFI where values closer to 1 point to good fit. We did not include the computational details here, but interested readers will find them in the above-cited papers.

The final analysis contained GOFs for N = 5,956,844 converged models. We used R 3.6.3 (R Core Team, 2020) for all analyses. We documented all used R packages in our R code. Two packages were particularly central to our analyses: We generated data with MASS 7.3-53 (Venables & Ripley, 2002) and fit the analysis models to the data with lavaan 0.6-7 (Rosseel, 2012). We took all GOFs from the lavaan output except for the "manual" computation of χ^2/df. We set seeds within the R-code for complete reproducibility and monitored the R-package versions via renv 0.12.2 (Ushey, 2020). We did not preregister the design and analysis of this study. The full code is available on the Open Science Framework (OSF) at https://osf.io/e6kxa/?view_only=946034c00dee431897f67ca7ded58918.

Statistical Analyses

The outcomes of interest were the *sensitivity* of GOFs to model misspecification and their *susceptibility* to influences other than model misspecification, such as the type of estimator or sample size. We analyzed the sensitivity and susceptibility via descriptive and inferential statistics along four main steps. First and foremost, we inspected the distributions of GOFs across the different scenarios. Second, we looked at zero-order correlations (Kendall's tau-b to account for ordinal level data) between the GOFs and simulation characteristics to get a first impression of their sensitivity and susceptibility. Third, we looked at the characteristics' main and interaction effects on GOFs, including linear and quadratic terms, in multivariate regression. The multivariate regression conveys the size of any effects while preserving the natural units of the variables (i.e., unstandardized beta-weights) to ease the interpretation.

We limited the multivariate regression to only include two-way interactions for three reasons: First, we faced a technical restriction regarding higher levels of interactions. We had to resort to the biglm function from the biglm package in R (Lumley, 2013) explicitly designed to handle big data. However, the biglm function is limited regarding the number of independent variables, including interaction effects. As our thorough analyses comprised large data and many variables, these technical restrictions confined our analysis to two-way interactions. A second reason was that the purpose of the regression is only to solidify, from a multivariate perspective, and quantify the various influences on GOFs that simpler analysis (e.g., the descriptive statistics) might suggest. A two-way interaction already suffices to show whether GOFs are subject to complex influences of various characteristics. Yet another reason why we focused on two-way interactions is to preserve straightforward interpretability. Whereas two-way interactions are readily interpretable, three- or even four-way interactions would complicate matters beyond a point where they add much value.

Fourth and finally, we visually inspected selected large influences on GOFs. We selected those characteristics for visualization, which appeared to have a large (or complex) impact on GOFs in the preceding analyses. The visualization permits in-depth interpretation of higher-order interaction effects. The final visualization of the most relevant influences is key to our analysis.

Simulation Results

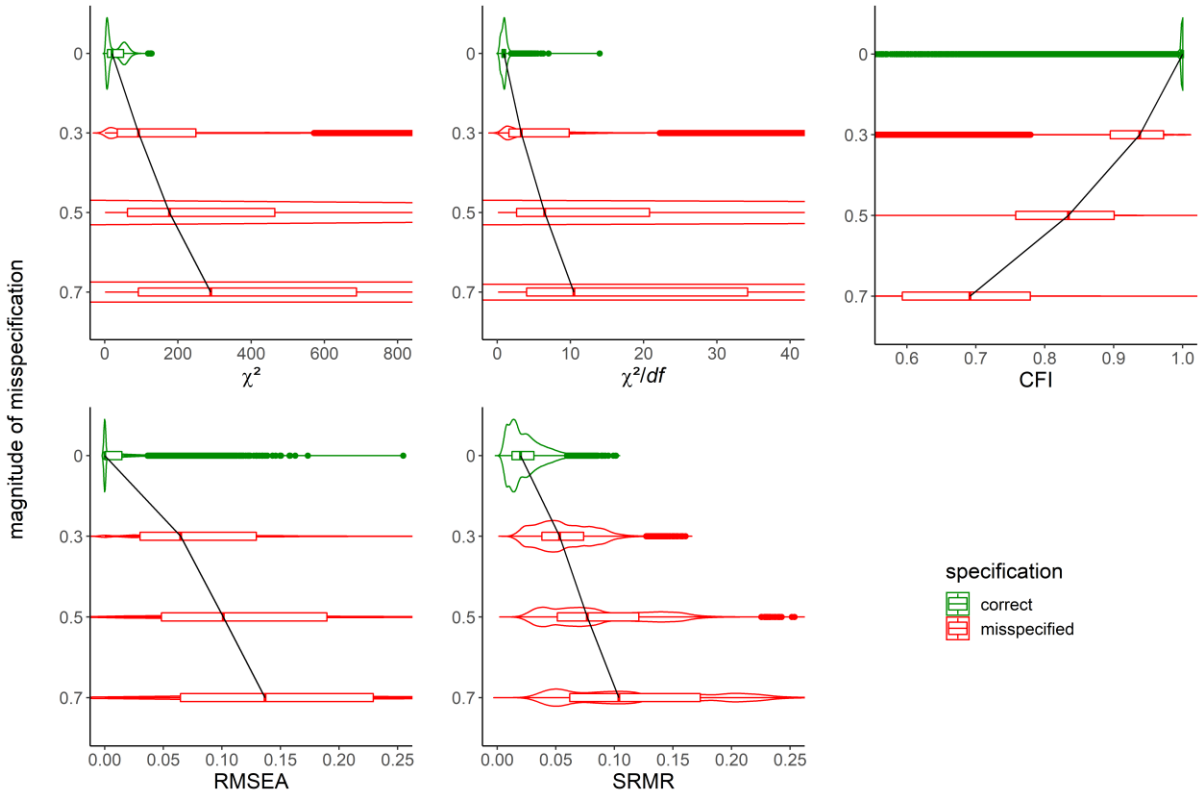
Sensitivity of GOFs: Descriptive Statistics

We first inspected how GOFs distribute across correctly specified and misspecified models in different scenarios (pooled across all simulation characteristics). Figure 3 compares the GOF distributions (i.e., χ^2, χ^2/df, CFI, RMSEA, and SRMR) as violin plots for either correctly specified or misspecified models regarding factor dimensionality (i.e., one-factor analysis models for either a one-factor or two-factor population model). The Y-axis represents different degrees of severity of the misspecification, with the correctly specified model as a point of reference shown on top in green. The X-axis represents the relevant range of values for each GOF. The black trace line vertically connects the GOF medians from different scenarios to reflect trends. We displayed each GOF in its original metric and direction. Similarly, Figure 4 shows the magnitude and proportion of cross-loadings in the population model that went unmodeled in the analysis model. We further split the figure into uncorrelated and correlated factor scenarios (factor correlation = .00 or .30) shown in Panels A and B, respectively. Tables A1 and A2 in Additional File 1 of the Supplementary Material provide detailed descriptive statistics.

As expected, all GOFs signaled worse model fit with increasing *magnitude* of misspecification in Figures 3 and 4, as evidenced by medians shifting toward unfavorable fit values. That is, all GOFs detected the misspecification of factor dimensionality and the misspecification due to increasingly unmodeled cross-loadings.

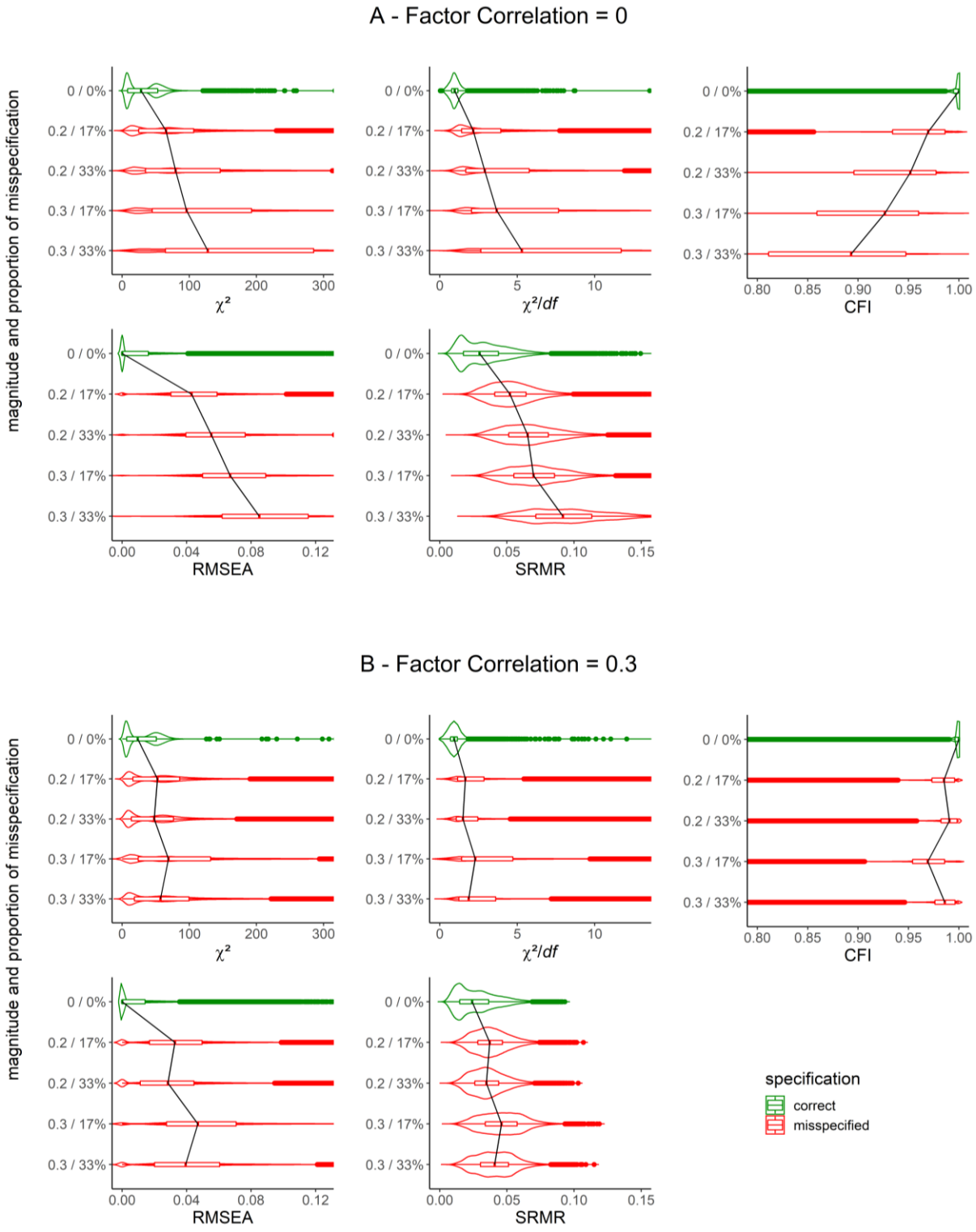
However, we observed distinct influences of the *proportion* of unmodeled cross-loadings on GOFs in uncorrelated and correlated factor scenarios in Figure 4. For uncorrelated factors, an increasing proportion of misspecification also shifted the GOF distribution toward more unfavorable values. For correlated factors, higher proportions of unmodeled cross-loadings resulted in lower medians of each GOF distribution (as the zigzag trace line indicates). Consequently, as the number of indicators with unmodeled cross-loadings increased, GOFs tended to indicate better, not worse, model fit. We attend to this pattern in more detail in the Discussion.

Figure 3: *Distribution of GOFs for Scenarios with Correctly Specified and Misspecified Factor Dimensionality Through the Manipulation of the Factor Correlation*



Note. We displayed each GOF in its original metric and direction. We restricted the X-axis to increase the readability.

Figure 4: *Distribution of GOFs for Scenarios with Correctly Specified and Misspecified Models Regarding Cross-Loadings*



Note. The levels of the Y-axis refer to the magnitude of misspecification and proportion of misspecification, separated by a slash. We displayed each GOF in its original metric and direction. We split the figure by factor correlation and restricted the X-axis to increase the readability.

Sensitivity and Susceptibility of GOFs: Multivariate Analysis with Joint Effects of Characteristics

Next, we quantified how GOFs responded to the different characteristics in correctly specified and misspecified models. We computed Kendall's tau-b as a measure of the bivariate association between each simulation characteristic and GOF to get a first impression of their sensitivity and susceptibility. For space reasons, the bivariate analysis is not included here but in Additional File 2 of the Supplementary Material. Then, for all GOFs, we examined the joint effects of the characteristics combined, including their two-way interaction effects, in a multivariate regression analysis using a least squares estimator (Lumley, 2013; Miller, 1992). We modeled quadratic effects in addition to linear ones for independent variables with more than two levels.

Table 2: *Summary of the Sensitivities and Susceptibilities of GOFs to Model Misspecification and Other Influences*

Independent variables	Dependent variables									
	χ^2		χ^2/df		CFI		RMSEA		SRMR	
	correct (1F/2F)	misspecified (Dim./Load.)	correct (1F/2F)	Misspecified (Dim./Load.)	correct (1F/2F)	misspecified (Dim./Load.)	correct (1F/2F)	misspecified (Dim./Load.)	correct (1F/2F)	misspecified (Dim./Load.)
Main effects										
Misspecification magnitude		−		−		−		−		−
Misspecification proportion [a]		−		−		−		−		−
Estimator (Reference ML)										
MLR		− (Dim.)		− (Dim.)	+ (2F)			− (Dim.)		
DWLS	+	− (Load.)	+	− (Load.)	−	+ / −	+ (1F)	− (Load.)	− (2F)	− (Load.)
WLSMV		− (Load.)		− (Load.)		− (Load.)		− (Load.)		
Number of indicators	−									
Response options										
Asymmetric (Reference symmetric)	−		−				− (1F)			
Loading magnitude		−		−		+ (Load.)		−		− (Dim.)
Sample size									+	
Correlated factors (.30, Reference .00) [a]		+		+		+		+		+
Large two-way interaction effects										
Misspecification magnitude×										
DWLS		− (Load.)		− (Load.)						
WLSMV		− (Load.)								
Correlated factors [a]		+		+		+		+		+
Misspecification proportion [a] ×										
DWLS		−		−						
Correlated factors [a]		+		+		+		+		+
MLR×										
Asymmetric	+		+	− (Dim.)			+ (1F)			
DWLS×										
Number of indicators	+									
Asymmetric	+		+				+ (1F)			
Loading magnitude	+		+							
Correlated factors [a]	+	+	+	+	+	+	+	+	+	
WLSMV×										
Asymmetric	+		+				+ (1F)			
Correlated factors [a]						+				

Independent variables	Dependent variables									
	χ^2		χ^2/df		CFI		RMSEA		SRMR	
	correct (1F/2F)	misspecified (Dim./Load.)	correct (1F/2F)	Misspecified (Dim./Load.)	correct (1F/2F)	misspecified (Dim./Load.)	correct (1F/2F)	misspecified (Dim./Load.)	correct (1F/2F)	misspecified (Dim./Load.)
Loading magnitude× Correlated factors [a]						–				

Note. We recoded GOFs so that lower values represent worse fit (i.e., χ^2, χ^2/df, RMSEA, and SRMR were multiplied by −1). Thus, "+" points to improving fit and "–" to worse fit with the increase of a characteristic. Blank gray cells indicate that the scenario was not/could not be tested in our simulation. Blank white cells indicate that we found no (relatively large) effect. Brackets indicate effects that apply only to certain scenarios (printed in light gray color). If we found different effects per type of correctly specified or misspecified model, we separated the effects with a slash (1F/2F and Dim./Load., respectively). 1F = one-factor CFA. 2F = two-factor CFA. Correct = correctly specified models. Misspecified = misspecified models. Dim. = misspecified factor dimensionality. Load. = unmodeled cross-loadings. MLR = MLR (Yuan & Bentler, 2000). [a]Only for GOFs from two-factor models (2F) and models with unmodeled cross-loadings (Load.). The multiplication sign (×) indicates interaction terms. SRMR is only available for comparing ML and DWLS because SRMR point estimates are identical for models with ML and MLR estimators and models with DWLS and WLSMV estimators (Maydeu-Olivares et al., 2018). We based the summary table on the findings from Table A3 for correctly specified models and Table A4 for misspecified models in Additional File 3 from the Supplementary Material. The table includes main and large two-way interaction effects.

Correctly Specified Models. Several Table 2 columns summarize the findings for correctly specified one- or two-factor models in terms of the direction of effects (not actual results or effect sizes), taken from the detailed regression results in Table A3 in Additional File 3 of the Supplementary Material. We only included relatively large effects from the regression in Table A3 in the Table 2. Those large effects of single independent variables might be small on an absolute scale. Still, they stood out as strong relative to all other effects (and might even aggregate with seemingly small effects of other independent variables). We discuss those effects in the following. We start by describing strong main effects. Then, we move over to findings of interaction effects that only multivariate analysis can uncover. Conclusively, we analyze the variance of GOFs explained by the included data and analysis characteristics (R^2).

Multivariate regression revealed that GOFs were relatively susceptible to various characteristics even in correctly specified models. SRMR depended on the sample size and suggested a better fit with increasing sample size. χ^2 depended on the number of indicators. It suggested better fit with a decreasing number of indicators. χ^2, χ^2/df, and RMSEA (the latter especially in scenarios with one-factor models) suggested better fit for symmetric instead of asymmetric response distributions. The type of estimator impacted all GOFs. However, effects were mixed for different GOFs. Whereas χ^2, χ^2/df, and RMSEA (the latter in scenarios with one-factor models) indicated better fit when using DWLS instead of ML, CFI and SRMR (the latter especially in scenarios with two-factor models) pointed to worse fit with DWLS.

The type of estimator also moderated several effects on GOFs. The number-of-indicator dependency of χ^2 became less strong when switching from ML to DWLS. Likewise, when using MLR, DWLS, or WLSMV instead of ML, the effect of the distribution vanished. Further, regression models revealed interactions between the estimator (DWLS vs. ML) and other characteristics: With DWLS, increasing loading magnitudes suggested better fit according to χ^2 and χ^2/df, but not other GOFs. In the presence of correlated factors, DWLS indicated better model fit than in the presence of uncorrelated factors according to all tested GOFs.

The explained variance (R^2) in the multivariate regression quantifies the joint explanatory power of all simulated characteristics on GOFs, which should ideally be low (as GOFs are otherwise susceptible to those characteristics). We found R^2 to be consistently higher for correctly specified one-factor than two-factor models for all GOFs (see Table A3 in Additional File 3 of the Supplementary Material). The characteristics of our simulation explained the largest share of variance in χ^2 and SRMR of correctly specified one- and two-factor models ($.815 \leq R^2 \leq .894$), meaning that χ^2 and SRMR depended most strongly on the

various simulation characteristics. By comparison, all tested GOFs derived from χ^2 (i.e., χ^2/df, CFI, and RMSEA) were less influenced by data- and analysis-specific characteristics than χ^2 (or SRMR, for that matter), which in turn limited the GOF variability for correctly specified models that those characteristics might have explained ($.061 \leq R^2 \leq .266$). Overall, the effects of characteristics on GOFs in the multivariate analysis were relatively small in absolute terms (though they might aggregate with seemingly small effects of other characteristics, see Table A3 in Additional File 3 of the Supplementary Material).

Misspecified Models. The summary in Table 2 shows columns for models with misspecified factor dimensionality or unmodeled cross-loadings. Table 2 includes main and large two-way interaction effects of model misspecification and other characteristics based on the detailed regression results in Table A4 in Additional File 3 of the Supplementary Material. We marked those effects as relatively large (or relevant) that were equal to or larger than the main effects of the magnitude or, if applicable, the proportion of misspecification. We discuss those effects in the following. We first describe the sensitivity of GOFs to the main effects of the magnitude or proportion of misspecification, followed by describing the interaction effects between the misspecification and other characteristics. Third, we explore the susceptibility of GOFs to data and analysis characteristics. Fourth, we analyze the explained variance (R^2) of GOFs taking all intended influences (i.e., magnitude and proportion of misspecification) and those of other characteristics together.

All GOFs were sensitive to the magnitude of misspecification in all regression models. They indicated worse fit as the magnitude of the misspecification increased (i.e., more misspecification in factor dimensionality, higher unmodeled cross-loadings). Likewise, increasing the proportion of cross-loadings in the population model but leaving them unmodeled in the analysis model suggested decreasing model fit—as expected (holding all else equal).

Crucially, the sensitivity of GOFs to misspecification depended on several other characteristics—a problem that only multivariate analysis can unravel. This *differential* sensitivity of GOFs became evident through substantial two-way interaction effects of the magnitude and proportion of misspecification with the factor correlation (for all GOFs) and the type of estimator (for χ^2 and χ^2/df) in scenarios with unmodeled cross-loadings. We specifically draw the reader's attention to the interaction between the proportion of misspecification and the factor correlation—a trend already evident in the GOF distributions in Figure 4 and confirmed by the multivariate analysis summarized in Table 2. GOFs correctly suggested worse

fit with a higher proportion of unmodeled cross-loadings when factors were uncorrelated. When factors were correlated, GOFs somewhat paradoxically suggested better fit. The factor correlation moderated the effect of the proportion of unmodeled cross-loadings on GOFs.

Further, GOFs were susceptible to many data and analysis characteristics. As the loading magnitude increased, the multivariate regression showed that most GOFs typically indicated worse fit (i.e., χ^2, χ^2/df, RMSEA, and SRMR; the latter especially in scenarios with misspecified factor dimensionality). Thus, low loadings concealed misfit. Only CFI pointed to better model fit with increasing loading magnitudes in scenarios with unmodeled cross-loadings—an effect that seemed to vanish with correlated rather than uncorrelated factors. GOFs also pointed to worse fit in the presence of uncorrelated rather than correlated factors with unmodeled cross-loadings. We also observed a strong influence of the type of estimator on all GOFs (either in scenarios with misspecified factor dimensionality or unmodeled cross-loadings). The multivariate regression also revealed several substantial interactions with the type of estimator. Most GOFs were not simply susceptible to the type of estimator but differentially so, depending on correlating factors (for χ^2, χ^2/df, CFI, and RMSEA in scenarios with unmodeled cross-loadings) or asymmetric response distributions (for χ^2/df in scenarios with misspecified factor dimensionality).

Fourth, the magnitude and proportion of misspecification and all other characteristics together explained up to 96% of the variation in GOFs (usually more than 62% in most scenarios; see Table A4 in Additional File 3 of the Supplementary Material). As an exception to this rule, χ^2 and χ^2/df were not explained (R^2 = .002 at most) in scenarios with misspecified factor dimensionality. Thus, in scenarios with misspecified factor dimensionality, the R^2 pattern spoke favorably of χ^2 and the χ^2/df ratio as being immune to *systematic* influences of *data and analysis characteristics* but also, and problematically so, as being insensitive to *model misspecification* (at least in our extensive simulation that manipulated several other characteristics).

Sensitivity and Susceptibility of GOFs: Selected, Visualized Effects of Characteristics

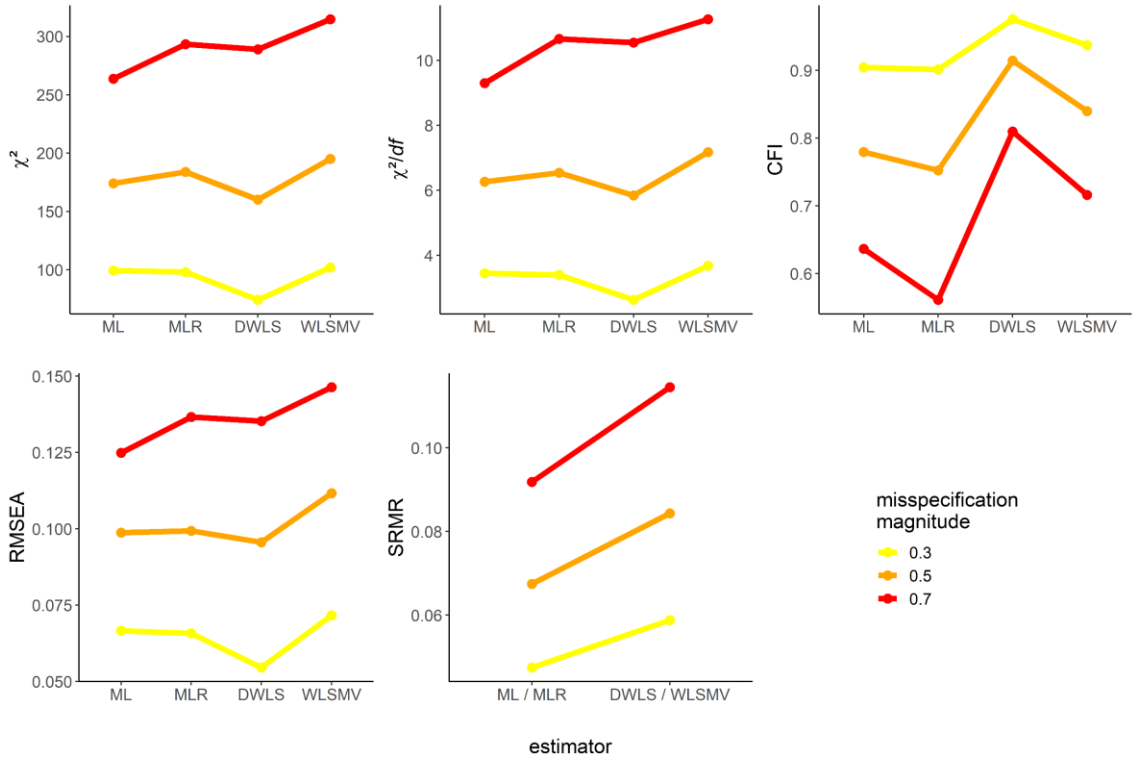
Finally, we visualized selected main and interaction effects on GOFs. Our previous analyses identified substantial influences on GOFs. The sensitivity of GOFs to misspecification (i.e., unmodeled cross-loadings) was mediated by the factor correlation as identified through the descriptive statistics (Figure 4) and multivariate regression (Table 2). The regression further revealed a substantial susceptibility of all GOFs to different types of estimators and loading magnitudes (for misspecified models). Influences on GOFs for correctly specified models had rather small effects in absolute terms. (Though those might aggregate with seemingly small effects of other characteristics, see Table A3 in Additional File 3 of the Supplementary Material.) Thus, we only selected effects on GOFs of misspecified models for an in-depth investigation. The visualization helps to illustrate the complex dependency of GOFs on these characteristics and the way they interact.

Figures 5–8 display those interactions via conditional median plots. The Y-axis shows the respective GOF and its values (original metric without altering the direction); the X-axis conveys the estimators or loading magnitudes. We disentangled the magnitude and, if applicable, proportion of misspecification by using differentially colored and, if applicable, shaped lines that connect medians for each scenario in the plot. We further split the figures by factor correlation for scenarios with unmodeled cross-loadings.

As a general trend, GOFs were sensitive to misspecification. They correctly indicated worse fit with increasing magnitudes of misspecification across all estimators and loading magnitudes (Figures 5–8). As expected, a higher proportion of unmodeled cross-loadings also went along with worse fit when factors were uncorrelated. By contrast, a higher proportion of unmodeled cross-loadings suggested better fit when factors were correlated (Figures 6 and 8; compare this to Figure 4; see also Discussion).

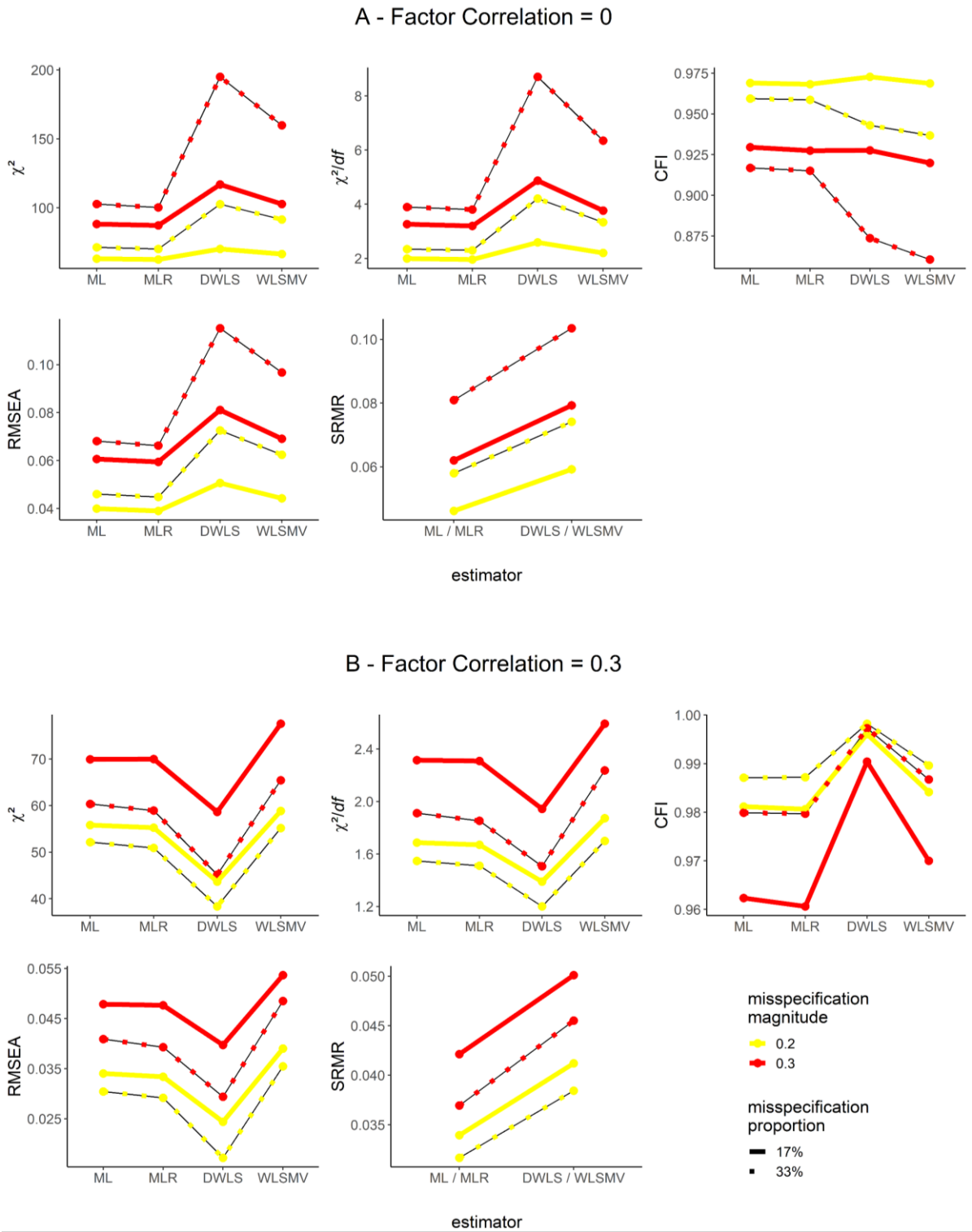
Next, we took a closer look at the susceptibility of GOFs to the type of estimator. A predominant trend was that GOFs were least sensitive to misspecification with DWLS compared to any other estimator (Figures 5 and 6), except for SRMR. However, the factor correlation moderated this trend. It is capable of being completely reversed. In the presence of uncorrelated factors, GOFs (i.e., χ^2, χ^2/df, and RMSEA) suggested worse model fit with DWLS than with other estimators (the only exception being CFI when using WLSMV; see Panel A in Figure 6).

Figure 5: *Median Values of GOFs Conditioned on the Type of Estimator and Misspecification for Scenarios with Misspecified Factor Dimensionality*



Note. We displayed each GOF in its original metric and direction. MLR = MLR (Yuan & Bentler, 2000).

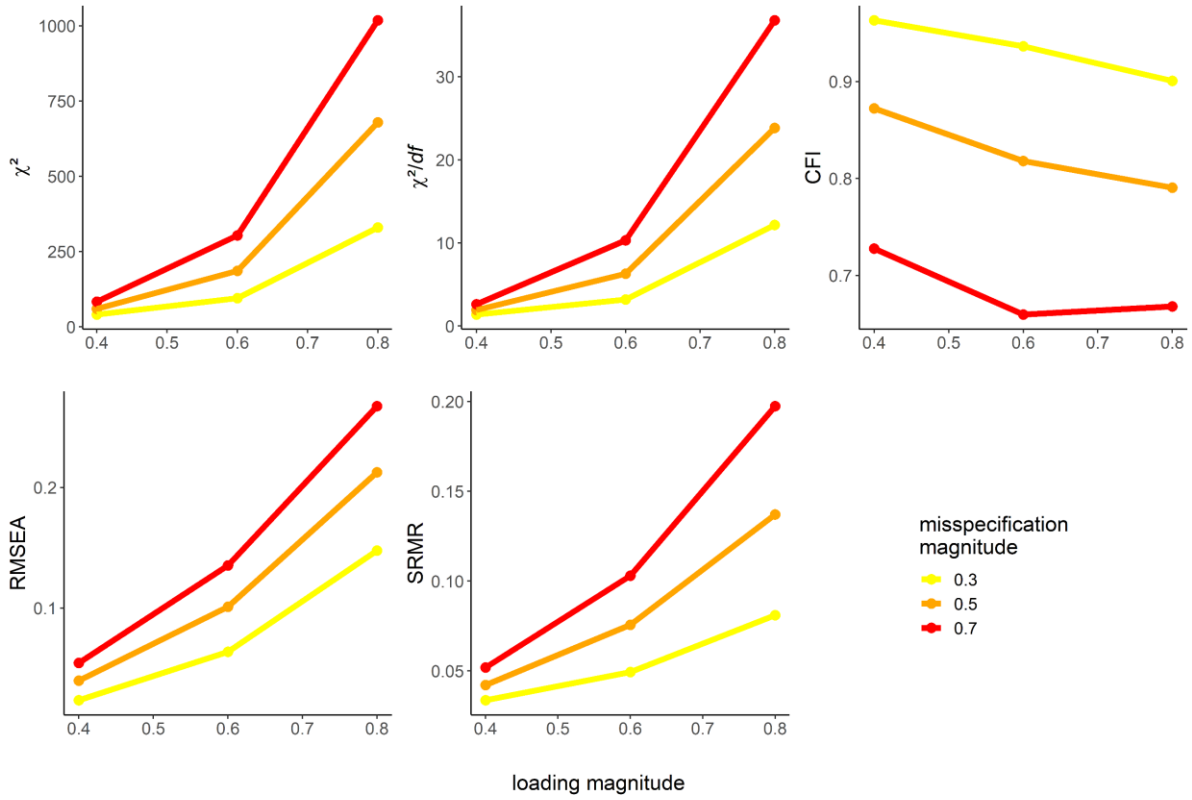
Figure 6: *Median Values of GOFs Conditioned on the Type of Estimator and Misspecification for Scenarios with Unmodeled Cross-Loadings*



Note. We displayed each GOF in its original metric and direction. We split the figure by factor correlation. MLR = MLR (Yuan & Bentler, 2000).

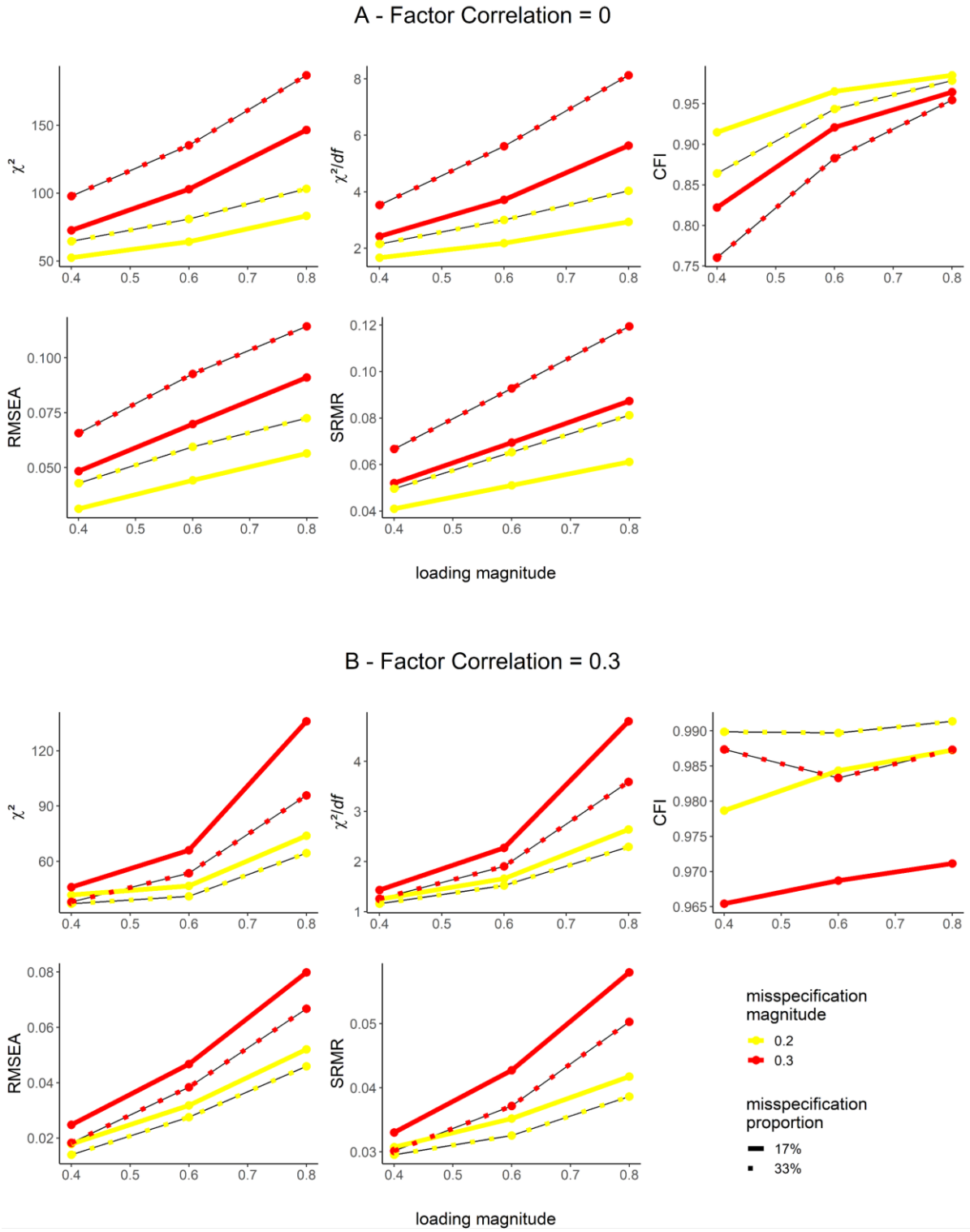
GOFs were not only susceptible to the type of estimator but also to the loading magnitude (i.e., the magnitude of primary loadings in the population model). GOFs became more sensitive to the magnitude of misspecification and, if applicable, proportion of misspecification with higher loading magnitudes (Figures 7 and 8). However, as visualized in Panel B in Figure 8, a higher proportion of misspecification resulted in more optimistic model fit in correlated-factor scenarios, as stated before. CFI was an exception to this pattern. Specifically, CFI showed a reverse or U-shaped relationship between loading magnitude and misspecification due to unmodeled cross-loadings, depending on the scenarios one looks at. Put differently, CFI pointed to better fit with increasing loading magnitudes in most scenarios of unmodeled cross-loadings (Panels A and B in Figure 8). Only at a large proportion of unmodeled cross-loadings (33%) and correlated factors (Panel B in Figure 8), CFI pointed to better fit with low and high loading magnitudes, while medium loadings magnitudes showed worse fit.

Figure 7: *Median Values of GOFs Conditioned on Loading Magnitude and Misspecification for Scenarios with Misspecified Factor Dimensionality*



Note. We displayed each GOF in its original metric and direction.

Figure 8: *Median Values of GOFs Conditioned on Loading Magnitude and Misspecification for Scenarios with Unmodeled Cross-Loadings*



Note. We displayed each GOF in its original metric and direction. We split the figure by factor correlation.

Discussion

GOFs were designed to detect model misspecification and help judge the tenability of latent variable models (e.g., Hu & Bentler, 1999). But how well do GOFs fulfill this purpose? We approached this question by conducting the most thorough simulation study on the *sensitivity* of GOFs to model misspecification in CFA models and their *susceptibility* to other data and analysis characteristics. Crucially, data and analysis characteristics other than misspecification should *not* influence GOFs, lest judgments of model fit may become seriously biased.

Five main insights emerged from our analysis of about 6×10^6 simulated data sets. First and unsurprisingly, GOFs were sensitive to misspecification of both factor dimensionality and cross-loadings: All GOFs correctly indicated worse fit as the degree of misspecified factor dimensionality increased (i.e., the correlation between two factors that were incorrectly modeled as a single factor decreased). GOFs also correctly indicated worse model fit as the magnitude and proportion of unmodeled cross-loadings grew (but only when the factors in the model were uncorrelated).

Second, the sensitivity of GOFs to model misspecification was not the same across all scenarios but varied considerably depending on several other data and analysis characteristics (especially the type of estimator and correlating factors). The most interesting finding was that, when factors were correlated, GOFs suggested better (rather than worse) model fit as the proportion of unmodeled cross-loadings grew. It may surprise applied researchers that the ability of GOFs to detect misspecification depends so strongly on the correlation of factors. In hindsight, this finding is plausible: Fitting a correlated two-factor analysis model that ignores substantial cross-loadings in the population model implies a different meaning and orientation of the two factors in the variable space. The factor with the indicators whose cross-loadings went unmodeled reflects a blend of both factors, such that the factor correlation increases. Concomitantly, the estimated factor loadings of indicators with unmodeled cross-loadings are higher than those of correctly modeled indicators (and—by design—residual variances decreased when cross-loadings were added to the population model). Therefore, a model with correlated factors and substantial cross-loadings that go unmodeled (i.e., are assumed to be zero) accounts for the unmodeled cross-loadings through other model parameters (i.e., the factor correlation and factor loadings). That results in seemingly good model fit despite misspecification. A strong correlation between the estimated factor correlation and the proportion of unmodeled cross-loadings corroborated this interpretation (tau-b = .54). Phrased

differently, the estimated factor correlations were higher than the induced one (i.e., factor correlation of .30 in the population model) and increased when the proportion of unmodeled cross-loadings increased (0%, 17%, 33% unmodeled cross-loadings: median of estimated factor correlations = .30, .46, .54, respectively).

Third, GOFs showed considerable susceptibility to data and analysis characteristics of correctly specified and misspecified models. *All* GOFs analyzed here were susceptible to influences other than model misspecification (especially influences of the type of estimator, loading magnitude, and factor correlation). The susceptibility of GOFs to data and analysis characteristics differed between correctly specified models, misspecified models, and different kinds of misspecified models. We replicated several findings of the susceptibility of GOFs to data and analysis characteristics that had been identified previously. Similar to previous studies, we identified a strong dependency of GOFs on the type of distribution (Reußner, 2019) and the type of estimator (Beauducel & Herzberg, 2006; Nye & Drasgow, 2011) in correctly specified models. Like previous studies, we also identified a strong dependency of GOFs on the magnitude of factor loadings (e.g., Beierl et al., 2018; Hancock & Mueller, 2011; Heene et al., 2011) and the type of factor correlation (only with unmodeled cross-loadings; Beauducel & Wittmann, 2005) in misspecified models.

Fourth, we also shed new light on former findings and unravel hidden complexities of the GOFs' susceptibility to data and analysis characteristics. Most interestingly, former studies (Xia & Yang, 2019) found that DWLS-based GOFs (i.e., χ^2, CFI, and RMSEA) depicted misspecified models more favorably than ML-based GOFs. Our results extended that finding and revealed an interaction with the factor correlation when cross-loadings went unmodeled. DWLS-based GOFs pointed to better fit than ML-based ones with correlated factors; uncorrelated factors reversed the effect.

Fifth, some known influences on GOFs were not as substantial as previously assumed when considering multiple influences in a multivariate analysis. For instance, Xia and Yang (2018) found that asymmetric response distributions lead to more optimistic model fit evaluations for DWLS-/WLSMV-based GOFs (i.e., χ^2, χ^2/df, CFI, and RMSEA) for misspecified models than do symmetric ones. The same applies to ML-based GOFs (i.e., CFI, RMSEA, and SRMR), as Reußner (2019) found. Though we replicated these principal findings, our main effects of asymmetry as well as the interaction effects between DLWS/WLSMV and asymmetry were relatively small compared to other effects in our multivariate analysis. Likewise, we only found a relatively strong sample size dependency for SRMR in correctly

specified models. Different from what was suggested by previous studies (e.g., Kenny et al., 2015; Sharma et al., 2005; Shi et al., 2019), the sample size dependency of other GOFs remained relatively small compared to other influences in the multivariate analysis. To fully compare our main and interaction effects with previous findings, we refer the reader to Table A5 in Additional File 3 of the Supplementary Material. These findings highlight that considering the interdependencies among the different influences on GOFs is essential to fully understand the differential sensitivity and susceptibility to extraneous influences on GOFs.

As outlined throughout the paper, we investigated the sensitivity and susceptibility of GOFs for many combinations of data/analysis characteristics and types of misspecification, considerably extending the scope of previous simulation studies. Still, our enlarged simulation could not cover all (potentially relevant) data and analysis characteristics or types of misspecification. An important limitation to be aware of is our self-imposed restriction to CFA models (see Garrido et al., 2016, for an extensive simulation about fit in exploratory structural equation models). Further, we limited ourselves to two types of misspecification (i.e., misspecification due to factor dimensionality and misspecification due to unmodeled cross-loadings), being fully aware that other types of misspecification regularly occur in empirical settings (such as unmodeled residual covariances; see Podsakoff et al., 2003). Such different types of misspecification are likely to impact GOFs differently (e.g., Savalei, 2012; Shi et al., 2019, 2018; Shi & Maydeu-Olivares, 2020). While covering many scenarios, we certainly did not cover all scenarios regularly found in empirical reality. For instance, models with more than two factors and more than 12 indicators get relevant for several psychological inventories (e.g., the Big Five Inventory by Soto & John, 2017, has 15 factors of facet traits nested in five domain factors and based on 60 indicators in total). Likewise, sample sizes larger than 2,000 regularly occur in large-scale assessments (e.g., Programme for the International Assessment of Adult Competencies, PIAAC, has a per-country sample size of at least 4,500; OECD, 2013).

Implications

We acknowledge that the sheer number of results from our simulation can be daunting. However, together these results convey a clear and straightforward message. The *sensitivity of GOFs to model misspecification* varies greatly across analysis scenarios. *GOFs are susceptible to various data and analysis characteristics*. GOF values reflect characteristics other than the magnitude and proportion of model misspecification. These conclusions align with those of several other studies as our comprehensive simulation study replicated several known influences on GOFs (such as their dependency on the type of estimator, e.g., Beauducel & Herzberg, 2006; Xia & Yang, 2019). In addition, we refined the current knowledge on the sensitivity and susceptibility of GOFs by unraveling several relevant moderator effects through large interactions (especially with the type of estimator and factor correlation) in our simulation study. Our findings underline even more strongly than previous findings that GOFs respond to various data and analysis characteristics in complex and hard-to-predict ways.

Thus, one must not blindly trust the values of GOFs as if they exclusively reflect (mis)fit, let alone rigidly apply fixed cutoffs for model evaluation. We believe this important insight should be internalized by all (applied) researchers and included in statistics and methods curricula dealing with model evaluation. Moreover, we understand that the findings may sound pessimistic and leave some readers wondering how to approach model evaluation in the future. However, all fundamental issues with GOFs that we and others identified (e.g., Marsh et al., 2004; McNeish & Wolf, 2021) have a silver lining. They can encourage researchers to think more deeply about the appropriateness of fixed cutoffs for GOFs and explore alternative procedures that will ultimately lead to more valid judgments about whether a model can be accepted.

Below, we first expand on the problem of fixed cutoffs for GOFs that springs from the differential sensitivity and susceptibility of GOFs to various data and analysis characteristics. Following this, we outline several promising avenues for model evaluation that do not rely on problematic fixed cutoffs.

(Fixed Cutoffs for) GOFs Are More Problematic Than Commonly Assumed

Considering the findings of our simulation, how solid as a basis for evaluating model fit are fixed cutoffs for GOFs? Our results suggest that relying on the same fixed cutoffs to judge model fit in real data applications can be highly problematic and misleading in many settings. Thanks to the breadth of scenarios we studied, we can further illustrate and quantify this problem. To do so, we estimated the frequency distribution of GOFs for correctly specified models separately for each simulation scenario. The 95% quantile (for χ^2, χ^2/df, RMSEA, and SRMR; 5% quantile for CFI) of each frequency distribution corresponds to a 5% probability of concluding that a model is misspecified when it is, in fact, correctly specified (i.e., 5% Type I error rate). We can use those quantiles as relevant cutoffs for GOFs. Additional File 4 of the Supplementary Material (Tables A6–A10) shows the tabulated quantiles.

Researchers often take CFI values above .950 to indicate good model fit (Hu & Bentler, 1999). This heuristic might be sufficiently accurate under some but certainly not under all circumstances. Especially low loading magnitudes undermine the nominal Type I error rate when using a cutoff of CFI > .950. In some scenarios, *much more lenient* values than .950 maintain a 5% error rate. For example, a cutoff as low as CFI = .813 is fully appropriate to demarcate correctly specified and misspecified models for a one-factor model estimated with ML at a sample size of N = 200, with loadings of .40 for six indicators and seven response options, in the presence of asymmetric data. In other scenarios, such as in the presence of high loadings, maintaining a 5% error rate requires *much stricter* values than .950 (e.g., a cutoff of .979 results with loadings of .80 in an otherwise identical setting). To be very clear, accepting (or rejecting) models under various scenarios at a fixed cutoff (.950) does not effectively control the Type I error rate. Fixed cutoffs cannot do justice to every possible setting. Consequently, we strongly discourage researchers from inferring the tenability of a model based on conventional, fixed cutoffs.

These examples highlight two caveats about fixed cutoffs, such as those by Hu and Bentler (1999), that have guided applied researchers' model evaluations for over two decades. Using cutoffs in settings not covered in the initial simulation studies is highly problematic. This pertains, for instance, to testing models with low compared to high factor loadings. For model evaluations through GOFs to be valid, researchers need to consider their specific data and analysis characteristics. In this regard, our findings reinforce previous warnings against overgeneralizing cutoffs, including those by Hu and Bentler (1999) in their original publication

suggesting the canonical cutoffs (see also Marsh et al., 2004; McNeish & Wolf, 2021; Nye & Drasgow, 2011).

Moving from Fixed to Tailored Cutoffs Is the Way Forward

Where does this leave applied researchers seeking to evaluate their model's fit? We recommend that researchers take three steps. First, researchers should consider and test alternative models to learn more about potentially better-suited models. Second, they need to inspect local (mis)fit, for instance, via the residual matrix and modification indices, to investigate whether a model is probably correctly specified or misspecified (see Pornprasertmanit, 2014, for a sophisticated strategy to evaluate local fit). Third, and most promisingly, researchers should inspect global fit not via fixed but via *tailored* (sometimes called "dynamic"; McNeish & Wolf, 2021, 2022) cutoffs for GOFs to evaluate the overall model fit free from bias, including any entailed misfit. Whereas considering alternative models and inspecting local fit are time-honored strategies, tailored cutoffs are a much more recent approach that, we believe, holds great promise and offers a much-needed remedy for the issues with GOFs identified in our present simulation. We believe applied research needs to move toward tailored cutoffs for GOFs that take into account the specific data and analysis characteristics. However, tailored cutoffs are recent and not yet widely used. To foster the much-needed move toward tailored cutoffs, we outline the procedures for evaluating tailored cutoffs in more detail here. We hope to encourage more researchers to consider this emerging strategy. We also provide practical examples and R code illustrating how tailored cutoffs can be implemented.

Tailoring cutoffs for GOFs to the specific data and analysis characteristics can be achieved in different ways. One strategy, which we call the table-based approach, is to consider tables from simulation studies with scenario-specific cutoffs, such as Tables A6 to A10 in Additional File 4 of the Supplementary Material. These tables contain cutoffs for combinations of data and analysis characteristics. They were created to read out the cutoff that can maintain error rates at the desired level in one's specific empirical setting (i.e., accounting for the data and analysis characteristics). This strategy is easy to apply and reminiscent of looking up critical values of, say, *z*-scores or *t*-statistics. One merely selects cutoffs for GOFs from the simulation scenario most closely resembling one's own empirical data and analysis characteristics. For example, for a one-factor model with six indicators, five response options, factor loadings around .60, and a symmetric response distribution estimated with WLSMV in a sample of 200 respondents, one would reject the tested model if the χ^2/df ratio is larger than

1.918, CFI is smaller than .972, RMSEA is larger than .068, or SRMR is larger than .048. However, the table-based approach is somewhat limited: If one's actual data and analysis characteristics are dissimilar to those of simulation scenarios, cutoffs are not given.

Two other strategies to arrive at tailored cutoffs are superior to the simplistic first strategy. Those are regression-/equation-based (e.g., Nye & Drasgow, 2011) and simulation-based approaches (e.g., McNeish & Wolf, 2021, 2022; Millsap, 2007, 2013; Pornprasertmanit, 2014).

In the regression-/equation-based approach,[1] a regression formula predicts the tailored cutoff (Nye & Drasgow, 2011). The formula originates from a single simulation study containing information about how data and analysis characteristics influence a GOF. Users plug characteristics of their own empirical setting into the formula to obtain a cutoff.

To exemplify the regression-/equation-based approach, we derived regression formulae for tailored cutoffs based on the results of our present simulation. The procedure was as follows: We took the cutoffs of Tables A6 to A10 in Additional File 4 of the Supplementary Material as dependent variables and regressed them on all data and analysis characteristics and their quadratic terms and two-way interactions separately for each GOF. The data and analysis characteristics, as well as their quadratic terms and two-way interactions, explained a large share of the variation in cutoffs for GOFs ($R^2 \geq .810$). We saved the regression coefficients in Table 3. The sum of the regression coefficients times the characteristics (i.e., the regression formula) predicts an appropriate cutoff for each GOF. To arrive at appropriate cutoffs for one's own empirical problem, one plugs their empirical data and analysis characteristics into the regression formulae using the coefficients from Table 3. We included a user-friendly R script in Additional File 5 of the Supplementary Material for this purpose. In principle, the regression formulae allow researchers to derive appropriate cutoffs even if their empirical data and analysis characteristics do not perfectly match the ones from the simulation studies.

This approach constitutes a clear advancement over the status quo of rigidly using fixed cutoffs, whatever the preferred heuristic for a GOF is. Further, it is more general than the simplistic table-based approach described first. It is also highly efficient because no new simulation must be carried out (as in the simulation-based approach described next). However,

[1] One can also loosely subsume another approach under the regression-/equation-based category: Researchers can derive tailored cutoffs by relying on statistical assumptions of the χ^2 distribution without and with misspecification (Moshagen & Erdfelder, 2016). Except for the distribution of χ^2, GOF distributions are unknown. As many GOFs (e.g., RMSEA) incorporate the χ^2, one can infer their distribution without and with model misspecification from the χ^2 distribution. A certain quantile of the GOF distribution without misspecification may serve as a cutoff.

the potential downside is that the starting point is still a single simulation study that can never cover all possible real-world settings, no matter how thorough. Although extrapolation is possible in principle, researchers should only use the regression formulae for tailored cutoffs when empirical settings do not strongly deviate from the simulation scenarios.

Table 3: *Regression Coefficients to Derive Tailored Cutoffs*

Independent variables	Dependent variable				
	χ^2	χ^2/df	CFI	RMSEA	SRMR
Intercept	−23.94201	3.28519	−0.53129	0.13285	0.05279
	Main effects				
Estimator (Reference: ML)					
MLR	6.72418	0.45189	−0.21041	0.01536	−
DWLS	5.84976	−0.68404	0.19662	−0.03062	0.03774
WLSMV	−4.68805	−0.27096	0.06079	−0.00865	−
Number of indicators	11.08965	−0.04753	0.04016	−0.00235	0.00278
Response options	−7.16670	−0.35058	0.12387	−0.00896	−0.00963
Response options^2	0.72250	0.03496	−0.00936	0.00098	0.00084
Asymmetric	−0.27294	0.02331	−0.04904	−0.00024	−0.00115
Loading magnitude	−25.73792	−3.58376	4.12967	−0.08865	0.02653
Loading magnitude^2	20.41717	2.96247	−2.75074	0.05766	−0.09506
Sample size	0.00906	0.45022	2.27580	−0.12606	−0.05619
Sample size^2	1.20211	−0.15723	−0.82698	0.04331	0.01882
Number of factors	−12.26618	−0.19792	−0.32211	−0.00594	0.01323
	Two−way interaction effects				
Estimator (Reference: ML)					
MLR×Number of indicators	−0.49485	−0.01090	0.00247	−0.00041	−
MLR×Response options	0.17085	0.00216	0.00384	0.00005	−
MLR×Response options^2	−0.02131	−0.00052	−0.00024	−0.00001	−
MLR×Asymmetric	−2.71568	−0.08311	−0.00135	−0.00225	−
MLR×Loading magnitude	−7.76460	−0.99175	0.45378	−0.03246	−
MLR×Loading magnitude^2	−2.61117	0.42949	−0.31994	0.01556	−
MLR×Sample size	−3.93101	−0.26550	0.09707	−0.00768	−
MLR×Sample size^2	1.45709	0.09907	−0.03794	0.00311	−
MLR×Number of factors	2.71283	0.11781	−0.00868	0.00304	−
DWLS×Number of indicators	−2.43747	0.00544	0.00158	−0.00038	−0.00022
DWLS×Response options	−0.39327	−0.02550	−0.00440	−0.00033	−0.00758
DWLS×Response options^2	0.02110	0.00195	0.00034	0.00003	0.00058
DWLS×Asymmetric	−3.01669	−0.09613	0.00452	−0.00244	0.00140
DWLS×Loading magnitude	−41.99689	−1.36226	−0.30944	−0.05998	−0.00280
DWLS×Loading magnitude^2	16.73726	0.48895	0.18350	0.02118	−0.00311
DWLS×Sample size	−2.10846	−0.11430	−0.11629	0.02896	−0.02165
DWLS×Sample size^2	0.75628	0.04419	0.04250	−0.00982	0.00756
DWLS×Number of factors	16.86537	0.64662	−0.01075	0.02281	0.00330

Independent variables	Dependent variable				
	χ^2	χ^2/df	CFI	RMSEA	SRMR
WLSMV×Number of indicators	−0.60239	−0.00413	0.00097	−0.00029	−
WLSMV×Response options	0.63654	0.01440	−0.00270	0.00054	−
WLSMV×Response options^2	−0.05539	−0.00115	0.00022	−0.00004	−
WLSMV×Asymmetric	−2.91980	−0.09484	0.00368	−0.00256	−
WLSMV×Loading magnitude	10.29574	0.24493	−0.06415	0.00577	−
WLSMV×Loading magnitude^2	−15.70961	−0.45035	0.03405	−0.01194	−
WLSMV×Sample size	3.01133	0.05021	−0.04314	0.00682	−
WLSMV×Sample size^2	−1.16888	−0.01938	0.01545	−0.00250	−
WLSMV×Number of factors	3.90897	0.15706	−0.00697	0.00452	−
Number of indicators×					
Response options	−0.25997	−0.00776	0.00002	−0.00033	−0.00019
Response options^2	0.02789	0.00081	−0.00003	0.00004	0.00002
Asymmetric	0.17890	0.00040	0.00034	0.00002	0.00003
Loading magnitude	−4.80064	−0.04388	−0.10488	−0.00698	−0.00316
Loading magnitude^2	3.83154	0.04484	0.07017	0.00652	0.00190
Sample size	−1.04664	0.00016	−0.01404	0.00655	−0.00160
Sample size^2	0.38895	0.00157	0.00500	−0.00224	0.00058
Number of factors	0.64889	−0.01164	0.00234	−0.00030	0.00003
Response options×					
Asymmetric	0.47743	0.01356	−0.00319	0.00043	0.00061
Loading magnitude	22.43204	1.53987	−0.33800	0.04504	0.01794
Loading magnitude^2	−19.13312	−1.33866	0.23297	−0.03818	−0.01639
Sample size	2.42094	−0.20111	−0.00951	−0.00561	0.01125
Sample size^2	−1.16974	0.07428	0.00109	0.00208	−0.00388
Number of factors	1.18404	0.04887	0.00555	0.00098	0.00068
Response options^2×					
Asymmetric	−0.04690	−0.00122	0.00031	−0.00005	−0.00006
Loading magnitude	−2.12451	−0.14685	0.02682	−0.00451	−0.00166
Loading magnitude^2	1.79085	0.12730	−0.01863	0.00381	0.00151
Sample size	−0.40475	0.01424	0.00006	0.00030	−0.00097
Sample size^2	0.17973	−0.00524	0.00020	−0.00011	0.00034
Number of factors	−0.10687	−0.00463	−0.00039	−0.00010	−0.00005
Asymmetric×					
Loading magnitude	3.71952	0.27558	0.11280	0.01327	0.00844
Loading magnitude^2	−1.17710	−0.18396	−0.07174	−0.00937	−0.00713
Sample size	0.29781	−0.00253	0.02437	−0.00219	−0.00393
Sample size^2	−0.16484	−0.00346	−0.00870	0.00072	0.00138
Number of factors	−0.81043	−0.03425	−0.00208	−0.00101	−0.00028
Loading magnitude×					
Sample size	16.43214	0.03858	−5.87140	0.01187	−0.04098
Sample size^2	−8.22119	−0.08583	2.14448	−0.00411	0.01586
Number of factors	1.82103	0.21559	0.65988	0.02703	−0.01793
Loading magnitude^2×					
Sample size	−15.03742	−0.14310	3.87122	0.00022	0.06458
Sample size^2	7.54665	0.12608	−1.41187	0.00015	−0.02336
Number of factors	5.02726	0.03351	−0.43878	−0.01413	0.04983
Sample size×					
Number of factors	0.39375	0.08076	0.09529	−0.00988	−0.02326
Sample size^2×					
Number of factors	−0.26943	−0.03848	−0.03378	0.00320	0.00784

Independent variables	Dependent variable				
	χ^2	χ^2/df	CFI	RMSEA	SRMR
Number of factors× Correlated factors	−2.51728	−0.05765	0.00487	−0.00223	−0.00481
R^2	.970	.810	.902	.903	.963
N	1,296	1,296	1,296	1,296	648

Note. MLR = MLR (Yuan & Bentler, 2000). The multiplication sign (×) indicates interaction terms. To correctly interpret the sample-size regression coefficient, divide the sample size by 1,000 before plugging it into the equation. Regression coefficients are unstandardized and uncentered. Independent variables with more than two simulated levels were included additionally in quadratic form. SRMR is only available for comparing ML and DWLS because SRMR point estimates are identical for models with ML and MLR estimators and models with DWLS and WLSMV estimators (Maydeu-Olivares et al., 2018). We omitted standard errors and *p*-values for clarity. Regression coefficients to derive tailored cutoffs for χ^2 were colored gray. They heavily depend on the degrees of freedom and are, thus, barely useful for models different from the ones in the paper. The sum of the regression coefficients times the characteristics (i.e., the regression formula) predicts an appropriate cutoff for each GOF.

If empirical settings strongly deviate from simulation scenarios, cutoffs should be used neither from cutoff tables nor regression formulae. Instead, one may adopt another approach and conduct a small-scale, scenario-specific simulation to investigate the behavior of GOFs. Several authors suggested this approach (most recently, McNeish & Wolf, 2021, 2022; for similar earlier work, see Millsap, 2007, 2013; Niemand & Mai, 2018; Pornprasertmanit, 2014; for nested models, see Pornprasertmanit et al., 2013). Before initializing the simulation, researchers define analysis and population models. Then, they simulate data from the population model (via a Monte Carlo simulation, similar to what we did in the present paper), fit the analysis model to the data, and record the GOFs. Similar to our tables in Additional File 4 of the Supplementary Material, researchers then extract cutoffs from the resulting GOF distributions. The analysis model can equal (or approximately equal; see Millsap, 2007, 2013; Pornprasertmanit, 2014) the population model, which corresponds to a correctly specified model. Cutoffs derived from the GOF distribution of correctly specified models control the Type I error rate (as implemented in the approaches of McNeish & Wolf, 2021, 2022; Millsap, 2007, 2013; Niemand & Mai, 2018; Pornprasertmanit, 2014). Including a misspecified model (i.e., where the analysis model differs considerably from the population model) allows controlling the Type II error rate (i.e., the probability of concluding that a model is correctly specified when it is, in fact, misspecified) in the derivation of tailored cutoffs (as implemented in the approaches of McNeish & Wolf, 2021, 2022; Pornprasertmanit, 2014). Further, including several misspecified models might help to evaluate model fit gradually (e.g., McNeish & Wolf, 2021, 2022).

Choosing simulation characteristics (e.g., analysis model, sample size, estimator) similar to those of the empirical setting of interest is the gold standard to arrive at tailored cutoffs. By simulating data, cutoffs can be tailored to the setting of interest. However, the flexibility of the simulation-based approach may not always be a merit but also a difficulty. The simulation-based approach demands specific knowledge about defining population and analysis models and running and analyzing simulations. Automated solutions (i.e., shiny apps) can considerably ease the process (e.g., McNeish & Wolf, 2021).

In sum, the table-, regression-/equation-, and simulation-based approaches are three alternative ways to arrive at tailored cutoffs for model evaluation. Although these procedures are more involved than judging model fit against fixed cutoffs for GOFs, we hope our simulation results have convinced the reader of the urgency of phasing out fixed cutoffs in favor of a more valid tailored approach.

Conclusion

GOFs were designed to detect model misspecification and support the evaluation of model fit. However, our simulation study highlights two fundamental problems with GOFs. First, GOFs reflect not only model misspecification; they are also susceptible to a range of data and analysis characteristics (other than model misspecification). Second, the sensitivity of GOFs to model misspecification also depends on such characteristics. In this regard, a particularly impressive (and alarming) finding was the strong dependence on absolute GOF values and their misspecification sensitivity to the factor correlation, the magnitude of factor loadings, and the type of estimator. Such characteristics are irrelevant from the applied researcher's point of view for judging model fit or identifying misspecification. Hence, they should ideally have no bearing at all on GOFs. However, our findings converge with—and even expand—previous smaller-scale simulations suggesting that a range of characteristics other than misspecification influence absolute GOF values.

The pattern of associations between those characteristics and GOFs is complex, as interaction effects attest; it varies for different GOFs and is hard to predict for specific constellations. This complexity means simple modifications cannot come to the rescue, such as adding or subtracting a constant from cutoff values. The problem lies with fixed cutoffs for GOFs *as such.* Fixed cutoffs cannot do justice to all combinations of data and analysis characteristics researchers encounter in applied settings.

Our findings make it abundantly clear that the conventional practice of relying on fixed cutoffs for GOFs is far more problematic than commonly assumed. Even though previous simulations had raised some of the issues highlighted in our study, the practice has not changed. Hu and Bentler (1999) already cautioned researchers to execute discretion when using their cutoffs (see also McNeish & Wolf, 2021). However, applied researchers continue to rely on these cutoffs even in settings markedly different from the scenarios covered by Hu and Bentler (1999) and related studies by Reußner (2019) and Rutkowski and Svetina (2014). More than 20 years later, our detailed simulation resonates with their initial warnings and brings several additional issues to light. Consequently, we urge researchers to be wary of the problems with fixed cutoffs.

We recommend researchers routinely adopt the time-honored strategies of inspecting (and reporting) local fit and comparing alternative models instead of relying exclusively on GOFs. Methodologists have long advocated these effective strategies, but these are far from being universally applied in published research. Overall, however, we believe the field needs to move away from relying on fixed cutoffs and toward cutoffs tailored to the specific data and analysis characteristics (e.g., McNeish & Wolf, 2021, 2022). Tailored cutoffs offer an appropriate response to the susceptibility of GOFs and the ensuing lack of validity of fixed cutoffs. To contribute to a much-needed shift toward tailored cutoffs, we discussed and developed emerging strategies for implementing tailored cutoffs and pointed to ongoing work that aims to improve these strategies further. We hope our simulation results will encourage researchers to embark on this path, ultimately resulting in valid and replicable research.

References

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13*(2), 186–203. https://doi.org/10.1207/s15328007sem1302_2

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling, 12*(1), 41–75. https://doi.org/10.1207/s15328007sem1201_3

Beierl, E. T., Bühner, M., & Heene, M. (2018). Is that measure really one-dimensional? *Methodology, 14*(4), 188–196. https://doi.org/10.1027/1614-2241/a000158

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M. (1995). *EQS structural equations program manual*. Multivariate Software.

Bilsky, W., Janik, M., & Schwartz, S. H. (2011). The structural organization of human values-evidence from three rounds of the European Social Survey (ESS). *Journal of Cross-Cultural Psychology, 42*(5), 759–776. https://doi.org/10.1177/0022022110362757

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology, 9*(2), 78–84. https://doi.org/10.1027/1614-2241/a000057

Bollen, K. A. (1989). *Structural equations with latent variables.* Wiley.

Bonett, D. G. (2012). Replication-extension studies. *Current Directions in Psychological Science, 21*(6), 409–412. https://doi.org/10.1177/0963721412459512

Boomsma, A. (2013). Reporting Monte Carlo studies in structural equation modeling. *Structural Equation Modeling, 20*(3), 518–540. https://doi.org/10.1080/10705511.2013.797839

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

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. https://doi.org/10.1177/0049124108314720

Clark, L. A., & Watson, D. (2019). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

Comrey, A. L., & Lee, H. B. (1992). *Interpretation and Application of Factor Analytic Results* (2nd ed.). Lawrence Erlbaum Associates.

DiStefano, C., McDaniel, H. L., Zhang, L., Shi, D., & Jiang, Z. (2019). Fitting large factor analysis models with ordinal data. *Educational and Psychological Measurement, 79*(3), 417–436. https://doi.org/10.1177/0013164418818242

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

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychological Methods, 21*(1), 93–111. https://doi.org/10.1037/met0000064

Groskurth, K., Nießen, D., Rammstedt, B., & Lechner, C. M. (2021). An English-language adaptation and validation of the Political Efficacy Short Scale (PESS). *Measurement Instruments for the Social Sciences, 3*, Article 1. https://doi.org/10.1186/s42409-020-00018-z

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. https://doi.org/10.1177/0013164410384856

Hayduk, L. (2014). Seeing perfectly fitting factor models that are causally misspecified: Understanding that close-fitting models can be worse. *Educational and Psychological Measurement, 74*(6), 905–926. https://doi.org/10.1177/0013164414527449

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. https://doi.org/10.1037/a0024917

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

Jackson, D. L., Gillaspy Jr, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: an overview and some recommendations. *Psychological Methods, 14*(1), 6–23. https://doi.org/10.1037/a0014694

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Scientific Software International.

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. https://doi.org/10.1207/S15328007SEM1003_1

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. https://doi.org/10.1177/0049124114543236

Kim, J. H., Lee, J., Richardson, T. V., Lee, D. H., McMahon, B. T., Kim, H., & Sametz, R. R. (2021). Psychometric validation of Adapted Inventory of Virtues and Strengths. *Rehabilitation Counseling Bulletin*. Advance online publication. https://doi.org/10.1177/0034355221993553

Lee, J., & Cagle, J. G. (2017). Validating the 11-item revised University of California Los Angeles scale to assess loneliness among older adults: An evaluation of factor structure and other measurement properties. *The American Journal of Geriatric Psychiatry, 25*(11), 1173–1183. https://doi.org/10.1016/j.jagp.2017.06.004

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

Lumley, T. (2013). Biglm: Bounded memory linear and generalized linear models. R package version 0.9-1. https://CRAN.R-project.org/package=biglm

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. https://doi.org/10.1016/j.paid.2006.09.023

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. https://doi.org/10.1207/s15328007sem1103_2

Maydeu-Olivares, A., Shi, D., & Rosseel, Y. (2018). Assessing fit in structural equation models: A Monte-Carlo evaluation of RMSEA versus SRMR confidence intervals and tests of close fit. *Structural Equation Modeling, 25*(3), 389–402. https://doi.org/10.1080/10705511.2017.1389611

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000425

McNeish, D., & Wolf, M. G. (2022). Dynamic fit index cutoffs for one-factor models. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-022-01847-y

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. https://doi.org/10.1080/00223891.2017.1281286

Miller, A. J. (1992). Algorithm AS 274: Least squares routines to supplement those of Gentleman. *Journal of the Royal Statistical Society, Series C (Applied Statistics), 41*(2), 458–478. https://doi.org/10.2307/2347583.

Millsap, R. E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences, 42*(5), 875–881. https://doi.org/10.1016/j.paid.2006.09.021

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

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

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. https://doi.org/10.1080/10705511.2012.634724

Moshagen, M., & Auerswald, M. (2018). On congruence and incongruence of measures of fit in structural equation modeling. *Psychological Methods, 23*(2), 318–336. https://doi.org/10.1037/met0000122

Moshagen, M., & Erdfelder, E. (2016). A new strategy for testing structural equation models. *Structural Equation Modeling, 2*3(1), 54–60. https://doi.org/10.1080/10705511.2014.950896

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*(1), 115–132. https://doi.org/10.1007/BF02294210

Muthén, B., Du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.* http://www.statmodel.com/bmuthen/articles/Article_075.pdf

Muthén, L.K., & Muthén, B.O. (1998-2017). *Mplus user's guide* (8th ed). Muthén & Muthén.

Niemand, T., & Mai, R. (2018). Flexible cutoff values for fit indices in the evaluation of structural equation models. *Journal of the Academy of Marketing Science, 46*, 1148–1172. https://doi.org/10.1007/s11747-018-0602-9

Nießen, D., Partsch, M. V., Kemper, C. J., & Rammstedt, B. (2019). An English-Language Adaptation of the Social Desirability–Gamma Short Scale (KSE-G). *Measurement Instruments for the Social Sciences, 2*(1), Article 2. https://doi.org/10.1186/s42409-018-0005-1

Nye, C. D., & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods, 14*(3), 548–570. https://doi.org/10.1177/1094428110368562

OECD (2013). *OECD skills outlook 2013: First results from the survey of adult skills*. Paris: OECD Publishing. https://doi.org/10.1787/9789264204256-en

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Pornprasertmanit, S. (2014). *The unified approach for model evaluation in structural equation modeling* [Unpublished doctoral dissertation]. University of Kansas. http://hdl.handle.net/1808/16828

Pornprasertmanit, S., Wu, W., & Little, T. D. (2013). 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). Springer. https://doi.org/10.1007/978-1-4614-9348-8_12

R Core Team (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/index.html

Rammstedt, B., & Beierlein, C. (2014). Can't we make it any shorter? The limits of personality assessment and way to overcome them. *Journal of Individual Differences, 35*(4), 212–220. https://doi.org/10.1027/1614-0001/a000141

Reußner, M. (2019). *Die Güte der Gütemaße: Zur Bewertung von Strukturgleichungsmodellen* [The fit of fit indices: The evaluation of model fit for structural equation models]. Walter de Gruyter.

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

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

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31–57. https://doi.org/10.1177/0013164413498257

Savalei, V. (2012). The relationship between root mean square error of approximation and model misspecification in confirmatory factor analysis models. *Educational and Psychological Measurement, 72*(6), 910–932. https://doi.org/10.1177/0013164412452564

Savalei, V. (2020). Improving fit indices in structural equation modeling with categorical data. *Multivariate Behavioral Research, 56*(3), 390–407. https://doi.org/10.1080/00273171.2020.1717922

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23–74.

Sharma, S., Mukherjee, S., Kumar, A., & Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research, 58*(7), 935–943. https://doi.org/10.1016/j.jbusres.2003.10.007

Shi, D., & Maydeu-Olivares, A. (2020). The effect of estimation methods on SEM fit indices. *Educational and Psychological Measurement, 80*(3), 421–445. https://doi.org/10.1177/0013164419885164

Shi, D., DiStefano, C., McDaniel, H. L., & Jiang, Z. (2018). Examining chi-square test statistics under conditions of large model size and ordinal data. *Structural Equation Modeling, 25*(6), 924–945. https://doi.org/10.1080/10705511.2018.1449653

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. https://doi.org/10.1177/0013164418783530

Shi, D., Maydeu-Olivares, A., & DiStefano, C. (2018). The relationship between the standardized root mean square residual and model misspecification in factor analysis models. *Multivariate Behavioral Research, 53*(5), 676–694. https://doi.org/10.1080/00273171.2018.1476221

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment, 31*(4), 557–566. https://doi.org/10.1037/pas0000648

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*(2), 173–180. https://doi.org/10.1207/s15327906mbr2502_4

Ullman, J. B. (2014). Structural equation modeling. In B. G. Tabachnick, & L. S. Fidell, (Eds.). *Using multivariate statistics* (6th ed.). Pearson Education.

Ushey, K. (2020). *renv: Project environments.* R package version 0.12.2. https://cran.r-project.org/web/packages/renv/index.html

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York.

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods, 8*(1), 16–37. https://doi.org/10.1037/1082-989X.8.1.16

Xia, Y., & Yang, Y. (2018). The influence of number of categories and threshold values on fit indices in structural equation modeling with ordered categorical data. *Multivariate Behavioral Research, 53*(5), 731–755. https://doi.org/10.1080/00273171.2018.1480346

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods, 51*, 409–428. https://doi.org/10.3758/s13428-018-1055-2

Yuan, K.-H., & Bentler, P. M. (2000). 5. Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 30*(1), 165–200. https://doi.org/10.1111/0081-1750.00078

Additional File 1: Descriptive Statistics of GOFs

Table A1: *Descriptive Statistics of GOFs for Correctly Specified or Misspecified Models Regarding Factor Dimensionality*

Specification	*Mdn*	*M*	*SD*	*Min*	*Max*
			χ^2		
correct: 0	19.918	29.555	23.900	0.224	127.951
misspecified: .30	92.502	263.896	477.293	0.728	3,509.264
misspecified: .50	176.865	584.419	5,521.009	0.957	1,864,111.000
misspecified: .70	289.410	885.368	45,863.610	1.121	29,896,940.000
			χ^2/df		
correct: 0	0.918	0.936	0.405	0.025	13.982
misspecified: .30	3.279	9.953	16.231	0.081	171.069
misspecified: .50	6.472	25.068	562.245	0.106	207,123.400
misspecified: .70	10.472	40.016	5,093.109	0.125	3,321,882.000
			CFI		
correct: 0	1	.993	0.024	0	1
misspecified: .30	.938	.926	0.063	0	1
misspecified: .50	.834	.814	0.125	0	1
misspecified: .70	.692	.683	0.142	0	1
			RMSEA		
correct: 0	0	0.009	0.015	0	0.255
misspecified: .30	0.065	0.080	0.060	0	0.622
misspecified: .50	0.101	0.123	0.103	0	18.855
misspecified: .70	0.137	0.155	0.119	0	40.755
			SRMR		
correct: 0	0.020	0.023	0.014	0.002	0.101
misspecified: .30	0.053	0.056	0.023	0.007	0.161
misspecified: .50	0.077	0.086	0.042	0.012	0.254
misspecified: .70	0.104	0.118	0.063	0.012	0.351

Table A2: *Descriptive Statistics of GOF for Correctly Specified or Misspecified Models Regarding Cross-Loadings*

	Mdn		*M*		*SD*		*Min*		*Max*	
	Factor Correlation									
Specification	.00	.03	.00	.03	.00	.03	.00	.03	.00	.03
					χ^2					
correct: 0, 0%	28.129	23.052	31.716	29.467	24.395	23.759	0.277	0.195	321.851	978.836
mis.: .20, 17%	65.239	52.658	87.120	67.502	92.751	73.443	0.458	0.416	1,361.707	1,008.538
mis.: .20, 33%	79.558	47.684	126.889	58.648	164.171	61.815	0.588	0.212	2,287.394	601.259
mis.: .30, 17%	96.065	68.946	162.217	116.498	202.464	154.999	0.962	0.298	2,504.531	3,545.111
mis.: .30, 33%	127.529	57.021	251.330	86.083	370.583	111.319	2.211	0.289	4,612.157	1,120.618
					χ^2/df					
correct: 0, 0%	0.958	0.923	1.001	0.949	0.407	0.508	0.031	0.024	13.538	122.354
mis.: .20, 17%	2.155	1.654	3.272	2.593	3.009	2.667	0.051	0.052	38.937	126.067
mis.: .20, 33%	2.899	1.490	4.879	2.218	5.276	2.131	0.065	0.027	66.259	66.132
mis.: .30, 17%	3.678	2.287	6.324	4.454	6.844	5.860	0.107	0.037	77.623	443.139
mis.: .30, 33%	5.300	1.868	9.930	3.396	12.073	4.149	0.246	0.036	131.666	41.791
					CFI					
correct: 0, 0%	1	1	.988	.990	0.039	0.035	0	0	1	1
mis.: .20, 17%	.969	.985	.949	.977	0.063	0.038	0	0	1	1
mis.: .20, 33%	.952	.991	.924	.984	0.078	0.029	0	0	1	1
mis.: .30, 17%	.927	.969	.898	.964	0.088	0.039	0	0	1	1
mis.: .30, 33%	.893	.986	.865	.982	0.103	0.024	0	0	1	1
					RMSEA					
correct: 0, 0%	0	0	0.010	0.009	0.016	0.015	0	0	0.251	0.779
mis.: .20, 17%	0.043	0.033	0.045	0.035	0.025	0.026	0	0	0.272	0.791
mis.: .20, 33%	0.055	0.028	0.059	0.030	0.030	0.024	0	0	0.318	0.571
mis.: .30, 17%	0.067	0.047	0.071	0.051	0.031	0.034	0	0	0.325	1.487
mis.: .30, 33%	0.085	0.039	0.092	0.042	0.040	0.030	0	0	0.397	0.418
					SRMR					
correct: 0, 0%	0.030	0.024	0.032	0.027	0.017	0.015	0.003	0.002	0.163	0.093
mis.: .20, 17%	0.052	0.037	0.054	0.038	0.018	0.013	0.006	0.004	0.185	0.107
mis.: .20, 33%	0.065	0.035	0.068	0.036	0.022	0.013	0.010	0.003	0.206	0.103
mis.: .30, 17%	0.070	0.046	0.071	0.046	0.022	0.016	0.014	0.005	0.199	0.119
mis.: .30, 33%	0.092	0.041	0.094	0.041	0.029	0.014	0.020	0.004	0.246	0.115

Note. The magnitude of unmodeled cross-loadings, followed by the proportion of indicators with unmodeled cross-loadings in percent. Mis = misspecified.

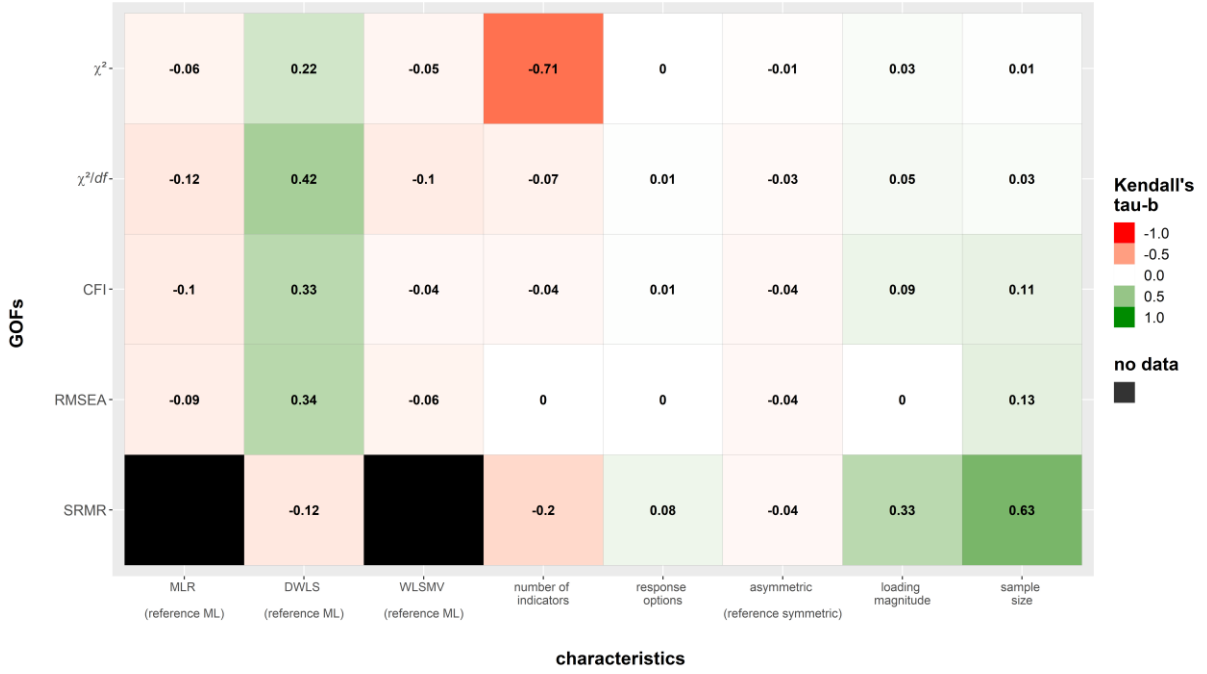
Additional File 2: Bivariate Associations of GOFs with Characteristics

Here, we quantified how GOFs responded to the different characteristics in correctly specified and misspecified models. We computed Kendall's tau-b as a measure of association between each simulation characteristic and GOF. Bivariate correlations collapse across non-focal scenarios (i.e., those characteristics that are not part of the specific bivariate correlation but vary in the data). Thus, bivariate findings may be misleading as they can mask other, more complex interdependencies between simulation characteristics that only multivariate analysis can reveal. Nonetheless, bivariate correlations help gain a first impression of how simulation characteristics and GOFs were associated.

To facilitate the readability of the following correlation tables, presented in a colored heatmap style, we recoded GOFs such that lower values consistently represent worse model fit (i.e., χ^2, χ^2/df, RMSEA, and SRMR were multiplied by −1). For increasing values of the simulated characteristics, positive correlations (colored in green) point to GOF values that indicate improving fit and negative correlations (colored in red) point to diminishing fit: the stronger the absolute correlation coefficient, the more intense the color. Correlations of GOFs with the magnitude and proportion of misspecification help explore their sensitivity to misspecification. Correlations of GOFs with other characteristics alert to their unintended susceptibility in correctly specified or misspecified models.

Correctly Specified Models. Figure A1 shows the correlations for correctly specified one-factor models. GOFs and characteristics correlated moderately to strongly in many instances (.20 ≤ |tau-b| ≤ .71). With the DWLS estimator instead of the ML estimator, GOFs typically pointed to better model fit (.22 ≤ tau-b ≤ .42; except for SRMR with tau-b = −.12). As the number of indicators increased, χ^2 indicated less acceptable models (tau-b = −.71). Akin to χ^2, SRMR was also susceptible to the number of indicators, suggesting worse model fit as models used more and more indicators (tau-b = −.20). SRMR was strongly influenced by sample size. Its values indicated better model fit when sample size increased (tau-b = .63). SRMR also pointed to better model fit with increasing loading magnitude (tau-b = .33). Overall, Figure A1 demonstrates the extent of GOF susceptibility to data and analysis characteristics for correctly specified one-factor models.

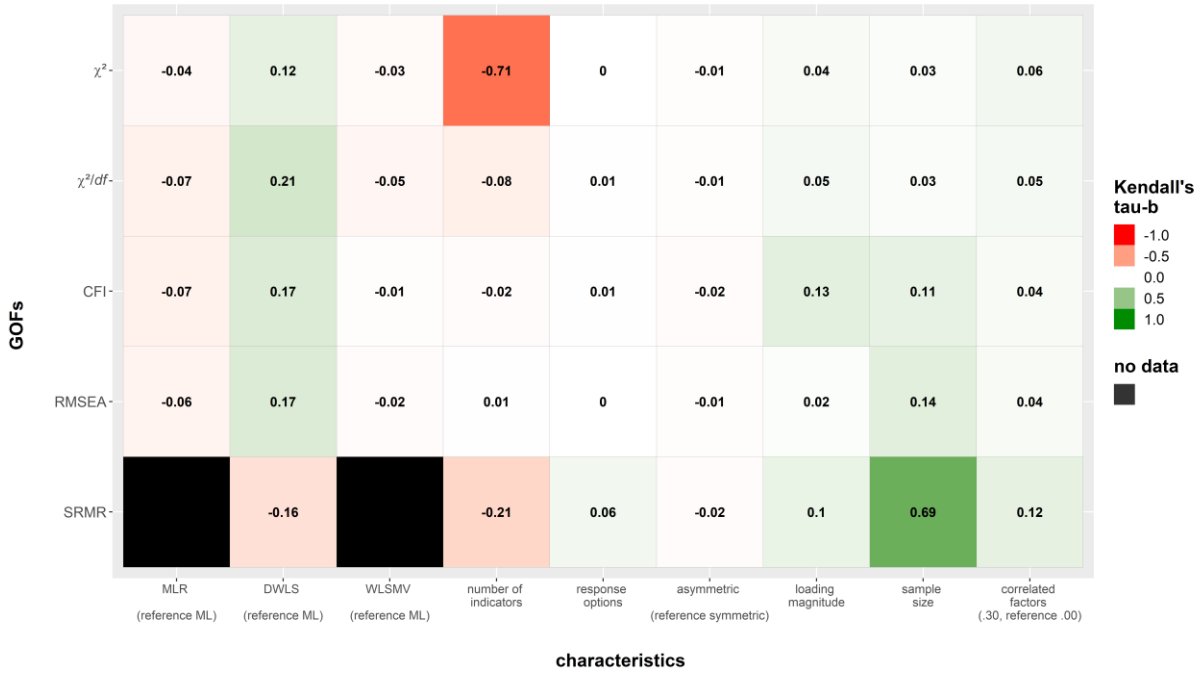
Figure A1: *Zero-Order Correlation Between Characteristics and GOFs of Correctly Specified One-Factor Models*



Note. We recoded GOFs so that lower values represent worse fit (i.e., χ^2, χ^2/df, RMSEA, and SRMR were multiplied by −1). MLR = MLR (Yuan & Bentler, 2000). We only displayed SRMR for comparing ML and DWLS because SRMR point estimates are identical for models estimated with ML or MLR and models estimated with DWLS or WLSMV (Maydeu-Olivares et al., 2018).

Figure A2 shows the findings for two-factor models. Akin to GOFs of one-factor models, though to a lesser extent (.12 ≤ tau-b ≤ .21), GOFs of two-factor models were related to the type of estimator, indicating better model fit when estimated with DWLS compared to ML, except for SRMR (tau-b = −.16). As for GOFs of one-factor models, there were substantial correlations between χ^2 and the number of indicators (tau-b = −.71), as well between SRMR and sample size (tau-b = .69) or the number of indicators (tau-b = −.21) for two-factor models. The correlation coefficients resulting from two-factor models were very similar to those from one-factor models (and the mathematical signs of the strongest coefficients were identical), suggesting that the number of factors did not substantially impact GOFs from correctly specified models. We found the biggest difference between SRMR and loading magnitude for one- and two-factor models (.33 vs. .10, respectively).

Figure A2: *Zero-Order Correlation Between Characteristics and GOFs of Correctly Specified Two-Factor Models*

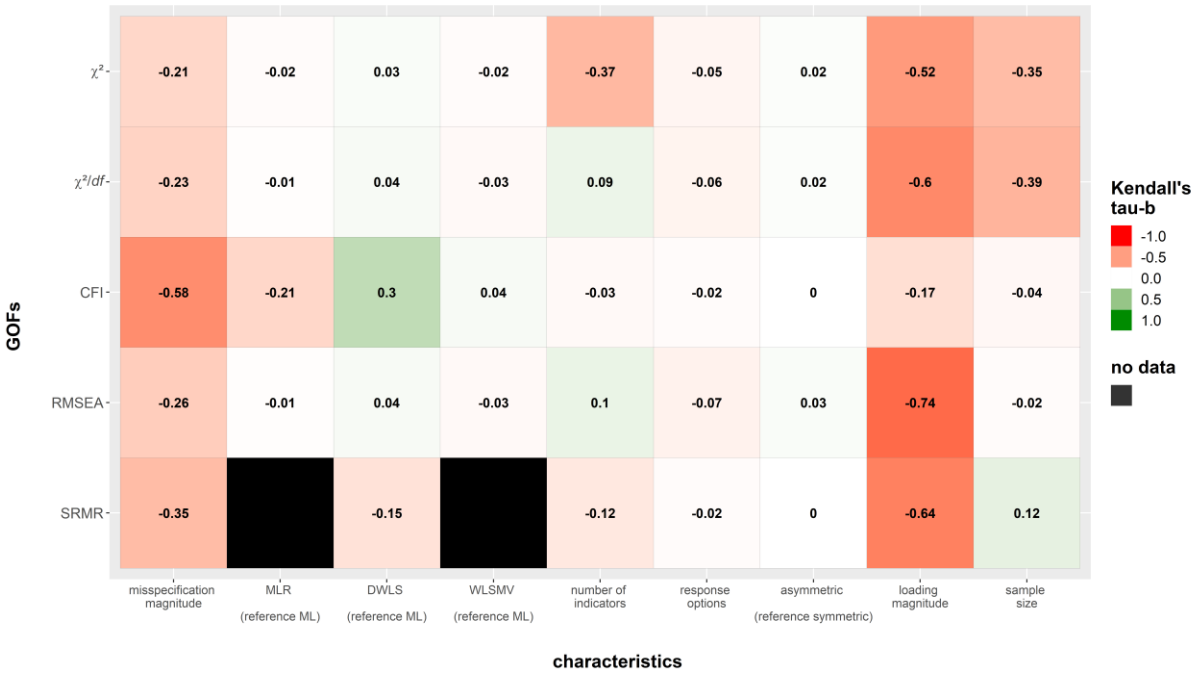


Note. We recoded GOFs so that lower values represent worse fit (i.e., χ^2, χ^2/df, RMSEA, and SRMR were multiplied by −1). MLR = MLR (Yuan & Bentler, 2000). We only displayed SRMR for comparing ML and DWLS because SRMR point estimates are identical for models estimated with ML or MLR and models estimated with DWLS or WLSMV (Maydeu-Olivares et al., 2018).

Misspecified Models. Figure A3 shows the correlations between simulation characteristics and GOFs of analysis models that were misspecified regarding the factor dimensionality of the population models. All GOFs correlated substantially with the magnitude of misspecification ($-.58 \leq$ tau-b $\leq -.21$), as they should. GOFs indicated consistently worse fit with increasing magnitudes of misspecification.

Unfortunately for the applied researcher, GOFs correlated with data and analysis characteristics other than misspecification. In fact, GOF susceptibility to other characteristics was extensive, sometimes exceeding GOF sensitivity to misspecification: All GOFs (except CFI) indicated substantially worse model fit as loading magnitudes increased ($-.74 \leq$ tau-b $\leq -.52$). The χ^2 and χ^2/df increased with sample size (i.e., they have stronger power to detect misspecification). They indicated worse fit with larger samples ($-.39 \leq$ tau-b $\leq -.35$), as intended by their mathematical definition. The χ^2 also increased with the number of indicators (per definition and true in our study, tau-b = −.37), against which the χ^2/df ratio is sufficiently guarded (tau-b = .09).

Figure A3: *Zero-Order Correlation Between Characteristics and GOFs of Models with Misspecified Factor Dimensionality*

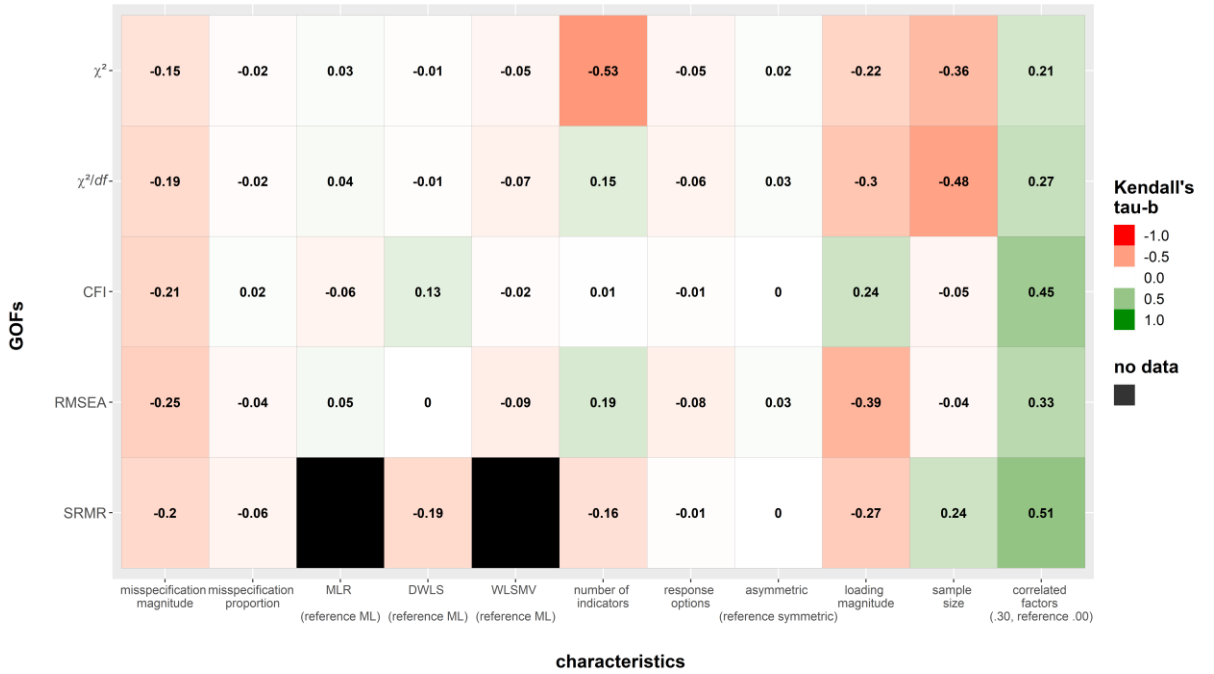


Note. We recoded GOFs so that lower values represent worse fit (i.e., χ^2, χ^2/df, RMSEA, and SRMR were multiplied by −1). MLR = MLR (Yuan & Bentler, 2000). We only displayed SRMR for comparing ML and DWLS because SRMR point estimates are identical for models estimated with ML or MLR and models estimated with DWLS or WLSMV (Maydeu-Olivares et al., 2018).

For models that were misspecified regarding cross-loadings (Figure A4), all GOFs correlated with the magnitude of misspecification (i.e., the size of unmodeled cross-loadings; $-.25 \leq$ tau-b $\leq -.15$). Differently, all GOFs barely correlated with the proportion of misspecification ($.02 \leq |\text{tau-b}| \leq .06$). In sum, all GOFs were sensitive for the magnitude of misspecification. Still, they did not seem sensitive to the proportion of misspecification (which might be spurious due to a moderation effect of the factor correlation, see Figure 4).

Problematically, GOFs of models with unmodeled cross-loadings were susceptible to many characteristics. All else being equal, GOFs showed better fit when factors were correlated instead of uncorrelated ($.21 \leq$ tau-b $\leq .51$). Further, GOFs suggested worse fit with higher loading magnitudes ($-.39 \leq$ tau-b $\leq -.22$), while only CFI pointed to better fit (tau-b = .24). χ^2 was strongly associated with the number of indicators (tau-b = −.53) and the sample size (tau-b = −.36). Also, χ^2/df and SRMR varied immensely with sample size, the former pointing to worse (tau-b = −.48) and the latter to better fit (tau-b = .24) with increasing sample size. GOF correlations with other characteristics were smaller than the correlations with the magnitude of misspecification.

Figure A4: *Zero-Order Correlation Between Characteristics and GOFs of Models with Unmodeled Cross-Loadings*



Note. We recoded GOFs so that lower values represent worse fit (i.e., χ^2, χ^2/df, RMSEA, and SRMR were multiplied by −1). MLR = MLR (Yuan & Bentler, 2000). We only displayed SRMR for comparing ML and DWLS because SRMR point estimates are identical for models estimated with ML or MLR and models estimated with DWLS or WLSMV (Maydeu-Olivares et al., 2018).

Additional File 3: Main and Interaction Effects on GOFs

Table A3: *Conditional Main and Interaction Effects on GOFs Resulting from Correctly Specified Models*

Independent variables	Dependent variables									
	χ^2		χ^2/df		CFI		RMSEA		SRMR	
	1F	2F	1F	2F	1F	2F	1F	2F	1F	2F
Intercept	**−31.882**	**−31.862**	**−1.007**	**−1.007**	**1.002**	**1.003**	**−0.003**	**−0.004**	**−0.009**	**−0.013**
	Main effects									
Estimator (Reference ML)										
MLR	**0.587**	**0.409**	**0.026**	**0.024**	**0.003**	**0.007**	−0.000	**0.001**	–	–
DWLS	**13.755**	**3.685**	**0.440**	**0.062**	**−0.006**	**−0.005**	**0.003**	**−0.001**	**−0.001**	**−0.002**
WLSMV	0.271	−0.001	0.011	−0.004	**−0.003**	−0.001	**−0.002**	−0.000	–	–
Number of indicators	**−7.957**	**−7.830**	0.000	−0.001	**−0.001**	−0.000	**0.000**	**0.000**	**−0.000**	**−0.000**
Response options	**0.182**	**0.168**	**0.007**	**0.005**	**0.000**	0.000	**0.000**	**0.000**	−0.000	**−0.000**
Response options^2	−0.042	0.060	−0.003	0.001	0.000	−0.000	−0.000	0.000	−0.000	**0.000**
Asymmetric (Reference symmetric)	**−3.594**	**−1.300**	**−0.111**	**−0.034**	**−0.001**	−0.000	**−0.004**	**−0.001**	**−0.001**	0.000
Loading magnitude	**−1.032**	**−0.604**	**−0.035**	**−0.017**	**−0.001**	**−0.001**	**−0.001**	**−0.001**	**0.001**	**−0.000**
Loading magnitude^2	**−0.165**	**−0.080**	**−0.006**	−0.001	**0.000**	0.000	−0.000	−0.000	0.000	**−0.000**
Sample size	**0.084**	**0.061**	**0.003**	**0.002**	**0.001**	**0.001**	**0.001**	**0.001**	**0.002**	**0.003**
Sample size^2	**−0.008**	0.001	**−0.000**	0.000	**−0.000**	**−0.000**	**−0.000**	**−0.000**	**−0.000**	**−0.000**
Correlated factors (.30, Reference .00)	–	**0.825**	–	**−0.017**	–	**−0.003**	–	**−0.001**	–	**0.001**
	Two-way interaction effects									
Estimator										
MLR×Number of indicators	**0.610**	**0.257**	**0.002**	**0.008**	**0.001**	**0.001**	**0.000**	**0.000**	–	–
DWLS×Number of indicators	**3.691**	**1.995**	0.001	**0.017**	**0.000**	**0.001**	**0.000**	**0.001**	**−0.000**	−0.000
WLSMV×Number of indicators	**0.697**	**0.313**	0.000	**0.006**	**0.000**	**0.000**	**0.000**	**0.000**	–	–
MLR×Response options	0.016	0.026	0.000	0.002	0.000	**0.000**	0.000	0.000	–	–
DWLS×Response options	**0.403**	**0.179**	**0.013**	**0.004**	**−0.000**	**−0.000**	0.000	−0.000	**0.001**	**0.001**
WLSMV×Response options	−0.053	**−0.060**	−0.002	−0.002	**−0.000**	**−0.000**	−0.000	**−0.000**	–	–
MLR×Response options^2	0.023	0.009	0.000	0.001	−0.000	−0.000	0.000	0.000	–	–
DWLS×Response options^2	**−0.098**	**−0.057**	**−0.004**	−0.002	0.000	0.000	−0.000	−0.000	**−0.000**	**−0.000**
WLSMV×Response options^2	0.025	0.017	0.000	0.000	0.000	0.000	0.000	0.000	–	–
MLR×Asymmetric	**3.462**	**1.240**	**0.108**	**0.030**	**0.002**	0.000	**0.004**	**0.001**	–	–
DWLS×Asymmetric	**3.545**	**1.233**	**0.112**	**0.028**	**0.003**	**0.002**	**0.005**	**0.001**	−0.000	**−0.001**
WLSMV×Asymmetric	**3.539**	**1.293**	**0.112**	**0.030**	**0.003**	**0.001**	**0.005**	**0.001**	–	–
MLR×Loading magnitude	**1.131**	**0.491**	**0.039**	**0.017**	**0.001**	**0.002**	**0.002**	**0.001**	–	–
DWLS×Loading magnitude	**3.876**	**2.141**	**0.123**	**0.061**	**−0.003**	**−0.002**	**0.002**	**0.001**	**0.001**	**0.000**
WLSMV×Loading magnitude	**1.308**	**0.511**	**0.042**	**0.012**	**−0.001**	**−0.000**	**0.002**	**0.001**	–	–
MLR×Loading magnitude^2	**0.152**	0.001	**0.004**	−0.002	**−0.000**	**−0.001**	**0.000**	−0.000	–	–
DWLS×Loading magnitude^2	**0.072**	**0.120**	0.002	**0.004**	**0.001**	**0.000**	−0.000	**−0.000**	**0.000**	0.000
WLSMV×Loading magnitude^2	**0.228**	**0.067**	**0.007**	0.002	**0.000**	−0.000	**0.000**	**0.000**	–	–
MLR×Sample size	**0.053**	**0.075**	**0.003**	**0.004**	**0.000**	**0.001**	**−0.000**	**0.000**	–	–
DWLS×Sample size	−0.010	**0.058**	−0.000	**0.002**	**−0.002**	**−0.001**	**−0.001**	**−0.000**	**0.000**	**0.001**
WLSMV×Sample size	**−0.035**	0.016	−0.001	0.001	**−0.001**	**−0.000**	**−0.000**	**−0.000**	–	–
MLR×Sample size^2	**−0.006**	**−0.009**	**−0.000**	**−0.000**	**−0.000**	**−0.000**	0.000	**−0.000**	–	–
DWLS×Sample size^2	−0.000	**−0.007**	−0.000	−0.000	**0.000**	**0.000**	**0.000**	**0.000**	**−0.000**	**−0.000**
WLSMV×Sample size^2	0.003	−0.002	0.000	−0.000	**0.000**	**0.000**	**0.000**	**0.000**	–	–
MLR×Correlated factors	–	0.118	–	**−0.011**	–	**−0.003**	–	−0.000	–	–
DWLS×Correlated factors	–	**5.954**	–	**0.247**	–	**0.004**	–	**0.007**	–	**0.002**
WLSMV×Correlated factors	–	**0.395**	–	**0.018**	–	0.001	–	0.000	–	–
Number of indicators×										
Response options	**0.016**	0.003	−0.000	−0.000	−0.000	**−0.000**	−0.000	**−0.000**	**0.000**	**0.000**
Response options^2	**−0.020**	**−0.011**	**−0.001**	−0.000	**−0.000**	−0.000	**−0.000**	**−0.000**	**−0.000**	**−0.000**
Asymmetric	**−0.224**	**−0.113**	−0.000	**−0.002**	0.000	0.000	**−0.000**	**−0.000**	**−0.000**	**−0.000**
Loading magnitude	**0.113**	**0.091**	0.000	0.000	**−0.000**	**−0.000**	**−0.000**	**−0.000**	**0.000**	**0.000**
Loading magnitude^2	**−0.018**	0.001	−0.000	−0.000	**0.000**	**0.000**	−0.000	**−0.000**	**0.000**	**0.000**

Independent variables	Dependent variables									
	χ^2		χ^2/df		CFI		RMSEA		SRMR	
	1F	2F	1F	2F	1F	2F	1F	2F	1F	2F
Number of indicators×										
Sample size	**0.022**	**0.026**	−0.000	**0.000**	**−0.000**	**−0.000**	**−0.000**	**−0.000**	**0.000**	**0.000**
Sample size^2	**−0.002**	**−0.003**	−0.000	**−0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**−0.000**	**−0.000**
Correlated factors	–	**0.254**	–	**−0.004**	–	**−0.000**	–	**−0.000**	–	**−0.000**
Response options×										
Asymmetric	**−0.057**	−0.009	−0.002	−0.001	−0.000	−0.000	−0.000	−0.000	**−0.000**	−0.000
Loading magnitude	**0.022**	**0.031**	**0.001**	**0.001**	**−0.000**	**−0.000**	0.000	0.000	**0.000**	**0.000**
Loading magnitude^2	−0.001	−0.000	0.000	−0.000	**0.000**	**0.000**	0.000	−0.000	0.000	0.000
Sample size	**0.021**	**0.023**	**0.001**	**0.001**	**−0.000**	**−0.000**	**0.000**	**0.000**	**−0.000**	**−0.000**
Sample size^2	**−0.003**	**−0.003**	**−0.000**	**−0.000**	0.000	**0.000**	**−0.000**	**−0.000**	**0.000**	**0.000**
Correlated factors	–	0.034	–	**0.002**	–	**−0.000**	–	0.000	–	0.000
Response options^2×										
Asymmetric	0.035	0.019	0.002	0.001	0.000	0.000	0.000	0.000	**0.000**	**0.000**
Loading magnitude	0.002	−0.005	−0.000	−0.000	**0.000**	**0.000**	−0.000	−0.000	**−0.000**	−0.000
Loading magnitude^2	−0.002	−0.003	0.000	−0.000	−0.000	**−0.000**	0.000	−0.000	0.000	−0.000
Sample size	0.001	0.004	**−0.000**	−0.000	**0.000**	0.000	**−0.000**	−0.000	**0.000**	**0.000**
Sample size^2	−0.000	**−0.002**	0.000	−0.000	−0.000	−0.000	**0.000**	−0.000	**−0.000**	**−0.000**
Correlated factors	–	0.014	–	**0.002**	–	0.000	–	**0.000**	–	0.000
Asymmetric×										
Loading magnitude	**−0.314**	**−0.097**	**−0.009**	**−0.002**	**0.000**	**0.001**	**−0.000**	**−0.000**	**0.000**	0.000
Loading magnitude^2	−0.022	−0.010	−0.001	0.000	**−0.000**	**−0.000**	−0.000	−0.000	**0.000**	0.000
Sample size	0.011	−0.006	0.001	−0.000	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Sample size^2	−0.000	0.001	−0.000	0.000	**−0.000**	**−0.000**	**−0.000**	−0.000	**−0.000**	**−0.000**
Correlated factors	–	**−0.221**	–	**−0.007**	–	0.000	–	**−0.000**	–	−0.000
Loading magnitude×										
Sample size	**−0.011**	−0.005	**−0.001**	−0.000	**−0.001**	**−0.001**	**−0.000**	−0.000	**−0.000**	**−0.000**
Sample size^2	0.001	0.001	**0.000**	0.000	**0.000**	**0.000**	**0.000**	0.000	**0.000**	**0.000**
Correlated factors	–	**0.407**	–	**0.014**	–	**−0.001**	–	**0.000**	–	**0.001**
Loading magnitude^2×										
Sample size	0.000	−0.005	−0.000	−0.000	**0.000**	**0.000**	0.000	−0.000	**−0.000**	**−0.000**
Sample size^2	0.000	0.001	0.000	0.000	**−0.000**	**−0.000**	−0.000	0.000	**0.000**	**0.000**
Correlated factors	–	**0.064**	–	**0.003**	–	**0.001**	–	**0.000**	–	**0.000**
Sample size×										
Correlated factors	–	0.014	–	0.000	–	**−0.000**	–	**−0.000**	–	**−0.000**
Sample size^2×										
Correlated factors	–	**−0.004**	–	−0.000	–	**0.000**	–	**0.000**	–	**0.000**
R^2	.894	.860	.266	.061	.213	.194	.219	.130	.893	.815
Number of observations	431,910	849,004	431,910	849,004	431,910	849,004	431,914	849,052	215,957	424,526

Note. We recoded GOFs so that lower values represent worse fit (i.e., χ^2, χ^2/df, RMSEA, and SRMR were multiplied by –1). Relatively large effects are in black, and all others are in light gray color. Those large effects of single independent variables might be small on an absolute scale. Still, they stood out as strong relative to all other effects (and might even aggregate with seemingly small effects of other independent variables). Despite the relative strength of those independent variables, the cutpoint of what constitutes a large and a small effect was arbitrarily set. As linear and quadratic independent variables depend on each other, we always marked both if they (or one of them) were large. 1F = one-factor CFA. 2F = two-factor CFA. The multiplication sign (×) indicates interaction terms. For the correct interpretation of the sample size regression coefficient, multiply the outcome by 100. To correctly interpret the loading magnitude regression coefficient, divide the outcome by 10. The mean-centered independent variables (except binary variables) eased the computation of interaction effects. Regression coefficients are unstandardized. Bold regression coefficients are with $p < .001$. Independent variables with more than two simulated levels were entered additionally in quadratic form. SRMR is only available for comparing ML and DWLS because SRMR point estimates are identical for models with ML and MLR estimators and models with DWLS and WLSMV estimators (Maydeu-Olivares et al., 2018).

Table A4: *Conditional Main and Interaction Effects on GOFs Resulting from Misspecified Models*

Independent variables	Dependent variables									
	χ^2		χ^2/df		CFI		RMSEA		SRMR	
	Dim.	Load.	Dim.	Load.	Dim.	Load.	Dim.	Load.	Dim.	Load.
Intercept	−167.174	**−112.681**	1.707	**−4.785**	**0.797**	**0.936**	**−0.109**	**−0.061**	**−0.065**	**−0.057**
	Main effects									
Misspecification magnitude	−130.557	**−79.952**	−8.623	**−3.347**	**−0.077**	**−0.057**	**−0.017**	**−0.027**	**−0.013**	**−0.023**
Misspecification magnitude^2	−39.354	–	−3.593	–	**−0.006**	–	**0.001**	–	**−0.000**	–
Misspecification proportion	–	**−27.943**	–	**−1.187**	–	**−0.015**	–	**−0.008**	–	**−0.011**
Estimator (Reference ML)										
MLR	−514.771	0.555	−47.487	0.022	**−0.053**	**0.004**	**−0.017**	**0.000**	–	–
DWLS	1.962	**−138.532**	1.299	**−4.315**	**0.104**	**−0.023**	**0.002**	**−0.033**	**−0.012**	**−0.012**
WLSMV	−96.210	**−71.258**	−2.679	**−2.038**	**0.026**	**−0.026**	**−0.013**	**−0.018**	–	–
Number of indicators	−45.523	**−23.113**	3.209	**0.549**	**−0.003**	**0.004**	**0.005**	**0.004**	**−0.001**	**−0.001**
Response options	−25.684	**−6.994**	−0.792	**−0.287**	**−0.004**	**0.000**	**−0.005**	**−0.002**	**−0.003**	**−0.002**
Response options^2	1.878	**2.220**	−0.516	**0.099**	**0.001**	0.000	**0.002**	**0.001**	**0.001**	**0.001**
Asymmetric (Reference symmetric)	−89.771	**10.346**	−10.766	**0.471**	−0.000	0.000	**0.005**	**0.003**	**0.002**	**0.003**
Loading magnitude	**−321.908**	**−28.669**	−14.605	**−1.313**	**−0.042**	**0.027**	**−0.048**	**−0.011**	**−0.024**	**−0.010**
Loading magnitude^2	−120.149	**−2.277**	−6.660	**−0.148**	**−0.006**	**−0.005**	**−0.007**	**−0.000**	**−0.003**	**−0.000**
Sample size	−43.171	**−10.725**	−2.315	**−0.468**	**−0.000**	**0.000**	**−0.000**	**−0.000**	**0.001**	**0.001**
Sample size^2	−0.320	0.005	−0.026	0.000	0.000	**−0.000**	**0.000**	**0.000**	**−0.000**	**−0.000**
Correlated factors (.30, Reference .00)	–	**49.824**	–	**2.236**	–	**0.044**	–	**0.021**	–	**0.029**
	Two-way interaction effects									
Misspecification magnitude×										
Misspecification proportion	–	**−9.422**	–	**−0.373**	–	0.000	–	**−0.000**	–	**−0.002**
MLR	−100.276	**1.710**	−10.052	**0.033**	**−0.020**	**−0.001**	**−0.003**	0.000	–	–
DWLS	−74.353	**−56.634**	−2.114	**−1.648**	**0.020**	**−0.001**	**−0.008**	**−0.007**	**−0.003**	**−0.003**
WLSMV	−20.426	**−33.341**	−0.707	**−0.953**	**0.005**	**−0.005**	**−0.003**	**−0.005**	–	–
Number of indicators	**−17.863**	**−12.204**	1.211	**0.283**	**0.000**	**0.001**	**0.001**	**0.001**	**−0.000**	**−0.000**
Response options	−9.313	**−5.940**	−0.357	**−0.246**	−0.000	**−0.000**	**−0.001**	**−0.001**	**−0.000**	**−0.000**
Response options^2	15.766	**1.615**	1.566	**0.067**	0.000	0.000	**0.000**	**0.000**	**0.000**	**0.000**
Asymmetric	−21.413	**8.468**	−3.397	**0.346**	**0.001**	**0.001**	**0.001**	**0.001**	**0.000**	**0.001**
Loading magnitude	**−67.084**	**−22.930**	−2.676	**−0.976**	**−0.000**	**0.008**	**−0.005**	**−0.005**	**−0.006**	**−0.004**
Loading magnitude^2	1.861	**−2.612**	1.115	**−0.145**	**0.003**	**−0.000**	**0.000**	**−0.000**	**−0.001**	**−0.000**
Sample size	**−17.483**	**−7.797**	−0.877	**−0.308**	**−0.001**	**−0.001**	**−0.000**	−0.000	**−0.000**	**−0.000**
Sample size^2	−0.156	**0.045**	−0.017	0.000	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Correlated factors	–	**62.662**	–	**2.477**	–	**0.048**	–	**0.014**	–	**0.015**
Misspecification magnitude^2×										
MLR	24.642	–	0.677	–	**0.004**	–	**0.002**	–	–	–
DWLS	−18.140	–	−0.682	–	**−0.004**	–	**−0.001**	–	**0.000**	–
WLSMV	−5.441	–	−0.256	–	**−0.003**	–	**−0.000**	–	–	–
Number of indicators	0.820	–	0.006	–	**−0.000**	–	**−0.000**	–	**−0.000**	–
Response options	1.643	–	−0.006	–	**0.000**	–	**0.000**	–	0.000	–
Response options^2	5.857	–	0.705	–	**−0.000**	–	−0.000	–	0.000	–
Asymmetric	−18.668	–	−2.137	–	**0.000**	–	−0.000	–	**0.000**	–
Loading magnitude	13.016	–	1.016	–	**0.003**	–	**0.001**	–	**−0.000**	–
Loading magnitude^2	15.872	–	1.401	–	**0.001**	–	**0.000**	–	**−0.000**	–
Sample size	0.590	–	0.024	–	**−0.000**	–	**0.000**	–	**−0.000**	–
Sample size^2	−0.087	–	−0.013	–	**0.000**	–	−0.000	–	**0.000**	–
Misspecification proportion×										
MLR	–	**0.608**	–	**0.051**	–	**0.001**	–	**0.001**	–	–
DWLS	–	**−37.402**	–	**−1.313**	–	**−0.009**	–	**−0.006**	–	**−0.001**
WLSMV	–	**−21.313**	–	**−0.811**	–	**−0.009**	–	**−0.005**	–	–
Number of indicators	–	**−2.433**	–	**0.057**	–	**0.000**	–	**0.000**	–	**0.000**
Response options	–	**−0.857**	–	**−0.029**	–	**0.000**	–	**−0.000**	–	**−0.000**
Response options^2	–	**0.267**	–	**0.009**	–	**−0.000**	–	**−0.000**	–	**0.000**
Asymmetric	–	**1.656**	–	**0.066**	–	0.000	–	**0.000**	–	**0.000**
Loading magnitude	–	**−2.314**	–	**−0.072**	–	**0.002**	–	**0.000**	–	**−0.001**
Loading magnitude^2	–	**0.717**	–	**0.031**	–	**0.000**	–	**0.000**	–	−0.000

Independent variables	Dependent variables									
	χ^2		χ^2/df		CFI		RMSEA		SRMR	
	Dim.	Load.	Dim.	Load.	Dim.	Load.	Dim.	Load.	Dim.	Load.
Misspecification proportion×										
Sample size	–	**−1.576**	–	**−0.061**	–	**−0.000**	–	**0.000**	–	**−0.000**
Sample size^2	–	**0.011**	–	0.000	–	**0.000**	–	−0.000	–	**0.000**
Correlated factors	–	**49.403**	–	**2.008**	–	**0.025**	–	**0.015**	–	**0.013**
Estimator										
MLR×Number of indicators	47.694	**0.922**	6.461	**0.015**	**0.009**	**0.001**	**0.002**	**0.000**	–	–
DWLS×Number of indicators	−29.986	**−16.612**	−0.829	**−0.102**	**0.005**	0.000	**−0.002**	**−0.001**	**0.000**	**0.000**
WLSMV×Number of indicators	−18.247	**−11.131**	−0.167	**−0.067**	**0.001**	**−0.002**	**−0.001**	**−0.001**	–	–
MLR×Response options	−8.729	0.184	0.153	0.004	0.000	0.000	−0.000	0.000	–	–
DWLS×Response options	−12.392	**−6.535**	−0.221	**−0.174**	**0.003**	−0.000	**−0.001**	**−0.001**	**0.003**	**0.002**
WLSMV×Response options	−3.316	**−3.389**	−0.098	**−0.097**	**0.001**	**−0.000**	**−0.001**	**−0.001**	–	–
MLR×Response options^2	73.658	−0.009	7.895	−0.000	0.000	−0.000	0.000	0.000	–	–
DWLS×Response options^2	2.547	**1.561**	0.022	**0.040**	**−0.001**	−0.000	0.000	**0.000**	**−0.001**	**−0.001**
WLSMV×Response options^2	0.718	**0.744**	0.009	**0.018**	**−0.000**	0.000	0.000	**0.000**	–	–
MLR×Asymmetric	−133.306	**4.543**	−18.182	**0.151**	−0.000	−0.000	**0.006**	**0.002**	–	–
DWLS×Asymmetric	40.572	**8.876**	1.309	**0.215**	**0.004**	**0.001**	**0.004**	**0.002**	**−0.001**	**−0.003**
WLSMV×Asymmetric	36.456	**4.791**	1.338	**0.103**	**0.005**	**0.001**	**0.004**	**0.001**	–	–
MLR×Loading magnitude	−66.301	**1.634**	−5.498	**0.042**	**−0.005**	**0.001**	**−0.002**	**0.001**	–	–
DWLS×Loading magnitude	−50.201	**−24.485**	−0.466	**−0.655**	**0.035**	**0.005**	**−0.001**	**−0.002**	**−0.004**	**−0.001**
WLSMV×Loading magnitude	−21.672	**−8.976**	−0.458	**−0.233**	**0.029**	**0.008**	**−0.001**	**−0.001**	–	–
MLR×Loading magnitude^2	58.306	**0.282**	6.659	0.005	**−0.001**	**−0.001**	0.000	−0.000	–	–
DWLS×Loading magnitude^2	−19.863	**−1.374**	−0.259	**−0.016**	**0.007**	**0.001**	**−0.001**	**0.000**	**−0.001**	**−0.000**
WLSMV×Loading magnitude^2	9.300	**3.151**	0.356	**0.110**	**0.012**	**0.001**	**0.002**	**0.001**	–	–
MLR×Sample size	−27.493	**0.329**	−2.630	**0.010**	**0.000**	**0.001**	−0.000	**0.000**	–	–
DWLS×Sample size	−8.532	**−8.106**	−0.061	**−0.220**	**−0.001**	**−0.001**	**−0.001**	**−0.001**	**0.000**	**0.000**
WLSMV×Sample size	−7.576	**−5.017**	−0.184	**−0.136**	−0.000	**−0.000**	−0.000	**−0.000**	–	–
MLR×Sample size^2	−0.891	−0.011	−0.078	−0.000	**−0.000**	**−0.000**	−0.000	**−0.000**	–	–
DWLS×Sample size^2	0.014	**0.064**	−0.000	**0.001**	**0.000**	**0.000**	**0.000**	**0.000**	**−0.000**	**−0.000**
WLSMV×Sample size^2	0.013	**0.036**	0.000	0.001	0.000	**0.000**	0.000	**0.000**	–	–
MLR×Correlated factors	–	−0.251	–	**−0.050**	–	**−0.002**	–	**−0.000**	–	–
DWLS×Correlated factors	–	**138.007**	–	**4.672**	–	**0.032**	–	**0.037**	–	**0.010**
WLSMV×Correlated factors	–	**31.241**	–	**0.861**	–	**0.025**	–	**0.006**	–	–
Number of indicators×										
Response options	−8.834	**−1.280**	0.113	**0.035**	−0.000	**−0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Response options^2	−3.563	**0.334**	−0.693	**−0.010**	0.000	**−0.000**	**−0.000**	**−0.000**	−0.000	**−0.000**
Asymmetric	24.401	**1.780**	1.497	**−0.051**	−0.000	−0.000	**−0.000**	**−0.000**	**−0.000**	**−0.000**
Loading magnitude	**−47.821**	**−4.470**	1.357	**0.151**	**0.000**	**−0.001**	**0.002**	**0.001**	**−0.000**	**0.000**
Loading magnitude^2	**−18.627**	**−0.341**	−0.250	**0.027**	−0.000	**0.000**	**0.000**	**0.000**	**−0.000**	**0.000**
Sample size	**−8.761**	**−1.714**	0.385	**0.043**	**0.000**	**−0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Sample size^2	0.039	**0.003**	0.005	**−0.000**	**−0.000**	0.000	**−0.000**	**−0.000**	**−0.000**	**−0.000**
Correlated factors	–	**12.477**	–	**−0.287**	–	**−0.004**	–	**−0.002**	–	**−0.000**
Response options×										
Asymmetric	7.955	**1.669**	0.067	**0.077**	**0.001**	0.000	**0.001**	**0.001**	**0.001**	**0.000**
Loading magnitude	−24.744	**−2.691**	−0.906	**−0.117**	**−0.000**	**−0.000**	**−0.002**	**−0.001**	**−0.001**	**−0.000**
Loading magnitude^2	−6.760	**−0.442**	−0.262	**−0.024**	**0.000**	0.000	**−0.000**	**−0.000**	**−0.000**	**−0.000**
Sample size	−4.875	**−0.835**	−0.172	**−0.033**	**0.000**	−0.000	**0.000**	**0.000**	**−0.000**	**−0.000**
Sample size^2	−0.035	**0.006**	−0.002	−0.000	**−0.000**	−0.000	**−0.000**	**−0.000**	**0.000**	**0.000**
Correlated factors	–	**4.480**	–	**0.160**	–	**−0.000**	–	**0.001**	–	**0.001**
Response options^2×										
Asymmetric	30.370	**−0.640**	3.721	**−0.028**	**−0.000**	0.000	**−0.001**	**−0.000**	**−0.000**	**−0.000**
Loading magnitude	7.021	**0.729**	0.267	**0.031**	**0.000**	**0.000**	**0.001**	**0.000**	**0.000**	**0.000**
Loading magnitude^2	−10.971	**0.109**	−1.359	**0.006**	−0.000	0.000	**0.000**	**0.000**	**0.000**	**0.000**
Sample size	3.512	**0.246**	0.292	**0.010**	−0.000	**0.000**	−0.000	**0.000**	**0.000**	**0.000**
Sample size^2	0.113	−0.004	0.014	−0.000	0.000	**−0.000**	0.000	**−0.000**	**−0.000**	**−0.000**
Correlated factors	–	**−1.207**	–	**−0.047**	–	**0.000**	–	**−0.000**	–	**−0.000**
Asymmetric×										
Loading magnitude	28.194	**3.714**	1.210	**0.158**	**−0.001**	**0.001**	**0.002**	**0.001**	−0.000	**0.000**
Loading magnitude^2	42.004	**0.509**	4.232	**0.027**	**−0.001**	−0.000	0.000	**0.000**	**−0.000**	0.000
Sample size	0.202	**1.264**	−0.370	**0.050**	**0.000**	**0.000**	0.000	**0.000**	**0.000**	**0.000**
Sample size^2	−0.558	−0.009	−0.063	0.000	**−0.000**	**−0.000**	−0.000	**−0.000**	**−0.000**	**−0.000**

Independent variables	Dependent variables									
	χ^2		χ^2/df		CFI		RMSEA		SRMR	
	Dim.	Load.	Dim.	Load.	Dim.	Load.	Dim.	Load.	Dim.	Load.
Asymmetric×										
Correlated factors	–	**−7.959**	–	**−0.308**	–	−0.000	–	**−0.002**	–	**−0.001**
Loading magnitude×										
Sample size	**−32.467**	**−3.158**	−1.358	**−0.131**	**−0.000**	**−0.000**	0.000	**−0.000**	**−0.000**	**−0.000**
Sample size^2	0.006	**0.015**	0.002	**−0.000**	**0.000**	**0.000**	−0.000	**0.000**	**0.000**	**0.000**
Correlated factors	–	**9.909**	–	**0.340**	–	**−0.030**	–	**−0.001**	–	**0.005**
Loading magnitude^2×										
Sample size	−6.888	**−0.310**	−0.153	**−0.019**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.000
Sample size^2	0.186	**−0.007**	0.021	0.000	**−0.000**	**−0.000**	**−0.000**	**−0.000**	**−0.000**	0.000
Correlated factors	–	**−3.413**	–	**−0.153**	–	**0.005**	–	**−0.001**	–	**−0.000**
Sample size×										
Correlated factors	–	**8.014**	–	**0.310**	–	**0.000**	–	**−0.000**	–	**0.000**
Sample size^2×										
Correlated factors	–	**−0.037**	–	**−0.001**	–	**−0.000**	–	−0.000	–	**−0.000**
R^2	.002	.736	.000	.780	.719	.619	.684	.697	.956	.845
Number of observations	1,286,474	3,386,579	1,286,474	3,386,579	1,286,474	3,386,579	1,289,172	3,386,706	644,586	1,693,353

Note. We recoded GOFs so that lower values represent worse fit (i.e., χ^2, χ^2/df, RMSEA, and SRMR were multiplied by –1). Relatively large effects are in black, and all other effects are in light gray color. We marked those effects as relatively large that were equal to or larger than the effects of misspecification. For scenarios with misspecified factor dimensionality (Dim.), we marked those effects that were equal to or larger than the linear plus quadratic main effect of the magnitude of misspecification (as both depend on each other). For scenarios with unmodeled cross-loadings (Load.), we marked those effects that were equal to or larger than either the main effect of the magnitude of misspecification or the proportion of misspecification (as both were independent in misspecified data). As linear and quadratic independent variables depend on each other, we always marked both if they (or one of them) were large. MLR = MLR (Yuan & Bentler, 2000). Dim. = misspecified factor dimensionality. Load. = unmodeled cross-loadings. The multiplication sign (×) indicates interaction terms. For the correct interpretation of the sample size regression coefficient, multiply the outcome by 100. For the correct interpretation of the misspecification magnitude, misspecification proportion, and loading magnitude regression coefficient, divide the outcome by 10. The mean-centered independent variables (except binary variables) eased the computation of interaction effects. Regression coefficients are unstandardized. Bold regression coefficients are with $p < .001$. Independent variables with more than two simulated levels were entered additionally in quadratic form. (Be careful: Only the magnitude of misspecification for models with misspecified factor dimensionality had more than two simulated levels, but not the magnitude of misspecification for models with unmodeled cross-loadings.) SRMR is only available for comparing ML and DWLS because SRMR point estimates are identical for models with ML and MLR estimators and models with DWLS and WLSMV estimators (Maydeu-Olivares et al., 2018).

Table A5: *Summary of the Sensitivities and Susceptibilities of GOFs to Model Misspecification and Other Influences Including Findings from Previous Studies*

Independent variables	Dependent variables									
	χ^2		χ^2/df		CFI		RMSEA		SRMR	
	correct (1F/2F)	misspecified (Dim./Load.)	correct (1F/2F)	misspecified (Dim./Load.)	correct (1F/2F)	misspecified (Dim./Load.)	correct (1F/2F)	misspecified (Dim./Load.)	correct (1F/2F)	misspecified (Dim./Load.)
	[findings from previous studies]									
Main effects										
Misspecification magnitude		− [−]		− [−]		− [−]		− [−]		− [−]
Misspecification proportion[a]		[−]		− [−]		− [−]		− [−]		− [−]
Estimator (Reference ML)										
MLR		− (Dim.)		− (Dim.)	+ (2F)			− (Dim.)		
DWLS	+ [+]	− (Load.) [+]	+	− (Load.)	[+]	+ / − [+]	+ (1F) [+]	− (Load.) [+]	− (2F) [−]	− (Load.)
WLSMV	[+]	− (Load.) [+]		− (Load.)	[+]	− (Load.) [+]	[+]	− (Load.) [+]		
Number of indicators	− [−]	[−]		[−]	[−]	[−]		[+]		[0/−]
Response options										
Asymmetric (Reference symmetric)	−		−		[−]	[+]	− (1F) [−]	[+]	[−]	[+]
Loading magnitude		− [−]		−	[+]	+ (Load.)		− [−]	[+]	− (Dim.) [−]
Sample size	[+]	[−]	[+]	[−]	[+]	[+]	[+]	[−]	+ [+]	[+]
Correlated factors (.30, Reference .00)[a]	[0]	+ [+]		+	[0]	+	[0]	+	[0]	+ [+]
Large two-way interaction effects										
Misspecification magnitude×										
DWLS		− (Load.)		− (Load.)						
WLSMV		− (Load.)								
Correlated factors[a]		+		+		+		+		+
Misspecification proportion[a] ×										
DWLS		−		−						
Correlated factors[a]		+		+		+		+		+
MLR×										
Asymmetric	+		+	− (Dim.)			+ (1F)			
DWLS×										
Number of indicators	+									
Response options		[−]		[−]		[−]		[−]	[+]	
Asymmetric	+	[+]	+	[+]		[+]	+ (1F)	[+]		
Loading magnitude	+		+							
Correlated factors[a]	+	+	+	+	+	+	+	+	+	
WLSMV×										
Response options	[−]	[−]		[−]	[−]	[−]	[−]	[−]		
Asymmetric	+	[+]	+	[+]		[+]	+ (1F)	[+]		
Correlated factors[a]						+				
Number of indicators×										
Sample size			[U]				[U]			

Independent variables	Dependent variables									
	χ^2		χ^2/df		CFI		RMSEA		SRMR	
	correct (1F/2F)	misspecified (Dim./Load.)	correct (1F/2F)	misspecified (Dim./Load.)	correct (1F/2F)	misspecified (Dim./Load.)	correct (1F/2F)	misspecified (Dim./Load.)	correct (1F/2F)	misspecified (Dim./Load.)
	[findings from previous studies]									
Loading magnitude× Correlated factors [a]						–				

Note. We recoded GOFs so that lower values represent worse fit (i.e., χ^2, χ^2/df, RMSEA, and SRMR were multiplied by −1). Thus, "+" points to improving fit with the increase of a characteristic, "–" to worse fit, "∪" to initially worse, then better fit. Blank gray cells indicate that the scenario was not/could not be tested in our simulation. Blank white cells indicate that we found no (relatively large) effect. Brackets indicate effects that apply only to certain scenarios (printed in light gray color). If we found different effects per type of correctly specified or misspecified model, we separated the effects with a slash (1F/2F and Dim./Load., respectively). Effects from previous studies are in square brackets (including no effects, "0"). 1F = one-factor CFA. 2F = two-factor CFA. Correct = correctly specified models. Misspecified = misspecified models. Dim. = misspecified factor dimensionality. Load. = unmodeled cross-loadings. MLR = MLR (Yuan & Bentler, 2000). [a]Only for GOFs from two-factor models (2F) and models with unmodeled cross-loadings (Load.). The multiplication sign (×) indicates interaction terms. SRMR is only available for comparing ML and DWLS because SRMR point estimates are identical for models with ML and MLR estimators and models with DWLS and WLSMV estimators (Maydeu-Olivares et al., 2018). We based the summary table on the findings from Table A3 for correctly specified models and Table A4 for misspecified models in this Additional File 3 from the Supplementary Material. The table includes main and large two-way interaction effects.

Additional File 4: 5%/95% Quantiles of GOF Distributions from Correctly Specified Models

Table A6: *95% Quantiles of χ^2 Distributions Resulting from Correctly Specified Models*

sample size	number of indicators	loading magnitude	response options	distribution	ML			MLR			DWLS			WLSMV		
					1F	2F		1F	2F		1F	2F		1F	2F	
						.30	.00		.30	.00		.30	.00		.30	.00
200	6	.40	3	sym	16.4	13.8	17.3	19.0	20.1	17.3	13.2	11.9	17.8	16.1	13.8	17.3
200	6	.40	3	asym	17.1	13.9	16.6	19.6	21.3	16.6	13.3	13.0	17.8	16.2	14.7	17.3
200	6	.40	5	sym	17.4	14.1	17.4	19.4	20.4	17.5	13.6	13.0	19.0	17.2	14.7	18.1
200	6	.40	5	asym	17.7	14.7	17.4	19.4	21.1	17.8	13.8	12.6	18.6	17.6	14.8	18.1
200	6	.40	7	sym	18.3	14.0	17.4	19.9	19.3	17.5	14.3	13.1	19.1	18.1	14.9	18.4
200	6	.40	7	asym	18.2	14.1	17.3	21.0	18.9	17.7	13.9	12.6	18.7	17.7	14.2	18.0
200	6	.60	3	sym	17.4	15.7	17.0	17.2	17.2	17.3	10.4	12.5	18.2	16.0	15.6	16.6
200	6	.60	3	asym	19.2	15.8	17.3	17.9	16.4	17.6	10.9	12.1	19.0	16.8	15.3	17.4
200	6	.60	5	sym	18.3	16.3	16.9	18.7	17.2	16.9	10.4	12.9	19.7	17.3	16.7	17.6
200	6	.60	5	asym	20.4	17.2	17.6	19.3	18.4	17.8	11.5	13.1	20.0	17.9	16.8	18.0
200	6	.60	7	sym	17.7	15.7	17.0	17.9	16.5	17.3	10.2	11.5	21.1	17.2	15.0	18.7
200	6	.60	7	asym	20.1	16.1	16.7	19.0	16.4	17.0	10.9	12.6	20.4	17.6	16.0	18.4
200	6	.80	3	sym	17.9	16.1	16.7	17.1	16.3	17.2	8.3	10.2	20.9	15.5	15.8	16.7
200	6	.80	3	asym	21.9	17.3	17.3	17.9	16.6	17.2	8.3	9.9	22.6	16.3	15.6	17.7
200	6	.80	5	sym	18.1	16.6	16.8	17.8	16.8	16.6	7.8	9.6	24.2	16.8	17.0	17.9
200	6	.80	5	asym	22.2	16.4	17.7	18.2	16.1	18.0	7.8	9.6	23.0	16.9	16.1	17.4
200	6	.80	7	sym	18.9	16.0	17.9	18.1	16.3	17.8	7.4	8.9	27.0	16.5	16.1	19.6
200	6	.80	7	asym	22.1	16.2	17.1	17.7	16.1	17.4	7.3	9.4	24.7	16.6	15.9	18.2
200	12	.40	3	sym	77.3	72.8	74.6	78.4	75.8	77.8	61.6	64.6	70.9	73.4	70.2	71.9
200	12	.40	3	asym	75.8	73.2	74.9	75.0	75.9	77.4	59.6	63.9	70.2	71.2	70.0	71.5
200	12	.40	5	sym	74.9	72.5	75.4	76.4	75.6	77.0	58.5	64.6	73.8	72.5	71.6	74.8
200	12	.40	5	asym	76.7	73.9	73.7	75.0	75.5	75.4	59.0	65.8	72.2	71.8	72.1	73.6
200	12	.40	7	sym	77.6	74.8	76.3	78.9	76.9	78.1	61.4	68.1	75.9	76.1	74.9	76.7
200	12	.40	7	asym	79.4	75.2	75.2	77.6	77.0	77.6	60.6	67.1	74.6	75.8	74.5	75.7
200	12	.60	3	sym	75.3	72.1	75.1	74.9	72.8	75.3	46.3	55.8	73.5	69.8	69.4	73.3
200	12	.60	3	asym	83.4	78.5	77.2	76.1	75.5	75.8	47.1	57.6	74.7	71.2	71.4	75.3
200	12	.60	5	sym	75.9	73.4	73.6	74.9	73.7	74.2	43.8	56.3	74.6	71.1	72.6	74.5
200	12	.60	5	asym	83.7	78.4	77.8	75.9	75.3	76.4	44.2	58.2	75.0	70.5	73.5	75.1
200	12	.60	7	sym	77.6	75.7	76.8	75.8	75.9	76.8	44.1	57.2	80.1	73.0	73.7	78.1
200	12	.60	7	asym	84.1	78.6	77.2	76.4	76.2	75.3	45.6	58.4	78.8	73.3	74.5	77.2
200	12	.80	3	sym	80.2	74.5	77.0	75.6	73.7	76.7	37.7	46.9	88.0	69.8	70.5	75.1
200	12	.80	3	asym	95.9	81.9	81.7	76.9	74.3	76.8	35.6	46.0	86.8	69.6	70.2	75.8
200	12	.80	5	sym	77.9	74.3	76.2	75.4	73.9	75.8	32.4	42.9	92.3	69.1	71.1	75.7
200	12	.80	5	asym	93.3	81.2	80.0	76.4	74.2	75.6	31.5	43.5	91.2	69.3	71.6	75.8
200	12	.80	7	sym	81.2	77.3	77.6	76.3	76.2	76.4	31.1	44.5	106.2	69.8	74.0	80.4
200	12	.80	7	asym	96.9	84.0	81.6	79.4	76.0	77.2	31.5	44.4	101.6	70.9	74.6	79.4
500	6	.40	3	sym	17.3	15.4	17.5	18.0	18.8	17.5	13.6	13.9	18.1	17.0	15.5	17.7
500	6	.40	3	asym	17.6	15.4	17.4	17.8	19.8	17.4	13.6	13.7	17.7	17.1	15.3	17.4
500	6	.40	5	sym	16.8	15.2	17.0	17.3	17.5	17.1	12.9	13.6	17.6	16.8	15.5	17.2
500	6	.40	5	asym	18.4	15.6	17.0	18.7	17.3	17.0	13.4	13.7	17.6	16.9	15.9	17.3
500	6	.40	7	sym	15.9	15.5	17.5	16.3	17.3	17.6	12.3	13.9	18.5	16.4	15.6	18.0
500	6	.40	7	asym	17.4	15.8	17.6	17.3	17.4	17.6	13.0	13.7	18.5	16.9	15.4	18.1
500	6	.60	3	sym	17.8	15.7	16.3	17.8	15.9	16.2	10.7	12.2	18.4	17.2	15.9	17.0
500	6	.60	3	asym	19.0	15.9	17.6	17.3	16.1	17.5	10.6	12.3	19.0	16.6	15.6	17.6
500	6	.60	5	sym	17.0	16.2	17.0	16.8	16.3	17.3	9.9	12.6	20.0	16.9	16.8	18.0
500	6	.60	5	asym	18.4	17.0	17.6	16.8	17.0	17.5	10.2	12.5	19.6	17.0	16.5	17.7
500	6	.60	7	sym	17.6	15.8	17.4	17.6	15.8	17.2	9.8	11.7	19.5	17.4	15.9	17.4
500	6	.60	7	asym	19.2	16.4	17.0	17.2	16.3	17.1	10.0	12.3	19.7	17.3	16.2	18.0
500	6	.80	3	sym	18.1	15.8	16.7	17.4	15.8	16.8	8.5	9.9	23.0	16.8	15.5	18.2
500	6	.80	3	asym	21.8	16.2	16.8	17.1	15.6	16.7	7.9	9.5	22.9	16.4	15.7	18.1
500	6	.80	5	sym	17.9	16.3	16.8	17.5	16.1	16.9	7.2	9.3	22.8	16.6	16.2	17.2
500	6	.80	5	asym	21.6	16.9	17.4	17.1	16.1	17.4	6.9	9.1	20.6	16.2	15.7	16.0
500	6	.80	7	sym	18.5	15.8	16.6	17.6	15.9	16.7	7.1	8.7	22.8	16.8	15.9	17.1
500	6	.80	7	asym	22.5	16.6	17.6	18.0	16.0	17.4	7.1	9.0	22.8	16.8	16.2	17.3

sample size	number of indicators	loading magnitude	response options	distribution	ML			MLR			DWLS			WLSMV		
					1F	2F		1F	2F		1F	2F		1F	2F	
						.30	.00		.30	.00		.30	.00		.30	.00
500	12	.40	3	sym	73.4	72.1	73.3	73.6	73.6	74.2	57.7	63.6	70.1	71.2	70.9	72.8
500	12	.40	3	asym	75.6	72.9	75.1	73.1	72.5	74.9	57.1	64.4	72.7	71.1	71.1	75.1
500	12	.40	5	sym	73.8	72.3	74.7	74.2	73.0	75.2	57.1	62.8	71.4	72.5	71.8	74.3
500	12	.40	5	asym	76.0	73.1	74.8	73.6	73.2	74.9	57.7	62.9	71.6	73.0	71.4	74.5
500	12	.40	7	sym	73.1	73.1	73.2	73.1	73.9	74.3	55.8	63.8	69.7	72.6	73.0	72.9
500	12	.40	7	asym	77.4	74.8	74.6	75.2	74.8	74.5	57.6	63.6	70.0	73.2	72.5	73.0
500	12	.60	3	sym	74.5	71.3	73.2	73.6	71.1	73.5	46.6	55.9	72.7	72.2	71.3	74.4
500	12	.60	3	asym	82.1	75.5	74.9	74.0	71.8	72.9	46.7	55.9	73.0	72.5	70.9	75.2
500	12	.60	5	sym	74.3	73.6	73.6	73.3	73.1	73.3	42.0	55.0	73.3	71.3	72.9	74.2
500	12	.60	5	asym	81.6	76.3	78.0	74.1	72.4	75.6	43.4	55.1	72.6	71.9	71.9	74.0
500	12	.60	7	sym	75.9	72.7	74.7	74.5	72.3	74.6	42.8	55.3	74.5	74.0	74.1	74.5
500	12	.60	7	asym	81.5	75.2	76.4	73.4	71.3	74.2	43.2	54.1	71.5	72.6	71.5	73.0
500	12	.80	3	sym	77.9	72.8	74.6	73.9	72.0	73.4	36.9	47.9	88.2	71.8	73.2	75.7
500	12	.80	3	asym	92.8	80.5	79.0	72.8	72.4	73.0	34.1	46.2	89.7	70.6	72.2	77.9
500	12	.80	5	sym	75.8	73.1	73.7	72.5	72.2	72.8	30.7	41.7	93.7	70.0	71.9	76.2
500	12	.80	5	asym	91.1	80.6	81.4	73.1	72.8	75.7	30.5	41.8	89.8	70.0	71.6	74.9
500	12	.80	7	sym	78.2	72.6	73.6	74.3	71.1	72.4	29.6	40.9	89.8	71.5	72.5	73.9
500	12	.80	7	asym	94.6	81.3	79.4	75.7	73.0	73.9	30.0	41.9	89.0	71.1	73.0	74.3
2,000	6	.40	3	sym	17.2	14.8	17.6	17.3	15.3	17.6	13.4	14.0	17.7	16.9	15.4	17.4
2,000	6	.40	3	asym	17.1	15.3	16.7	16.6	15.7	16.7	13.1	13.7	17.1	16.5	15.4	16.9
2,000	6	.40	5	sym	16.7	15.2	16.7	16.9	15.8	16.6	12.7	13.9	17.0	16.7	15.8	16.6
2,000	6	.40	5	asym	17.0	15.5	16.9	16.5	15.7	16.9	12.7	13.7	17.5	16.4	15.6	17.2
2,000	6	.40	7	sym	16.7	15.1	16.5	16.8	15.6	16.4	12.7	13.4	16.6	16.9	15.3	16.2
2,000	6	.40	7	asym	17.2	15.2	16.4	16.7	15.6	16.3	13.1	13.0	17.2	17.0	14.9	16.9
2,000	6	.60	3	sym	17.7	15.6	16.7	17.3	15.7	16.6	10.9	12.2	18.1	17.2	15.9	16.8
2,000	6	.60	3	asym	18.8	15.6	16.1	16.9	15.3	16.1	10.4	12.2	17.9	16.6	15.3	16.6
2,000	6	.60	5	sym	16.9	15.6	16.8	16.7	15.6	16.7	9.7	11.8	18.2	16.8	15.8	16.5
2,000	6	.60	5	asym	18.2	16.0	16.6	16.4	15.5	16.6	9.6	11.6	18.8	16.4	15.2	17.0
2,000	6	.60	7	sym	17.4	15.0	16.7	16.9	15.0	16.6	9.7	11.2	18.3	16.9	15.1	16.5
2,000	6	.60	7	asym	18.6	15.5	17.2	16.7	15.0	17.2	9.5	11.1	18.8	16.3	14.9	17.0
2,000	6	.80	3	sym	19.2	15.4	16.9	18.1	15.4	17.0	8.8	10.4	22.6	17.7	16.4	17.8
2,000	6	.80	3	asym	20.7	16.4	16.7	16.5	15.6	16.7	7.8	10.0	21.7	16.6	15.9	17.1
2,000	6	.80	5	sym	17.4	15.7	16.4	16.7	15.7	16.4	7.1	8.7	23.2	16.6	15.4	17.2
2,000	6	.80	5	asym	20.9	16.1	16.4	16.7	15.3	16.5	6.8	8.6	21.3	16.3	15.0	16.3
2,000	6	.80	7	sym	18.1	15.0	17.0	17.1	15.0	17.0	6.7	8.4	21.6	16.6	15.4	16.3
2,000	6	.80	7	asym	21.1	16.2	16.1	16.7	15.4	16.2	6.6	8.4	22.1	16.3	15.4	16.8
2,000	12	.40	3	sym	73.9	71.6	72.7	73.7	71.5	72.8	58.2	63.9	67.7	73.0	71.7	71.2
2,000	12	.40	3	asym	75.8	73.2	75.2	72.8	71.6	74.2	57.8	62.7	70.4	72.6	70.5	73.9
2,000	12	.40	5	sym	69.8	70.6	70.7	69.4	70.5	70.7	53.9	61.3	68.3	70.2	70.6	72.0
2,000	12	.40	5	asym	75.8	72.7	73.2	72.6	71.3	72.4	55.3	61.6	68.3	71.3	70.6	72.3
2,000	12	.40	7	sym	73.7	72.0	72.7	73.3	72.0	73.0	55.1	61.7	68.8	72.6	71.2	72.9
2,000	12	.40	7	asym	75.6	74.7	73.9	72.8	73.0	73.0	55.0	62.8	69.0	72.3	72.2	72.8
2,000	12	.60	3	sym	73.3	70.6	73.1	71.7	70.3	73.0	44.9	54.9	71.4	71.2	70.7	73.7
2,000	12	.60	3	asym	81.1	73.0	76.1	72.5	69.0	73.1	44.9	53.2	71.3	71.8	68.7	73.9
2,000	12	.60	5	sym	73.4	70.3	71.4	71.7	69.9	71.2	41.9	52.3	70.5	72.1	70.8	72.5
2,000	12	.60	5	asym	80.5	74.4	75.1	71.9	70.3	72.6	42.5	53.7	69.1	72.4	71.4	71.8
2,000	12	.60	7	sym	75.1	72.8	72.9	73.1	72.0	72.5	40.7	53.8	70.0	72.2	72.5	72.0
2,000	12	.60	7	asym	81.6	76.0	73.4	73.4	71.8	70.8	42.0	54.1	69.3	72.8	72.4	71.6
2,000	12	.80	3	sym	76.7	73.8	75.9	72.6	72.2	74.4	35.5	45.9	86.3	70.8	71.4	75.4
2,000	12	.80	3	asym	91.0	80.6	78.5	72.5	72.1	72.5	33.6	45.0	83.0	71.4	71.6	73.9
2,000	12	.80	5	sym	77.3	72.6	72.2	73.8	71.3	71.2	31.0	41.1	90.7	73.2	72.4	74.7
2,000	12	.80	5	asym	90.7	79.1	77.2	71.8	70.7	71.1	30.2	41.6	92.7	71.8	71.9	76.9
2,000	12	.80	7	sym	77.4	72.4	73.0	72.7	70.6	71.5	28.7	38.5	89.8	71.5	70.5	73.7
2,000	12	.80	7	asym	91.2	79.3	80.0	72.1	71.1	73.4	28.8	40.7	83.7	71.0	72.2	71.9

Note. MLR = MLR (Yuan & Bentler, 2000). 1F = one-factor CFA. 2F = two-factor CFA: factor correlation either .30 (correlated) or .00 (uncorrelated).

Table A7: *95% Quantiles of χ^2/df Distributions Resulting from Correctly Specified Models*

sample size	number of indicators	loading magnitude	response options	distribution	ML			MLR			DWLS			WLSMV		
					1F	2F		1F	2F		1F	2F		1F	2F	
						.30	.00		.30	.00		.30	.00		.30	.00
200	6	.40	3	sym	1.818	1.731	1.927	2.110	2.512	1.927	1.467	1.492	1.977	1.791	1.722	1.925
200	6	.40	3	asym	1.897	1.733	1.849	2.177	2.658	1.848	1.480	1.626	1.973	1.797	1.834	1.923
200	6	.40	5	sym	1.928	1.763	1.934	2.157	2.548	1.940	1.516	1.628	2.109	1.916	1.842	2.016
200	6	.40	5	asym	1.966	1.842	1.933	2.156	2.634	1.976	1.535	1.570	2.070	1.951	1.856	2.016
200	6	.40	7	sym	2.031	1.745	1.937	2.206	2.413	1.949	1.584	1.636	2.122	2.014	1.860	2.041
200	6	.40	7	asym	2.025	1.764	1.927	2.335	2.361	1.961	1.542	1.578	2.078	1.966	1.771	2.001
200	6	.60	3	sym	1.935	1.960	1.894	1.909	2.144	1.919	1.159	1.564	2.027	1.772	1.953	1.846
200	6	.60	3	asym	2.132	1.981	1.925	1.987	2.056	1.953	1.209	1.510	2.115	1.870	1.917	1.935
200	6	.60	5	sym	2.037	2.032	1.882	2.083	2.156	1.882	1.157	1.606	2.192	1.918	2.087	1.960
200	6	.60	5	asym	2.266	2.153	1.951	2.147	2.303	1.976	1.281	1.634	2.226	1.991	2.105	1.997
200	6	.60	7	sym	1.962	1.958	1.885	1.988	2.059	1.922	1.133	1.433	2.349	1.913	1.874	2.082
200	6	.60	7	asym	2.237	2.014	1.860	2.111	2.055	1.894	1.213	1.569	2.271	1.953	2.004	2.043
200	6	.80	3	sym	1.985	2.012	1.853	1.895	2.041	1.907	0.926	1.281	2.320	1.722	1.980	1.857
200	6	.80	3	asym	2.430	2.168	1.927	1.992	2.073	1.906	0.921	1.241	2.512	1.816	1.948	1.971
200	6	.80	5	sym	2.017	2.081	1.864	1.979	2.101	1.846	0.867	1.201	2.687	1.865	2.122	1.990
200	6	.80	5	asym	2.468	2.052	1.965	2.026	2.011	1.998	0.864	1.197	2.553	1.883	2.008	1.938
200	6	.80	7	sym	2.101	1.996	1.985	2.016	2.032	1.982	0.817	1.112	3.000	1.835	2.016	2.174
200	6	.80	7	asym	2.454	2.027	1.898	1.971	2.008	1.934	0.806	1.171	2.749	1.847	1.984	2.026
200	12	.40	3	sym	1.431	1.373	1.381	1.453	1.430	1.440	1.140	1.218	1.313	1.359	1.324	1.332
200	12	.40	3	asym	1.404	1.382	1.387	1.390	1.431	1.434	1.104	1.206	1.300	1.319	1.320	1.324
200	12	.40	5	sym	1.387	1.368	1.397	1.415	1.427	1.427	1.084	1.218	1.367	1.343	1.351	1.385
200	12	.40	5	asym	1.420	1.395	1.364	1.389	1.425	1.397	1.092	1.242	1.336	1.330	1.361	1.362
200	12	.40	7	sym	1.437	1.411	1.413	1.462	1.451	1.447	1.137	1.285	1.406	1.409	1.413	1.420
200	12	.40	7	asym	1.470	1.418	1.392	1.436	1.453	1.437	1.123	1.266	1.381	1.403	1.405	1.401
200	12	.60	3	sym	1.394	1.361	1.390	1.387	1.374	1.394	0.858	1.053	1.361	1.292	1.309	1.357
200	12	.60	3	asym	1.544	1.480	1.430	1.409	1.424	1.404	0.873	1.086	1.382	1.318	1.348	1.395
200	12	.60	5	sym	1.406	1.385	1.362	1.387	1.391	1.374	0.811	1.063	1.382	1.317	1.370	1.379
200	12	.60	5	asym	1.551	1.478	1.441	1.406	1.421	1.415	0.818	1.098	1.389	1.305	1.387	1.391
200	12	.60	7	sym	1.437	1.428	1.423	1.404	1.431	1.422	0.816	1.080	1.483	1.352	1.390	1.447
200	12	.60	7	asym	1.557	1.484	1.430	1.415	1.438	1.395	0.844	1.103	1.458	1.358	1.405	1.429
200	12	.80	3	sym	1.486	1.406	1.426	1.400	1.391	1.421	0.699	0.885	1.629	1.293	1.330	1.390
200	12	.80	3	asym	1.776	1.546	1.514	1.425	1.402	1.421	0.660	0.867	1.607	1.288	1.324	1.403
200	12	.80	5	sym	1.442	1.402	1.412	1.397	1.394	1.404	0.600	0.809	1.709	1.280	1.342	1.402
200	12	.80	5	asym	1.728	1.532	1.482	1.414	1.399	1.401	0.584	0.821	1.689	1.283	1.351	1.404
200	12	.80	7	sym	1.503	1.458	1.436	1.413	1.438	1.416	0.575	0.840	1.967	1.292	1.397	1.489
200	12	.80	7	asym	1.795	1.585	1.511	1.470	1.435	1.429	0.583	0.837	1.881	1.313	1.408	1.470
500	6	.40	3	sym	1.917	1.931	1.941	2.005	2.350	1.946	1.511	1.733	2.007	1.884	1.942	1.969
500	6	.40	3	asym	1.955	1.923	1.931	1.979	2.478	1.931	1.515	1.708	1.970	1.897	1.918	1.935
500	6	.40	5	sym	1.863	1.902	1.894	1.926	2.185	1.904	1.437	1.704	1.959	1.867	1.938	1.910
500	6	.40	5	asym	2.041	1.949	1.888	2.074	2.159	1.887	1.491	1.712	1.955	1.883	1.990	1.917
500	6	.40	7	sym	1.771	1.935	1.940	1.816	2.159	1.958	1.372	1.739	2.056	1.825	1.950	1.999
500	6	.40	7	asym	1.935	1.970	1.958	1.925	2.172	1.955	1.448	1.716	2.054	1.881	1.922	2.006
500	6	.60	3	sym	1.983	1.957	1.814	1.977	1.988	1.805	1.186	1.528	2.040	1.916	1.991	1.885
500	6	.60	3	asym	2.112	1.993	1.950	1.917	2.009	1.940	1.174	1.536	2.111	1.849	1.949	1.951
500	6	.60	5	sym	1.888	2.023	1.888	1.867	2.038	1.917	1.100	1.571	2.227	1.877	2.100	2.000
500	6	.60	5	asym	2.049	2.124	1.954	1.864	2.120	1.947	1.128	1.567	2.173	1.889	2.060	1.963
500	6	.60	7	sym	1.957	1.969	1.928	1.955	1.979	1.914	1.092	1.462	2.172	1.939	1.985	1.934
500	6	.60	7	asym	2.129	2.049	1.884	1.915	2.042	1.901	1.108	1.535	2.193	1.921	2.026	1.996
500	6	.80	3	sym	2.010	1.970	1.857	1.930	1.973	1.862	0.943	1.234	2.552	1.866	1.940	2.017
500	6	.80	3	asym	2.424	2.025	1.862	1.895	1.947	1.858	0.879	1.186	2.549	1.819	1.960	2.012
500	6	.80	5	sym	1.989	2.043	1.869	1.949	2.008	1.879	0.803	1.167	2.529	1.840	2.022	1.915
500	6	.80	5	asym	2.403	2.114	1.938	1.897	2.017	1.929	0.766	1.135	2.287	1.802	1.959	1.780
500	6	.80	7	sym	2.052	1.979	1.840	1.957	1.986	1.854	0.790	1.082	2.531	1.872	1.993	1.899
500	6	.80	7	asym	2.500	2.069	1.954	1.996	2.003	1.936	0.784	1.122	2.535	1.867	2.029	1.926
500	12	.40	3	sym	1.359	1.361	1.358	1.363	1.388	1.375	1.069	1.201	1.298	1.319	1.338	1.348
500	12	.40	3	asym	1.400	1.376	1.391	1.353	1.368	1.387	1.057	1.216	1.345	1.317	1.341	1.390
500	12	.40	5	sym	1.366	1.364	1.383	1.373	1.377	1.392	1.058	1.185	1.322	1.343	1.356	1.376
500	12	.40	5	asym	1.407	1.380	1.386	1.363	1.381	1.388	1.068	1.187	1.326	1.351	1.348	1.380
500	12	.40	7	sym	1.353	1.380	1.356	1.353	1.393	1.375	1.034	1.204	1.290	1.344	1.377	1.350
500	12	.40	7	asym	1.433	1.412	1.381	1.393	1.411	1.379	1.067	1.200	1.296	1.355	1.367	1.353
500	12	.60	3	sym	1.379	1.345	1.356	1.364	1.342	1.361	0.862	1.055	1.347	1.337	1.345	1.378
500	12	.60	3	asym	1.520	1.425	1.388	1.370	1.355	1.350	0.864	1.055	1.352	1.342	1.337	1.392
500	12	.60	5	sym	1.377	1.389	1.364	1.358	1.379	1.358	0.778	1.038	1.357	1.321	1.375	1.374
500	12	.60	5	asym	1.512	1.440	1.445	1.371	1.366	1.401	0.805	1.039	1.344	1.331	1.356	1.370
500	12	.60	7	sym	1.406	1.371	1.384	1.379	1.365	1.382	0.792	1.043	1.379	1.371	1.398	1.380
500	12	.60	7	asym	1.509	1.420	1.415	1.359	1.345	1.374	0.801	1.021	1.323	1.344	1.350	1.352
500	12	.80	3	sym	1.443	1.374	1.382	1.369	1.358	1.359	0.684	0.903	1.633	1.329	1.381	1.402
500	12	.80	3	asym	1.718	1.518	1.462	1.348	1.366	1.351	0.632	0.871	1.661	1.308	1.362	1.443
500	12	.80	5	sym	1.403	1.379	1.364	1.343	1.362	1.348	0.569	0.788	1.734	1.296	1.358	1.411
500	12	.80	5	asym	1.688	1.520	1.508	1.355	1.373	1.403	0.564	0.789	1.664	1.297	1.351	1.387
500	12	.80	7	sym	1.448	1.369	1.363	1.375	1.341	1.341	0.549	0.771	1.663	1.324	1.369	1.368

sample size	number of indicators	loading magnitude	response options	distribution	ML			MLR			DWLS			WLSMV		
					1F	2F		1F	2F		1F	2F		1F	2F	
						.30	.00		.30	.00		.30	.00		.30	.00
500	12	.80	7	asym	1.752	1.533	1.470	1.401	1.376	1.368	0.556	0.790	1.647	1.316	1.377	1.375
2,000	6	.40	3	sym	1.913	1.852	1.960	1.918	1.907	1.957	1.491	1.746	1.962	1.874	1.924	1.931
2,000	6	.40	3	asym	1.900	1.911	1.850	1.847	1.962	1.854	1.455	1.711	1.902	1.832	1.921	1.873
2,000	6	.40	5	sym	1.857	1.905	1.851	1.881	1.970	1.847	1.407	1.743	1.890	1.852	1.979	1.845
2,000	6	.40	5	asym	1.891	1.932	1.875	1.835	1.959	1.880	1.412	1.715	1.947	1.820	1.951	1.907
2,000	6	.40	7	sym	1.853	1.891	1.831	1.863	1.946	1.820	1.410	1.677	1.839	1.881	1.909	1.799
2,000	6	.40	7	asym	1.908	1.906	1.823	1.854	1.946	1.816	1.451	1.623	1.914	1.891	1.858	1.878
2,000	6	.60	3	sym	1.963	1.953	1.852	1.927	1.959	1.846	1.209	1.525	2.008	1.912	1.984	1.863
2,000	6	.60	3	asym	2.085	1.952	1.794	1.875	1.913	1.786	1.157	1.520	1.994	1.841	1.913	1.850
2,000	6	.60	5	sym	1.877	1.956	1.864	1.856	1.950	1.859	1.077	1.475	2.027	1.863	1.971	1.837
2,000	6	.60	5	asym	2.024	1.995	1.840	1.818	1.932	1.845	1.062	1.453	2.088	1.819	1.898	1.893
2,000	6	.60	7	sym	1.938	1.873	1.856	1.873	1.873	1.848	1.074	1.403	2.038	1.873	1.890	1.834
2,000	6	.60	7	asym	2.067	1.934	1.909	1.859	1.878	1.908	1.052	1.390	2.085	1.809	1.860	1.885
2,000	6	.80	3	sym	2.132	1.927	1.883	2.009	1.920	1.889	0.978	1.296	2.513	1.965	2.052	1.977
2,000	6	.80	3	asym	2.304	2.044	1.859	1.836	1.955	1.855	0.866	1.247	2.406	1.849	1.990	1.896
2,000	6	.80	5	sym	1.935	1.964	1.827	1.852	1.958	1.826	0.788	1.090	2.572	1.847	1.927	1.910
2,000	6	.80	5	asym	2.326	2.008	1.822	1.852	1.915	1.833	0.760	1.070	2.372	1.808	1.877	1.814
2,000	6	.80	7	sym	2.013	1.881	1.892	1.897	1.879	1.893	0.739	1.045	2.395	1.843	1.929	1.813
2,000	6	.80	7	asym	2.340	2.022	1.789	1.857	1.930	1.797	0.738	1.052	2.460	1.806	1.926	1.867
2,000	12	.40	3	sym	1.368	1.351	1.346	1.365	1.349	1.349	1.078	1.206	1.254	1.351	1.352	1.319
2,000	12	.40	3	asym	1.405	1.381	1.392	1.349	1.351	1.374	1.071	1.184	1.304	1.345	1.331	1.368
2,000	12	.40	5	sym	1.292	1.332	1.309	1.285	1.330	1.309	0.999	1.157	1.264	1.300	1.332	1.334
2,000	12	.40	5	asym	1.404	1.371	1.356	1.345	1.346	1.341	1.024	1.162	1.265	1.320	1.332	1.339
2,000	12	.40	7	sym	1.365	1.358	1.347	1.357	1.358	1.352	1.020	1.164	1.273	1.344	1.343	1.349
2,000	12	.40	7	asym	1.399	1.409	1.369	1.347	1.378	1.352	1.019	1.185	1.277	1.340	1.362	1.349
2,000	12	.60	3	sym	1.357	1.332	1.354	1.329	1.327	1.352	0.831	1.037	1.323	1.318	1.334	1.365
2,000	12	.60	3	asym	1.501	1.378	1.409	1.342	1.301	1.354	0.831	1.003	1.320	1.329	1.297	1.369
2,000	12	.60	5	sym	1.359	1.326	1.322	1.328	1.320	1.318	0.775	0.987	1.305	1.335	1.336	1.343
2,000	12	.60	5	asym	1.491	1.405	1.391	1.332	1.326	1.344	0.787	1.013	1.279	1.341	1.347	1.330
2,000	12	.60	7	sym	1.391	1.374	1.350	1.355	1.359	1.342	0.753	1.014	1.297	1.337	1.368	1.334
2,000	12	.60	7	asym	1.511	1.435	1.358	1.359	1.356	1.311	0.778	1.021	1.283	1.347	1.366	1.326
2,000	12	.80	3	sym	1.421	1.392	1.406	1.345	1.363	1.377	0.658	0.866	1.598	1.311	1.348	1.397
2,000	12	.80	3	asym	1.684	1.521	1.454	1.342	1.360	1.342	0.623	0.849	1.538	1.322	1.350	1.368
2,000	12	.80	5	sym	1.431	1.371	1.337	1.367	1.345	1.319	0.574	0.775	1.679	1.355	1.365	1.382
2,000	12	.80	5	asym	1.680	1.492	1.429	1.330	1.335	1.317	0.559	0.786	1.717	1.329	1.356	1.425
2,000	12	.80	7	sym	1.433	1.366	1.351	1.346	1.333	1.324	0.532	0.726	1.662	1.324	1.330	1.365
2,000	12	.80	7	asym	1.688	1.497	1.481	1.335	1.342	1.358	0.533	0.768	1.549	1.315	1.362	1.332

Note. MLR = MLR (Yuan & Bentler, 2000). 1F = one-factor CFA. 2F = two-factor CFA: factor correlation either .30 (correlated) or .00 (uncorrelated).

Table A8: *5% Quantiles of CFI Distributions Resulting from Correctly Specified Models*

sample size	number of indicators	loading magnitude	response options	distribution	ML			MLR			DWLS			WLSMV		
					1F	2F		1F	2F		1F	2F		1F	2F	
						.30	.00		.30	.00		.30	.00		.30	.00
200	6	.40	3	sym	.814	.785	.649	.746	.438	.641	.914	.866	.668	.847	.789	.663
200	6	.40	3	asym	.803	.769	.631	.690	.361	.631	.912	.835	.634	.836	.774	.621
200	6	.40	5	sym	.830	.810	.695	.771	.548	.702	.926	.878	.714	.856	.807	.705
200	6	.40	5	asym	.812	.779	.648	.737	.452	.641	.922	.872	.644	.846	.798	.635
200	6	.40	7	sym	.839	.834	.716	.797	.630	.723	.932	.891	.710	.870	.823	.704
200	6	.40	7	asym	.813	.803	.699	.743	.612	.692	.920	.876	.719	.852	.809	.713
200	6	.60	3	sym	.952	.919	.917	.948	.904	.915	.996	.967	.926	.972	.935	.927
200	6	.60	3	asym	.938	.924	.910	.938	.903	.898	.994	.969	.915	.967	.937	.915
200	6	.60	5	sym	.955	.932	.934	.953	.923	.934	.996	.974	.942	.972	.947	.943
200	6	.60	5	asym	.937	.922	.926	.938	.904	.924	.994	.970	.932	.967	.939	.933
200	6	.60	7	sym	.956	.943	.937	.955	.936	.934	.997	.984	.939	.976	.957	.940
200	6	.60	7	asym	.942	.932	.935	.942	.922	.927	.995	.977	.935	.972	.949	.936
200	6	.80	3	sym	.982	.974	.974	.982	.972	.972	1	.997	.983	.995	.985	.985
200	6	.80	3	asym	.973	.971	.973	.976	.968	.970	1	.997	.980	.993	.985	.982
200	6	.80	5	sym	.985	.980	.981	.985	.979	.980	1	.998	.986	.996	.988	.988
200	6	.80	5	asym	.977	.979	.978	.980	.977	.974	1	.998	.986	.995	.988	.987
200	6	.80	7	sym	.985	.982	.981	.985	.981	.980	1	.999	.986	.996	.990	.987
200	6	.80	7	asym	.979	.979	.980	.982	.977	.976	1	.999	.985	.995	.989	.987
200	12	.40	3	sym	.828	.789	.779	.815	.752	.739	.966	.910	.862	.886	.830	.813
200	12	.40	3	asym	.832	.775	.755	.829	.724	.711	.977	.907	.836	.897	.818	.785
200	12	.40	5	sym	.874	.829	.807	.867	.803	.789	.985	.932	.873	.920	.855	.828
200	12	.40	5	asym	.852	.799	.796	.853	.767	.776	.983	.915	.869	.910	.839	.822
200	12	.40	7	sym	.865	.820	.816	.860	.799	.799	.977	.920	.870	.910	.847	.827
200	12	.40	7	asym	.843	.788	.782	.844	.753	.750	.980	.908	.861	.909	.829	.815
200	12	.60	3	sym	.955	.944	.938	.956	.941	.936	1	.996	.972	.981	.966	.957
200	12	.60	3	asym	.937	.925	.929	.947	.927	.930	1	.994	.967	.978	.958	.950
200	12	.60	5	sym	.964	.955	.954	.964	.953	.952	1	.997	.977	.984	.968	.965
200	12	.60	5	asym	.948	.936	.939	.955	.939	.938	1	.993	.973	.982	.963	.958
200	12	.60	7	sym	.963	.951	.951	.963	.949	.950	1	.995	.974	.983	.967	.962
200	12	.60	7	asym	.947	.939	.944	.955	.941	.944	1	.993	.972	.980	.962	.959
200	12	.80	3	sym	.980	.978	.976	.982	.979	.976	1	1	.991	.996	.990	.989
200	12	.80	3	asym	.968	.971	.973	.977	.976	.974	1	1	.990	.995	.990	.987
200	12	.80	5	sym	.985	.983	.983	.986	.983	.982	1	1	.993	.997	.993	.991
200	12	.80	5	asym	.974	.977	.978	.981	.980	.980	1	1	.993	.996	.992	.990
200	12	.80	7	sym	.984	.982	.982	.986	.982	.982	1	1	.992	.996	.992	.989
200	12	.80	7	asym	.973	.976	.978	.980	.979	.979	1	1	.992	.996	.991	.989
500	6	.40	3	sym	.916	.865	.831	.907	.811	.828	.964	.908	.839	.934	.871	.838
500	6	.40	3	asym	.902	.858	.821	.894	.763	.814	.957	.903	.827	.926	.872	.825
500	6	.40	5	sym	.935	.903	.878	.931	.871	.878	.977	.932	.885	.950	.905	.884
500	6	.40	5	asym	.921	.887	.859	.917	.844	.857	.973	.918	.869	.946	.892	.867
500	6	.40	7	sym	.944	.895	.869	.941	.868	.869	.981	.931	.878	.957	.906	.877
500	6	.40	7	asym	.930	.877	.859	.924	.825	.852	.976	.921	.861	.948	.891	.859
500	6	.60	3	sym	.979	.969	.967	.979	.968	.967	.998	.988	.970	.988	.974	.971
500	6	.60	3	asym	.975	.965	.962	.977	.963	.960	.998	.987	.968	.987	.974	.968
500	6	.60	5	sym	.985	.974	.972	.984	.974	.972	.999	.991	.975	.990	.978	.975
500	6	.60	5	asym	.979	.969	.968	.981	.966	.966	.999	.989	.973	.989	.978	.974
500	6	.60	7	sym	.984	.978	.975	.984	.976	.974	.999	.993	.979	.990	.983	.980
500	6	.60	7	asym	.979	.974	.971	.980	.972	.969	.999	.991	.974	.990	.979	.975
500	6	.80	3	sym	.992	.990	.990	.993	.990	.989	1	.999	.992	.998	.995	.993
500	6	.80	3	asym	.990	.990	.990	.992	.989	.989	1	.999	.992	.997	.994	.993
500	6	.80	5	sym	.994	.992	.992	.994	.992	.992	1	.999	.995	.998	.995	.996
500	6	.80	5	asym	.991	.991	.991	.993	.990	.990	1	1	.995	.998	.995	.996
500	6	.80	7	sym	.994	.993	.993	.994	.993	.993	1	1	.995	.998	.996	.996
500	6	.80	7	asym	.991	.992	.992	.992	.991	.991	1	1	.995	.998	.995	.995
500	12	.40	3	sym	.940	.906	.906	.939	.902	.901	.993	.961	.942	.962	.927	.922
500	12	.40	3	asym	.930	.897	.888	.936	.894	.884	.995	.958	.925	.962	.922	.902
500	12	.40	5	sym	.950	.928	.922	.950	.924	.919	.996	.974	.951	.968	.944	.934
500	12	.40	5	asym	.939	.915	.909	.942	.912	.907	.995	.970	.946	.966	.938	.926
500	12	.40	7	sym	.954	.929	.926	.953	.927	.923	.998	.974	.958	.971	.943	.940
500	12	.40	7	asym	.939	.911	.913	.942	.911	.909	.995	.971	.954	.967	.939	.936
500	12	.60	3	sym	.982	.979	.977	.983	.978	.977	1	.998	.988	.991	.985	.982
500	12	.60	3	asym	.975	.972	.973	.979	.975	.974	1	.998	.987	.990	.984	.980

sample size	number of indicators	loading magnitude	response options	distribution	ML			MLR			DWLS			WLSMV		
					1F	2F		1F	2F		1F	2F		1F	2F	
						.30	.00		.30	.00		.30	.00		.30	.00
500	12	.60	5	sym	.986	.981	.982	.987	.981	.981	1	.999	.991	.993	.987	.987
500	12	.60	5	asym	.978	.976	.976	.982	.979	.976	1	.999	.990	.992	.986	.985
500	12	.60	7	sym	.986	.982	.981	.986	.982	.981	1	.999	.991	.993	.987	.987
500	12	.60	7	asym	.980	.978	.977	.984	.980	.979	1	.999	.991	.993	.988	.986
500	12	.80	3	sym	.992	.992	.991	.993	.992	.992	1	1	.996	.998	.996	.995
500	12	.80	3	asym	.988	.989	.990	.992	.991	.991	1	1	.996	.998	.996	.994
500	12	.80	5	sym	.994	.994	.994	.995	.994	.994	1	1	.997	.999	.997	.996
500	12	.80	5	asym	.990	.991	.991	.994	.992	.992	1	1	.997	.998	.997	.996
500	12	.80	7	sym	.994	.994	.994	.995	.994	.994	1	1	.998	.999	.997	.997
500	12	.80	7	asym	.990	.991	.992	.993	.993	.993	1	1	.997	.998	.997	.996
2,000	6	.40	3	sym	.978	.965	.954	.978	.963	.954	.991	.976	.960	.983	.967	.960
2,000	6	.40	3	asym	.976	.960	.952	.976	.958	.951	.990	.973	.956	.982	.964	.956
2,000	6	.40	5	sym	.984	.972	.969	.984	.970	.969	.995	.981	.970	.988	.974	.971
2,000	6	.40	5	asym	.981	.970	.961	.981	.968	.961	.994	.980	.966	.987	.973	.966
2,000	6	.40	7	sym	.985	.975	.971	.984	.974	.971	.995	.983	.972	.988	.976	.972
2,000	6	.40	7	asym	.981	.968	.966	.981	.967	.965	.994	.982	.970	.987	.975	.970
2,000	6	.60	3	sym	.995	.992	.991	.995	.992	.991	.999	.997	.993	.997	.994	.993
2,000	6	.60	3	asym	.994	.991	.991	.994	.991	.991	1	.996	.992	.997	.993	.993
2,000	6	.60	5	sym	.996	.994	.994	.996	.994	.994	1	.998	.995	.998	.995	.995
2,000	6	.60	5	asym	.995	.993	.993	.995	.993	.992	1	.998	.994	.998	.995	.994
2,000	6	.60	7	sym	.996	.995	.994	.996	.995	.994	1	.998	.995	.998	.996	.995
2,000	6	.60	7	asym	.995	.994	.993	.995	.994	.992	1	.998	.994	.998	.996	.994
2,000	6	.80	3	sym	.998	.998	.997	.998	.997	.997	1	1	.998	.999	.998	.998
2,000	6	.80	3	asym	.998	.997	.997	.998	.997	.997	1	1	.998	.999	.998	.998
2,000	6	.80	5	sym	.999	.998	.998	.999	.998	.998	1	1	.999	1	.999	.999
2,000	6	.80	5	asym	.998	.998	.998	.998	.998	.998	1	1	.999	.999	.999	.999
2,000	6	.80	7	sym	.999	.998	.998	.999	.998	.998	1	1	.999	1	.999	.999
2,000	6	.80	7	asym	.998	.998	.998	.998	.998	.998	1	1	.999	1	.999	.999
2,000	12	.40	3	sym	.985	.977	.975	.985	.977	.975	.998	.991	.986	.990	.982	.981
2,000	12	.40	3	asym	.982	.975	.971	.984	.976	.972	.998	.990	.984	.990	.981	.978
2,000	12	.40	5	sym	.990	.983	.983	.990	.983	.983	1	.995	.990	.994	.987	.985
2,000	12	.40	5	asym	.985	.978	.978	.986	.979	.978	1	.993	.988	.992	.985	.983
2,000	12	.40	7	sym	.988	.982	.982	.988	.982	.982	1	.994	.990	.993	.987	.986
2,000	12	.40	7	asym	.985	.978	.978	.987	.979	.979	1	.993	.988	.993	.985	.984
2,000	12	.60	3	sym	.996	.995	.994	.996	.995	.994	1	1	.997	.998	.996	.996
2,000	12	.60	3	asym	.994	.994	.993	.995	.995	.993	1	1	.997	.998	.997	.995
2,000	12	.60	5	sym	.997	.996	.996	.997	.996	.996	1	1	.998	.998	.997	.997
2,000	12	.60	5	asym	.995	.994	.994	.996	.995	.995	1	1	.998	.998	.997	.997
2,000	12	.60	7	sym	.997	.996	.996	.997	.996	.996	1	1	.998	.998	.997	.997
2,000	12	.60	7	asym	.995	.994	.995	.996	.995	.996	1	1	.998	.998	.997	.997
2,000	12	.80	3	sym	.998	.998	.998	.998	.998	.998	1	1	.999	1	.999	.999
2,000	12	.80	3	asym	.997	.997	.997	.998	.998	.998	1	1	.999	.999	.999	.999
2,000	12	.80	5	sym	.999	.998	.999	.999	.999	.999	1	1	.999	1	.999	.999
2,000	12	.80	5	asym	.998	.998	.998	.998	.998	.998	1	1	.999	1	.999	.999
2,000	12	.80	7	sym	.999	.999	.999	.999	.999	.999	1	1	.999	1	.999	.999
2,000	12	.80	7	asym	.998	.998	.998	.999	.998	.998	1	1	.999	1	.999	.999

Note. MLR = MLR (Yuan & Bentler, 2000). 1F = one-factor CFA. 2F = two-factor CFA: factor correlation either .30 (correlated) or .00 (uncorrelated).

Table A9: *95% Quantiles of RMSEA Distributions Resulting from Correctly Specified Models*

sample size	number of indicators	loading magnitude	response options	distribution	ML			MLR			DWLS			WLSMV		
					1F	2F		1F	2F		1F	2F		1F	2F	
						.30	.00		.30	.00		.30	.00		.30	.00
200	6	.40	3	sym	.064	.060	.068	.074	.087	.068	.048	.050	.070	.063	.060	.068
200	6	.40	3	asym	.067	.061	.065	.077	.091	.065	.049	.056	.070	.063	.065	.068
200	6	.40	5	sym	.068	.062	.068	.076	.088	.069	.051	.056	.075	.068	.065	.071
200	6	.40	5	asym	.070	.065	.068	.076	.090	.070	.052	.054	.073	.069	.066	.071
200	6	.40	7	sym	.072	.061	.068	.078	.084	.069	.054	.057	.075	.071	.066	.072
200	6	.40	7	asym	.072	.062	.068	.082	.082	.069	.052	.054	.074	.070	.062	.071
200	6	.60	3	sym	.068	.069	.067	.067	.076	.068	.028	.053	.072	.062	.069	.065
200	6	.60	3	asym	.075	.070	.068	.070	.073	.069	.032	.051	.075	.066	.068	.069
200	6	.60	5	sym	.072	.072	.066	.074	.076	.066	.028	.055	.077	.068	.074	.069
200	6	.60	5	asym	.080	.076	.069	.076	.081	.070	.038	.056	.078	.071	.075	.071
200	6	.60	7	sym	.069	.069	.067	.070	.073	.068	.026	.047	.082	.068	.066	.074
200	6	.60	7	asym	.079	.071	.066	.075	.073	.067	.033	.053	.080	.069	.071	.072
200	6	.80	3	sym	.070	.071	.065	.067	.072	.067	0	.038	.081	.060	.070	.066
200	6	.80	3	asym	.085	.076	.068	.070	.073	.067	0	.035	.087	.064	.069	.070
200	6	.80	5	sym	.071	.074	.066	.070	.074	.065	0	.032	.092	.066	.075	.071
200	6	.80	5	asym	.086	.073	.069	.072	.071	.071	0	.031	.088	.067	.071	.069
200	6	.80	7	sym	.074	.071	.070	.071	.072	.070	0	.024	.100	.065	.071	.077
200	6	.80	7	asym	.085	.072	.067	.070	.071	.068	0	.029	.094	.065	.070	.072
200	12	.40	3	sym	.046	.043	.044	.048	.046	.047	.027	.033	.040	.042	.040	.041
200	12	.40	3	asym	.045	.044	.044	.044	.046	.047	.023	.032	.039	.040	.040	.040
200	12	.40	5	sym	.044	.043	.045	.046	.046	.046	.020	.033	.043	.041	.042	.044
200	12	.40	5	asym	.046	.044	.043	.044	.046	.045	.022	.035	.041	.041	.043	.043
200	12	.40	7	sym	.047	.045	.045	.048	.047	.047	.026	.038	.045	.045	.046	.046
200	12	.40	7	asym	.048	.046	.044	.047	.048	.047	.025	.037	.044	.045	.045	.045
200	12	.60	3	sym	.044	.042	.044	.044	.043	.044	0	.016	.043	.038	.039	.042
200	12	.60	3	asym	.052	.049	.046	.045	.046	.045	0	.021	.044	.040	.042	.045
200	12	.60	5	sym	.045	.044	.043	.044	.044	.043	0	.018	.044	.040	.043	.044
200	12	.60	5	asym	.052	.049	.047	.045	.046	.046	0	.022	.044	.039	.044	.044
200	12	.60	7	sym	.047	.046	.046	.045	.046	.046	0	.020	.049	.042	.044	.047
200	12	.60	7	asym	.053	.049	.046	.046	.047	.044	0	.023	.048	.042	.045	.046
200	12	.80	3	sym	.049	.045	.046	.045	.044	.046	0	0	.056	.038	.041	.044
200	12	.80	3	asym	.062	.052	.051	.046	.045	.046	0	0	.055	.038	.040	.045
200	12	.80	5	sym	.047	.045	.045	.045	.044	.045	0	0	.060	.038	.041	.045
200	12	.80	5	asym	.060	.052	.049	.046	.045	.045	0	0	.059	.038	.042	.045
200	12	.80	7	sym	.050	.048	.047	.045	.047	.046	0	0	.070	.038	.045	.050
200	12	.80	7	asym	.063	.054	.051	.048	.047	.046	0	0	.067	.040	.045	.049
500	6	.40	3	sym	.043	.043	.043	.045	.052	.044	.032	.038	.045	.042	.043	.044
500	6	.40	3	asym	.044	.043	.043	.044	.054	.043	.032	.038	.044	.042	.043	.043
500	6	.40	5	sym	.042	.042	.042	.043	.049	.043	.030	.038	.044	.042	.043	.043
500	6	.40	5	asym	.046	.044	.042	.046	.048	.042	.031	.038	.044	.042	.045	.043
500	6	.40	7	sym	.039	.043	.043	.040	.048	.044	.027	.038	.046	.041	.044	.045
500	6	.40	7	asym	.043	.044	.044	.043	.048	.044	.030	.038	.046	.042	.043	.045
500	6	.60	3	sym	.044	.044	.040	.044	.044	.040	.019	.033	.046	.043	.045	.042
500	6	.60	3	asym	.047	.045	.044	.043	.045	.043	.019	.033	.047	.041	.044	.044
500	6	.60	5	sym	.042	.045	.042	.042	.046	.043	.014	.034	.050	.042	.047	.045
500	6	.60	5	asym	.046	.047	.044	.042	.047	.044	.016	.034	.048	.042	.046	.044
500	6	.60	7	sym	.044	.044	.043	.044	.044	.043	.014	.030	.048	.043	.044	.043
500	6	.60	7	asym	.048	.046	.042	.043	.046	.042	.015	.033	.049	.043	.045	.045
500	6	.80	3	sym	.045	.044	.041	.043	.044	.042	0	.022	.056	.042	.043	.045
500	6	.80	3	asym	.053	.045	.042	.042	.044	.041	0	.019	.056	.041	.044	.045
500	6	.80	5	sym	.044	.046	.042	.044	.045	.042	0	.018	.055	.041	.045	.043
500	6	.80	5	asym	.053	.047	.043	.042	.045	.043	0	.016	.051	.040	.044	.040
500	6	.80	7	sym	.046	.044	.041	.044	.044	.041	0	.013	.055	.042	.045	.042
500	6	.80	7	asym	.055	.046	.044	.045	.045	.043	0	.016	.055	.042	.045	.043
500	12	.40	3	sym	.027	.027	.027	.027	.028	.027	.012	.020	.024	.025	.026	.026
500	12	.40	3	asym	.028	.027	.028	.027	.027	.028	.011	.021	.026	.025	.026	.028
500	12	.40	5	sym	.027	.027	.028	.027	.027	.028	.011	.019	.025	.026	.027	.027
500	12	.40	5	asym	.029	.028	.028	.027	.028	.028	.012	.019	.026	.027	.026	.028
500	12	.40	7	sym	.027	.028	.027	.027	.028	.027	.008	.020	.024	.026	.027	.026
500	12	.40	7	asym	.029	.029	.028	.028	.029	.028	.012	.020	.024	.027	.027	.027
500	12	.60	3	sym	.028	.026	.027	.027	.026	.027	0	.011	.026	.026	.026	.028
500	12	.60	3	asym	.032	.029	.028	.027	.027	.026	0	.011	.027	.026	.026	.028

sample size	number of indicators	loading magnitude	response options	distribution	ML			MLR			DWLS			WLSMV		
					1F	2F		1F	2F		1F	2F		1F	2F	
						.30	.00		.30	.00		.30	.00		.30	.00
500	12	.60	5	sym	.027	.028	.027	.027	.028	.027	0	.009	.027	.025	.027	.027
500	12	.60	5	asym	.032	.030	.030	.027	.027	.028	0	.009	.026	.026	.027	.027
500	12	.60	7	sym	.029	.027	.028	.028	.027	.028	0	.009	.028	.027	.028	.028
500	12	.60	7	asym	.032	.029	.029	.027	.026	.027	0	.006	.025	.026	.026	.027
500	12	.80	3	sym	.030	.027	.028	.027	.027	.027	0	0	.036	.026	.028	.028
500	12	.80	3	asym	.038	.032	.030	.026	.027	.027	0	0	.036	.025	.027	.030
500	12	.80	5	sym	.028	.028	.027	.026	.027	.026	0	0	.038	.024	.027	.029
500	12	.80	5	asym	.037	.032	.032	.027	.027	.028	0	0	.036	.024	.027	.028
500	12	.80	7	sym	.030	.027	.027	.027	.026	.026	0	0	.036	.025	.027	.027
500	12	.80	7	asym	.039	.033	.031	.028	.027	.027	0	0	.036	.025	.027	.027
2,000	6	.40	3	sym	.021	.021	.022	.021	.021	.022	.016	.019	.022	.021	.021	.022
2,000	6	.40	3	asym	.021	.021	.021	.021	.022	.021	.015	.019	.021	.020	.021	.021
2,000	6	.40	5	sym	.021	.021	.021	.021	.022	.021	.014	.019	.021	.021	.022	.021
2,000	6	.40	5	asym	.021	.022	.021	.020	.022	.021	.014	.019	.022	.020	.022	.021
2,000	6	.40	7	sym	.021	.021	.020	.021	.022	.020	.014	.018	.020	.021	.021	.020
2,000	6	.40	7	asym	.021	.021	.020	.021	.022	.020	.015	.018	.021	.021	.021	.021
2,000	6	.60	3	sym	.022	.022	.021	.022	.022	.021	.010	.016	.022	.021	.022	.021
2,000	6	.60	3	asym	.023	.022	.020	.021	.021	.020	.009	.016	.022	.021	.021	.021
2,000	6	.60	5	sym	.021	.022	.021	.021	.022	.021	.006	.015	.023	.021	.022	.020
2,000	6	.60	5	asym	.023	.022	.020	.020	.022	.021	.006	.015	.023	.020	.021	.021
2,000	6	.60	7	sym	.022	.021	.021	.021	.021	.021	.006	.014	.023	.021	.021	.020
2,000	6	.60	7	asym	.023	.022	.021	.021	.021	.021	.005	.014	.023	.020	.021	.021
2,000	6	.80	3	sym	.024	.022	.021	.022	.021	.021	0	.012	.028	.022	.023	.022
2,000	6	.80	3	asym	.026	.023	.021	.020	.022	.021	0	.011	.027	.021	.022	.021
2,000	6	.80	5	sym	.022	.022	.020	.021	.022	.020	0	.007	.028	.021	.022	.021
2,000	6	.80	5	asym	.026	.022	.020	.021	.021	.020	0	.006	.026	.020	.021	.020
2,000	6	.80	7	sym	.023	.021	.021	.021	.021	.021	0	.005	.026	.021	.022	.020
2,000	6	.80	7	asym	.026	.023	.020	.021	.022	.020	0	.005	.027	.020	.022	.021
2,000	12	.40	3	sym	.014	.013	.013	.014	.013	.013	.006	.010	.011	.013	.013	.013
2,000	12	.40	3	asym	.014	.014	.014	.013	.013	.014	.006	.010	.012	.013	.013	.014
2,000	12	.40	5	sym	.012	.013	.012	.012	.013	.012	0	.009	.011	.012	.013	.013
2,000	12	.40	5	asym	.014	.014	.013	.013	.013	.013	.003	.009	.012	.013	.013	.013
2,000	12	.40	7	sym	.014	.013	.013	.013	.013	.013	.003	.009	.012	.013	.013	.013
2,000	12	.40	7	asym	.014	.014	.014	.013	.014	.013	.003	.010	.012	.013	.013	.013
2,000	12	.60	3	sym	.013	.013	.013	.013	.013	.013	0	.004	.013	.013	.013	.014
2,000	12	.60	3	asym	.016	.014	.014	.013	.012	.013	0	.001	.013	.013	.012	.014
2,000	12	.60	5	sym	.013	.013	.013	.013	.013	.013	0	0	.012	.013	.013	.013
2,000	12	.60	5	asym	.016	.014	.014	.013	.013	.013	0	.003	.012	.013	.013	.013
2,000	12	.60	7	sym	.014	.014	.013	.013	.013	.013	0	.003	.012	.013	.014	.013
2,000	12	.60	7	asym	.016	.015	.013	.013	.013	.012	0	.003	.012	.013	.014	.013
2,000	12	.80	3	sym	.015	.014	.014	.013	.013	.014	0	0	.017	.012	.013	.014
2,000	12	.80	3	asym	.018	.016	.015	.013	.013	.013	0	0	.016	.013	.013	.014
2,000	12	.80	5	sym	.015	.014	.013	.014	.013	.013	0	0	.018	.013	.014	.014
2,000	12	.80	5	asym	.018	.016	.015	.013	.013	.013	0	0	.019	.013	.013	.015
2,000	12	.80	7	sym	.015	.014	.013	.013	.013	.013	0	0	.018	.013	.013	.014
2,000	12	.80	7	asym	.019	.016	.016	.013	.013	.013	0	0	.017	.013	.013	.013

Note. MLR = MLR (Yuan & Bentler, 2000). 1F = one-factor CFA. 2F = two-factor CFA: factor correlation either .30 (correlated) or .00 (uncorrelated).

Table A10: *95% Quantiles of SRMR Distributions Resulting from Correctly Specified Models*

sample size	number of indicators	loading magnitude	response options	distribution	ML/MLR			DWLS/WLSMV		
					1F	2F		1F	2F	
						.30	.00		.30	.00
200	6	.40	3	sym	.048	.047	.056	.068	.066	.079
200	6	.40	3	asym	.049	.048	.056	.072	.071	.083
200	6	.40	5	sym	.048	.047	.056	.059	.059	.070
200	6	.40	5	asym	.049	.047	.056	.064	.061	.074
200	6	.40	7	sym	.049	.046	.056	.058	.057	.067
200	6	.40	7	asym	.050	.047	.057	.061	.059	.070
200	6	.60	3	sym	.040	.048	.057	.056	.067	.080
200	6	.60	3	asym	.043	.047	.058	.060	.068	.085
200	6	.60	5	sym	.039	.048	.057	.048	.057	.071
200	6	.60	5	asym	.043	.049	.058	.053	.061	.076
200	6	.60	7	sym	.039	.045	.059	.046	.053	.072
200	6	.60	7	asym	.042	.047	.059	.049	.057	.075
200	6	.80	3	sym	.028	.042	.060	.038	.058	.086
200	6	.80	3	asym	.031	.043	.063	.040	.059	.094
200	6	.80	5	sym	.024	.041	.064	.029	.047	.079
200	6	.80	5	asym	.028	.042	.064	.031	.051	.082
200	6	.80	7	sym	.023	.039	.067	.028	.045	.080
200	6	.80	7	asym	.027	.041	.064	.029	.048	.081
200	12	.40	3	sym	.057	.059	.061	.076	.078	.082
200	12	.40	3	asym	.057	.059	.061	.078	.082	.086
200	12	.40	5	sym	.054	.058	.062	.064	.067	.072
200	12	.40	5	asym	.056	.059	.061	.067	.073	.075
200	12	.40	7	sym	.055	.059	.062	.062	.066	.071
200	12	.40	7	asym	.057	.060	.062	.066	.070	.074
200	12	.60	3	sym	.046	.054	.061	.061	.072	.082
200	12	.60	3	asym	.048	.056	.062	.064	.076	.087
200	12	.60	5	sym	.043	.053	.061	.050	.061	.071
200	12	.60	5	asym	.047	.055	.062	.054	.066	.076
200	12	.60	7	sym	.043	.053	.063	.049	.059	.072
200	12	.60	7	asym	.046	.055	.063	.052	.064	.075
200	12	.80	3	sym	.032	.047	.067	.042	.062	.090
200	12	.80	3	asym	.035	.048	.067	.043	.064	.093
200	12	.80	5	sym	.027	.044	.067	.032	.050	.078
200	12	.80	5	asym	.031	.046	.068	.034	.054	.083
200	12	.80	7	sym	.027	.044	.071	.030	.048	.081
200	12	.80	7	asym	.031	.046	.069	.032	.052	.083
500	6	.40	3	sym	.031	.032	.036	.044	.045	.051
500	6	.40	3	asym	.032	.032	.036	.047	.047	.053
500	6	.40	5	sym	.030	.032	.036	.037	.039	.044
500	6	.40	5	asym	.032	.032	.035	.040	.041	.046
500	6	.40	7	sym	.029	.032	.036	.035	.038	.043
500	6	.40	7	asym	.031	.032	.036	.038	.040	.046
500	6	.60	3	sym	.026	.030	.036	.036	.041	.052
500	6	.60	3	asym	.027	.031	.037	.037	.043	.055
500	6	.60	5	sym	.024	.030	.038	.029	.037	.047
500	6	.60	5	asym	.026	.031	.037	.032	.039	.049
500	6	.60	7	sym	.024	.029	.038	.028	.034	.045
500	6	.60	7	asym	.026	.031	.037	.030	.037	.047
500	6	.80	3	sym	.018	.026	.041	.024	.036	.058
500	6	.80	3	asym	.019	.027	.041	.024	.037	.061
500	6	.80	5	sym	.015	.025	.041	.018	.030	.050
500	6	.80	5	asym	.017	.026	.039	.019	.032	.050
500	6	.80	7	sym	.015	.025	.041	.017	.028	.048
500	6	.80	7	asym	.017	.026	.040	.018	.031	.051
500	12	.40	3	sym	.035	.037	.039	.047	.050	.052
500	12	.40	3	asym	.036	.038	.040	.049	.053	.056

sample size	number of indicators	loading magnitude	response options	distribution	ML/MLR			DWLS/WLSMV		
					1F	2F		1F	2F	
						.30	.00		.30	.00
500	12	.40	5	sym	.035	.037	.039	.040	.043	.046
500	12	.40	5	asym	.035	.038	.040	.043	.046	.049
500	12	.40	7	sym	.034	.037	.039	.039	.042	.044
500	12	.40	7	asym	.035	.038	.039	.041	.044	.046
500	12	.60	3	sym	.029	.034	.039	.039	.045	.053
500	12	.60	3	asym	.031	.035	.039	.040	.048	.055
500	12	.60	5	sym	.027	.034	.039	.031	.039	.046
500	12	.60	5	asym	.029	.035	.039	.034	.041	.048
500	12	.60	7	sym	.027	.034	.039	.031	.038	.045
500	12	.60	7	asym	.029	.034	.039	.032	.039	.046
500	12	.80	3	sym	.020	.030	.042	.026	.040	.057
500	12	.80	3	asym	.022	.031	.043	.026	.041	.061
500	12	.80	5	sym	.017	.028	.043	.019	.031	.051
500	12	.80	5	asym	.019	.029	.043	.021	.034	.053
500	12	.80	7	sym	.016	.027	.042	.018	.030	.048
500	12	.80	7	asym	.019	.029	.042	.019	.032	.051
2,000	6	.40	3	sym	.015	.016	.018	.022	.023	.026
2,000	6	.40	3	asym	.016	.016	.018	.023	.024	.026
2,000	6	.40	5	sym	.015	.016	.018	.019	.020	.022
2,000	6	.40	5	asym	.015	.016	.018	.020	.021	.024
2,000	6	.40	7	sym	.015	.016	.018	.018	.019	.021
2,000	6	.40	7	asym	.015	.016	.018	.019	.019	.022
2,000	6	.60	3	sym	.013	.015	.018	.018	.021	.026
2,000	6	.60	3	asym	.013	.015	.018	.019	.022	.027
2,000	6	.60	5	sym	.012	.015	.018	.015	.018	.022
2,000	6	.60	5	asym	.013	.015	.018	.015	.019	.024
2,000	6	.60	7	sym	.012	.014	.018	.014	.017	.022
2,000	6	.60	7	asym	.013	.015	.018	.015	.018	.023
2,000	6	.80	3	sym	.009	.013	.020	.012	.018	.029
2,000	6	.80	3	asym	.009	.014	.020	.012	.019	.030
2,000	6	.80	5	sym	.007	.012	.020	.009	.015	.025
2,000	6	.80	5	asym	.008	.013	.020	.009	.016	.026
2,000	6	.80	7	sym	.007	.012	.020	.008	.014	.024
2,000	6	.80	7	asym	.008	.013	.020	.009	.015	.026
2,000	12	.40	3	sym	.018	.019	.019	.024	.025	.026
2,000	12	.40	3	asym	.018	.019	.020	.025	.026	.028
2,000	12	.40	5	sym	.017	.018	.019	.020	.021	.023
2,000	12	.40	5	asym	.018	.019	.019	.021	.023	.024
2,000	12	.40	7	sym	.017	.018	.019	.019	.021	.022
2,000	12	.40	7	asym	.018	.019	.020	.020	.022	.023
2,000	12	.60	3	sym	.014	.017	.020	.019	.022	.026
2,000	12	.60	3	asym	.015	.017	.020	.020	.023	.028
2,000	12	.60	5	sym	.014	.017	.019	.016	.019	.023
2,000	12	.60	5	asym	.015	.017	.020	.017	.021	.024
2,000	12	.60	7	sym	.013	.017	.019	.015	.019	.022
2,000	12	.60	7	asym	.014	.017	.019	.016	.020	.023
2,000	12	.80	3	sym	.010	.015	.021	.013	.019	.029
2,000	12	.80	3	asym	.011	.015	.021	.013	.020	.030
2,000	12	.80	5	sym	.008	.014	.022	.010	.016	.025
2,000	12	.80	5	asym	.010	.015	.022	.010	.017	.028
2,000	12	.80	7	sym	.008	.014	.021	.009	.015	.024
2,000	12	.80	7	asym	.009	.014	.021	.009	.016	.025

Note. MLR = MLR (Yuan & Bentler, 2000). 1F = one-factor CFA. 2F = two-factor CFA: factor correlation either .30 (correlated) or .00 (uncorrelated).

Additional File 5: Code to Derive Tailored Cutoffs via Regression Formulae in R

```
#######################
#Derive tailored cutoffs via regression formulae
#######################

#Please insert the data and analysis characteristics of your model of interest
#Example: ML estimator, 6 items, 7-point Likert scale, symmetric indicator distribution,
average loading magnitude of 0.8, N = 500, one-factor CFA

estimatorMLR <- 0 #1 if MLR estimator, 0 if not
estimatorDWLS <- 0 #1 if DWLS estimator, 0 if not
estimatorWLSMV <- 0 #1 if WLSMV estimator, 0 if not
nitem <- 6 # number of items
res.op <- 7 # number of response options
distrasym <- 0 #1 if asymmetric indicator distribution, 0 if symmetric distribution
loading.magnitude <- 0.8 #average loading magnitude
sample <- 500 # sample size
nfactorCFA <- 1 # number of latent variables/factors
correlated.factors <- 0 #only applies if nfactorCFA > 1: 1 if correlated factors, 0 if
uncorrelated factors


####################################
#Code to derive tailored cutoffs
####################################

sample <- sample/1000

#chisq
chisq <- (-23.94201+
            estimatorMLR*6.72418+
            estimatorDWLS*5.84976+
            estimatorWLSMV*-4.68805+
            nitem*11.08965+
            res.op*-7.16670+
            I((res.op)^2)*0.72250+
            distrasym*-0.27294+
            loading.magnitude*-25.73792+
            I((loading.magnitude)^2)*20.41717+
            sample*0.00906+
            I((sample)^2)*1.20211+
            nfactorCFA*-12.26618+
            estimatorMLR*nitem*-0.49485+
            estimatorMLR*res.op*0.17085+
            estimatorMLR*I((res.op)^2)*-0.02131+
            estimatorMLR*distrasym*-2.71568+
            estimatorMLR*loading.magnitude*-7.76460+
            estimatorMLR*I((loading.magnitude)^2)*-2.61117+
            estimatorMLR*sample*-3.93101+
            estimatorMLR*I((sample)^2)*1.45709+
            estimatorMLR*nfactorCFA*2.71283+
            estimatorDWLS*nitem*-2.43747+
            estimatorDWLS*res.op*-0.39327+
            estimatorDWLS*I((res.op)^2)*0.02110+
            estimatorDWLS*distrasym*-3.01669+
            estimatorDWLS*loading.magnitude*-41.99689+
            estimatorDWLS*I((loading.magnitude)^2)*16.73726+
            estimatorDWLS*sample*-2.10846+
            estimatorDWLS*I((sample)^2)*0.75628+
            estimatorDWLS*nfactorCFA*16.86537+
            estimatorWLSMV*nitem*-0.60239+
            estimatorWLSMV*res.op*0.63654+
            estimatorWLSMV*I((res.op)^2)*-0.05539+
            estimatorWLSMV*distrasym*-2.91980+
            estimatorWLSMV*loading.magnitude*10.29574+
            estimatorWLSMV*I((loading.magnitude)^2)*-15.70961+
            estimatorWLSMV*sample*3.01133+
            estimatorWLSMV*I((sample)^2)*-1.16888+
            estimatorWLSMV*nfactorCFA*3.90897+
```

```
            nitem*res.op*-0.25997+
            nitem*I((res.op)^2)*0.02789+
            nitem*distrasym*0.17890+
            nitem*loading.magnitude*-4.80064+
            nitem*I((loading.magnitude)^2)*3.83154+
            nitem*sample*-1.04664+
            nitem*I((sample)^2)*0.38895+
            nitem*nfactorCFA*0.64889+
            res.op*distrasym*0.47743+
            res.op*loading.magnitude*22.43204+
            res.op*I((loading.magnitude)^2)*-19.13312+
            res.op*sample*2.42094+
            res.op*I((sample)^2)*-1.16974+
            res.op*nfactorCFA*1.18404+
            I((res.op)^2)*distrasym*-0.04690+
            I((res.op)^2)*loading.magnitude*-2.12451+
            I((res.op)^2)*I((loading.magnitude)^2)*1.79085+
            I((res.op)^2)*sample*-0.40475+
            I((res.op)^2)*I((sample)^2)*0.17973+
            I((res.op)^2)*nfactorCFA*-0.10687+
            distrasym*loading.magnitude*3.71952+
            distrasym*I((loading.magnitude)^2)*-1.17710+
            distrasym*sample*0.29781+
            distrasym*I((sample)^2)*-0.16484+
            distrasym*nfactorCFA*-0.81043+
            loading.magnitude*sample*16.43214+
            loading.magnitude*I((sample)^2)*-8.22119+
            loading.magnitude*nfactorCFA*1.82103+
            I((loading.magnitude)^2)*sample*-15.03742+
            I((loading.magnitude)^2)*I((sample)^2)*7.54665+
            I((loading.magnitude)^2)*nfactorCFA*5.02726+
            sample*nfactorCFA*0.39375+
            I((sample)^2)*nfactorCFA*-0.26943+
            nfactorCFA*correlated.factors*-2.51728)


#chisq.df
chisq.df <- (3.28519+
                 estimatorMLR*0.45189+
                 estimatorDWLS*-0.68404+
                 estimatorWLSMV*-0.27096+
                 nitem*-0.04753+
                 res.op*-0.35058+
                 I((res.op)^2)*0.03496+
                 distrasym*0.02331+
                 loading.magnitude*-3.58376+
                 I((loading.magnitude)^2)*2.96247+
                 sample*0.45022+
                 I((sample)^2)*-0.15723+
                 nfactorCFA*-0.19792+
                 estimatorMLR*nitem*-0.01090+
                 estimatorMLR*res.op*0.00216+
                 estimatorMLR*I((res.op)^2)*-0.00052+
                 estimatorMLR*distrasym*-0.08311+
                 estimatorMLR*loading.magnitude*-0.99175+
                 estimatorMLR*I((loading.magnitude)^2)*0.42949+
                 estimatorMLR*sample*-0.26550+
                 estimatorMLR*I((sample)^2)*0.09907+
                 estimatorMLR*nfactorCFA*0.11781+
                 estimatorDWLS*nitem*0.00544+
                 estimatorDWLS*res.op*-0.02550+
                 estimatorDWLS*I((res.op)^2)*0.00195+
                 estimatorDWLS*distrasym*-0.09613+
                 estimatorDWLS*loading.magnitude*-1.36226+
                 estimatorDWLS*I((loading.magnitude)^2)*0.48895+
                 estimatorDWLS*sample*-0.11430+
                 estimatorDWLS*I((sample)^2)*0.04419+
                 estimatorDWLS*nfactorCFA*0.64662+
                 estimatorWLSMV*nitem*-0.00413+
                 estimatorWLSMV*res.op*0.01440+
                 estimatorWLSMV*I((res.op)^2)*-0.00115+
                 estimatorWLSMV*distrasym*-0.09484+
                 estimatorWLSMV*loading.magnitude*0.24493+
                 estimatorWLSMV*I((loading.magnitude)^2)*-0.45035+
                 estimatorWLSMV*sample*0.05021+
                 estimatorWLSMV*I((sample)^2)*-0.01938+
```

```
                estimatorWLSMV*nfactorCFA*0.15706+
                nitem*res.op*-0.00776+
                nitem*I((res.op)^2)*0.00081+
                nitem*distrasym*0.00040+
                nitem*loading.magnitude*-0.04388+
                nitem*I((loading.magnitude)^2)*0.04484+
                nitem*sample*0.00016+
                nitem*I((sample)^2)*0.00157+
                nitem*nfactorCFA*-0.01164+
                res.op*distrasym*0.01356+
                res.op*loading.magnitude*1.53987+
                res.op*I((loading.magnitude)^2)*-1.33866+
                res.op*sample*-0.20111+
                res.op*I((sample)^2)*0.07428+
                res.op*nfactorCFA*0.04887+
                I((res.op)^2)*distrasym*-0.00122+
                I((res.op)^2)*loading.magnitude*-0.14685+
                I((res.op)^2)*I((loading.magnitude)^2)*0.12730+
                I((res.op)^2)*sample*0.01424+
                I((res.op)^2)*I((sample)^2)*-0.00524+
                I((res.op)^2)*nfactorCFA*-0.00463+
                distrasym*loading.magnitude*0.27558+
                distrasym*I((loading.magnitude)^2)*-0.18396+
                distrasym*sample*-0.00253+
                distrasym*I((sample)^2)*-0.00346+
                distrasym*nfactorCFA*-0.03425+
                loading.magnitude*sample*0.03858+
                loading.magnitude*I((sample)^2)*-0.08583+
                loading.magnitude*nfactorCFA*0.21559+
                I((loading.magnitude)^2)*sample*-0.14310+
                I((loading.magnitude)^2)*I((sample)^2)*0.12608+
                I((loading.magnitude)^2)*nfactorCFA*0.03351+
                sample*nfactorCFA*0.08076+
                I((sample)^2)*nfactorCFA*-0.03848+
                nfactorCFA*correlated.factors*-0.05765)

#cfi
cfi <- (-0.53129+
            estimatorMLR*-0.21041+
            estimatorDWLS*0.19662+
            estimatorWLSMV*0.06079+
            nitem*0.04016+
            res.op*0.12387+
            I((res.op)^2)*-0.00936+
            distrasym*-0.04904+
            loading.magnitude*4.12967+
            I((loading.magnitude)^2)*-2.75074+
            sample*2.27580+
            I((sample)^2)*-0.82698+
            nfactorCFA*-0.32211+
            estimatorMLR*nitem*0.00247+
            estimatorMLR*res.op*0.00384+
            estimatorMLR*I((res.op)^2)*-0.00024+
            estimatorMLR*distrasym*-0.00135+
            estimatorMLR*loading.magnitude*0.45378+
            estimatorMLR*I((loading.magnitude)^2)*-0.31994+
            estimatorMLR*sample*0.09707+
            estimatorMLR*I((sample)^2)*-0.03794+
            estimatorMLR*nfactorCFA*-0.00868+
            estimatorDWLS*nitem*0.00158+
            estimatorDWLS*res.op*-0.00440+
            estimatorDWLS*I((res.op)^2)*0.00034+
            estimatorDWLS*distrasym*0.00452+
            estimatorDWLS*loading.magnitude*-0.30944+
            estimatorDWLS*I((loading.magnitude)^2)*0.18350+
            estimatorDWLS*sample*-0.11629+
            estimatorDWLS*I((sample)^2)*0.04250+
            estimatorDWLS*nfactorCFA*-0.01075+
            estimatorWLSMV*nitem*0.00097+
            estimatorWLSMV*res.op*-0.00270+
            estimatorWLSMV*I((res.op)^2)*0.00022+
            estimatorWLSMV*distrasym*0.00368+
            estimatorWLSMV*loading.magnitude*-0.06415+
            estimatorWLSMV*I((loading.magnitude)^2)*0.03405+
            estimatorWLSMV*sample*-0.04314+
            estimatorWLSMV*I((sample)^2)*0.01545+
```

```
            estimatorWLSMV*nfactorCFA*-0.00697+
            nitem*res.op*0.00002+
            nitem*I((res.op)^2)*-0.00003+
            nitem*distrasym*0.00034+
            nitem*loading.magnitude*-0.10488+
            nitem*I((loading.magnitude)^2)*0.07017+
            nitem*sample*-0.01404+
            nitem*I((sample)^2)*0.00500+
            nitem*nfactorCFA*0.00234+
            res.op*distrasym*-0.00319+
            res.op*loading.magnitude*-0.33800+
            res.op*I((loading.magnitude)^2)*0.23297+
            res.op*sample*-0.00951+
            res.op*I((sample)^2)*0.00109+
            res.op*nfactorCFA*0.00555+
            I((res.op)^2)*distrasym*0.00031+
            I((res.op)^2)*loading.magnitude*0.02682+
            I((res.op)^2)*I((loading.magnitude)^2)*-0.01863+
            I((res.op)^2)*sample*0.00006+
            I((res.op)^2)*I((sample)^2)*0.00020+
            I((res.op)^2)*nfactorCFA*-0.00039+
            distrasym*loading.magnitude*0.11280+
            distrasym*I((loading.magnitude)^2)*-0.07174+
            distrasym*sample*0.02437+
            distrasym*I((sample)^2)*-0.00870+
            distrasym*nfactorCFA*-0.00208+
            loading.magnitude*sample*-5.87140+
            loading.magnitude*I((sample)^2)*2.14448+
            loading.magnitude*nfactorCFA*0.65988+
            I((loading.magnitude)^2)*sample*3.87122+
            I((loading.magnitude)^2)*I((sample)^2)*-1.41187+
            I((loading.magnitude)^2)*nfactorCFA*-0.43878+
            sample*nfactorCFA*0.09529+
            I((sample)^2)*nfactorCFA*-0.03378+
            nfactorCFA*correlated.factors*0.00487)

#rmsea
rmsea <- (0.13285+
              estimatorMLR*0.01536+
              estimatorDWLS*-0.03062+
              estimatorWLSMV*-0.00865+
              nitem*-0.00235+
              res.op*-0.00896+
              I((res.op)^2)*0.00098+
              distrasym*-0.00024+
              loading.magnitude*-0.08865+
              I((loading.magnitude)^2)*0.05766+
              sample*-0.12606+
              I((sample)^2)*0.04331+
              nfactorCFA*-0.00594+
              estimatorMLR*nitem*-0.00041+
              estimatorMLR*res.op*0.00005+
              estimatorMLR*I((res.op)^2)*-0.00001+
              estimatorMLR*distrasym*-0.00225+
              estimatorMLR*loading.magnitude*-0.03246+
              estimatorMLR*I((loading.magnitude)^2)*0.01556+
              estimatorMLR*sample*-0.00768+
              estimatorMLR*I((sample)^2)*0.00311+
              estimatorMLR*nfactorCFA*0.00304+
              estimatorDWLS*nitem*-0.00038+
              estimatorDWLS*res.op*-0.00033+
              estimatorDWLS*I((res.op)^2)*0.00003+
              estimatorDWLS*distrasym*-0.00244+
              estimatorDWLS*loading.magnitude*-0.05998+
              estimatorDWLS*I((loading.magnitude)^2)*0.02118+
              estimatorDWLS*sample*0.02896+
              estimatorDWLS*I((sample)^2)*-0.00982+
              estimatorDWLS*nfactorCFA*0.02281+
              estimatorWLSMV*nitem*-0.00029+
              estimatorWLSMV*res.op*0.00054+
              estimatorWLSMV*I((res.op)^2)*-0.00004+
              estimatorWLSMV*distrasym*-0.00256+
              estimatorWLSMV*loading.magnitude*0.00577+
              estimatorWLSMV*I((loading.magnitude)^2)*-0.01194+
              estimatorWLSMV*sample*0.00682+
              estimatorWLSMV*I((sample)^2)*-0.00250+
```

```
               estimatorWLSMV*nfactorCFA*0.00452+
               nitem*res.op*-0.00033+
               nitem*I((res.op)^2)*0.00004+
               nitem*distrasym*0.00002+
               nitem*loading.magnitude*-0.00698+
               nitem*I((loading.magnitude)^2)*0.00652+
               nitem*sample*0.00655+
               nitem*I((sample)^2)*-0.00224+
               nitem*nfactorCFA*-0.00030+
               res.op*distrasym*0.00043+
               res.op*loading.magnitude*0.04504+
               res.op*I((loading.magnitude)^2)*-0.03818+
               res.op*sample*-0.00561+
               res.op*I((sample)^2)*0.00208+
               res.op*nfactorCFA*0.00098+
               I((res.op)^2)*distrasym*-0.00005+
               I((res.op)^2)*loading.magnitude*-0.00451+
               I((res.op)^2)*I((loading.magnitude)^2)*0.00381+
               I((res.op)^2)*sample*0.00030+
               I((res.op)^2)*I((sample)^2)*-0.00011+
               I((res.op)^2)*nfactorCFA*-0.00010+
               distrasym*loading.magnitude*0.01327+
               distrasym*I((loading.magnitude)^2)*-0.00937+
               distrasym*sample*-0.00219+
               distrasym*I((sample)^2)*0.00072+
               distrasym*nfactorCFA*-0.00101+
               loading.magnitude*sample*0.01187+
               loading.magnitude*I((sample)^2)*-0.00411+
               loading.magnitude*nfactorCFA*0.02703+
               I((loading.magnitude)^2)*sample*0.00022+
               I((loading.magnitude)^2)*I((sample)^2)*0.00015+
               I((loading.magnitude)^2)*nfactorCFA*-0.01413+
               sample*nfactorCFA*-0.00988+
               I((sample)^2)*nfactorCFA*0.00320+
               nfactorCFA*correlated.factors*-0.00223)


#srmr
srmr <- (0.05279+
             estimatorDWLS*0.03774+
             nitem*0.00278+
             res.op*-0.00963+
             I((res.op)^2)*0.00084+
             distrasym*-0.00115+
             loading.magnitude*0.02653+
             I((loading.magnitude)^2)*-0.09506+
             sample*-0.05619+
             I((sample)^2)*0.01882+
             nfactorCFA*0.01323+
             estimatorDWLS*nitem*-0.00022+
             estimatorDWLS*res.op*-0.00758+
             estimatorDWLS*I((res.op)^2)*0.00058+
             estimatorDWLS*distrasym*0.00140+
             estimatorDWLS*loading.magnitude*-0.00280+
             estimatorDWLS*I((loading.magnitude)^2)*-0.00311+
             estimatorDWLS*sample*-0.02165+
             estimatorDWLS*I((sample)^2)*0.00756+
             estimatorDWLS*nfactorCFA*0.00330+
             nitem*res.op*-0.00019+
             nitem*I((res.op)^2)*0.00002+
             nitem*distrasym*0.00003+
             nitem*loading.magnitude*-0.00316+
             nitem*I((loading.magnitude)^2)*0.00190+
             nitem*sample*-0.00160+
             nitem*I((sample)^2)*0.00058+
             nitem*nfactorCFA*0.00003+
             res.op*distrasym*0.00061+
             res.op*loading.magnitude*0.01794+
             res.op*I((loading.magnitude)^2)*-0.01639+
             res.op*sample*0.01125+
             res.op*I((sample)^2)*-0.00388+
             res.op*nfactorCFA*0.00068+
             I((res.op)^2)*distrasym*-0.00006+
             I((res.op)^2)*loading.magnitude*-0.00166+
             I((res.op)^2)*I((loading.magnitude)^2)*0.00151+
             I((res.op)^2)*sample*-0.00097+
```

```
            I((res.op)^2)*I((sample)^2)*0.00034+
            I((res.op)^2)*nfactorCFA*-0.00005+
            distrasym*loading.magnitude*0.00844+
            distrasym*I((loading.magnitude)^2)*-0.00713+
            distrasym*sample*-0.00393+
            distrasym*I((sample)^2)*0.00138+
            distrasym*nfactorCFA*-0.00028+
            loading.magnitude*sample*-0.04098+
            loading.magnitude*I((sample)^2)*0.01586+
            loading.magnitude*nfactorCFA*-0.01793+
            I((loading.magnitude)^2)*sample*0.06458+
            I((loading.magnitude)^2)*I((sample)^2)*-0.02336+
            I((loading.magnitude)^2)*nfactorCFA*0.04983+
            sample*nfactorCFA*-0.02326+
            I((sample)^2)*nfactorCFA*0.00784+
            nfactorCFA*correlated.factors*-0.00481)

print(paste0("The following cutoffs were derived based on the regression formulae: Chi2/df =
", round(chisq.df,3),
            ", CFI = ", round(cfi, 3), ", RMSEA = ", round(rmsea,3), ", SRMR = ",
round(srmr,3),". The Chi2 cutoff was ", round(chisq,3),
            ", but please note that it heavily depends on the degrees of freedom and should
not be used for models different from the ones in the paper.", collapse=""))
```

Making model judgments ROC(K)-solid: Tailored cutoffs for fit indices through simulation and ROC analysis in structural equation modeling

Katharina Groskurth[1,2], Nivedita Bhaktha[1], and Clemens M. Lechner[1]

[1] GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany

[2] Graduate School of Economic and Social Sciences, University of Mannheim, Germany

Abstract

To evaluate model fit in confirmatory factor analysis, researchers compare fit indices against fixed cutoff values (e.g., CFI ≥ .950 indicates a well-fitting model). Although using fixed cutoffs is widespread, methodologists have long cautioned against overgeneralizing such cutoffs, highlighting that one should only apply fixed cutoffs in settings highly similar to the simulation studies from which those cutoffs originate. Values of fit indices vary depending on the model, estimation, and data characteristics of the empirical setting. Conclusions regarding whether a model does or does not fit the data may ultimately be biased. As a solution, methodologists have proposed four principal approaches to obtain so-called tailored (or "dynamic") cutoffs that are explicitly generated for the setting of interest. We herein review these approaches. Notably, none of these approaches have yet provided guidelines on which fit index (out of all fit indices of interest) best evaluates whether the model fits the data (or not) in the specific setting. Thus, we proposed the so-called simulation-cum-ROC approach that combines a Monte Carlo simulation with receiver operating characteristic (ROC) analysis. The simulation-cum-ROC approach generates tailored cutoffs while identifying the most reliable fit indices in the given setting. We provided computational R code and a shiny app for an easy implementation of the approach. Thus, prior knowledge of Monte Carlo simulations and ROC analysis is not needed to generate tailored cutoffs with the simulation-cum-ROC approach.

The Problem with Fixed Cutoffs for A Specific Set of Fit Indices

To test the goodness of confirmatory factor analysis (CFA) models—and structural equation models (SEM) more generally—researchers typically rely on cutoffs for fit indices (Jackson et al., 2009; Kline, 2016). Apart from testing exact model fit via the chi-square test statistic (χ^2; e.g., Bollen, 1989), some of the most commonly used global fit indices are the comparative fit index (CFI; Bentler, 1990), the root mean square error of approximation (RMSEA; Steiger, 1990), and the standardized root mean residual (SRMR; Bentler, 1995). Most prominently, Hu and Bentler (1999) proposed, based on a simulation study, that CFI should be above or close to .950, RMSEA should be below or close to .060, and SRMR should be below or close to .080 to indicate good model fit. By comparing empirical values of fit indices against these cutoffs, researchers evaluate whether their model does fit the data (i.e., is assumed to be correctly specified) or does not fit the data (i.e., is assumed to be misspecified). This simple binary (yes/no) decision-making on model fit using the same, fixed cutoffs across diverse empirical settings has guided research for decades (e.g., Jackson et al., 2009).

However, evaluating model fit via fixed cutoffs for fit indices is more problematic than many researchers appear to realize. Fit indices are not only sensitive to misspecification, as intended, but undesirably susceptible to a range of model, estimation, and data characteristics. These characteristics include, for example, the magnitude of factor loadings, the type of estimator, the sample size, and interactions thereof, especially when the model is misspecified (e.g., Groskurth et al., 2022; Heene et al., 2011; Moshagen & Auerswald, 2018; Shi et al., 2019; Xia & Yang, 2018, 2019; for an overview, see Niemand & Mai, 2018). Likewise, the (non-) normality of the items' multivariate response distribution influences fit indices, regardless of whether the model is correctly specified or misspecified (e.g., Fouladi, 2000; Yuan & Bentler, 1999, 2000b; Yuan et al., 2004). Further complicating matters, different fit indices react differently to model misspecification, extraneous characteristics, and the interaction between them (Groskurth et al., 2022; Lai & Green, 2016; Moshagen & Auerswald, 2018).

Thus, two key challenges exist in using fit indices for model evaluation. First, *the performance of fit indices to detect model misspecification can vary greatly across empirical settings.* The differential performance threatens the ability of fit indices to demarcate between correctly specified and misspecified models (e.g., Reußner, 2019). No fit index universally outperforms others (for an overview, see Groskurth et al., 2022; Niemand & Mai, 2018). Second, by implication, *cutoffs for fit indices pertain only to a specific empirical setting* (i.e.,

a combination between model, estimation, and data characteristics). Cutoffs may no longer be valid in empirical settings that diverge markedly from the simulation studies generating the cutoffs (e.g., Hu & Bentler, 1999; McNeish & Wolf, 2021).

Therefore, it is impossible to arrive at general rules on the performance of specific fit indices, let alone fixed cutoffs that are universally applicable across settings. It is likewise impossible to devise a simulation study that includes all possible settings. Although Hu and Bentler (1999) already warned against overgeneralizing their cutoffs, their cautionary note seems to have been largely unheeded in applied research (e.g., Jackson et al., 2009; McNeish & Wolf, 2021). In practice, researchers apply cutoffs for fit indices rather uncritically. Treating the once-proposed cutoffs and sets of fit indices as "golden rules" can result in wrong conclusions regarding model fit (Marsh et al., 2004; for examples, see McNeish & Wolf, 2021). Such erroneous results threaten the integrity of scientific findings.

A solution that has long been proposed is to use tailored or "dynamic" cutoffs for fit indices customized to a specific setting of interest (Millsap, 2013; see also Kim & Millsap, 2014, based on Bollen & Stine, 1992). Tailored cutoffs are not yet widely used despite recently regaining traction (e.g., McNeish & Wolf, 2021, 2022). Toward the ultimate aim of helping researchers transition to more valid model evaluation practices via tailored cutoffs, the first goal of this study was to review and summarize existing approaches to generating tailored cutoffs. Such a systematic overview is currently missing from the literature. This review revealed that all approaches to generating tailored cutoffs have unique strengths and, while being generally superior to fixed cutoffs, share some limitations. Chief among these limitations is that none of the extant approaches allows for a performance evaluation of fit indices. Thus, they provide no guidelines on which fit index (out of all fit indices of interest) can best discriminate between correctly specified and misspecified models in a given setting.

Therefore, the second goal of our study was to introduce and illustrate a novel approach that builds on—and extends—prior approaches (e.g., McNeish & Wolf, 2021, 2022; Millsap, 2013; Pornprasertmanit, 2014). It combines a Monte Carlo simulation, an often-used procedure in psychometrics, with a receiver operating characteristic (ROC) analysis widely used in machine learning. Our so-called simulation-cum-ROC approach answers two questions: (1) Which fit indices, if any, perform well in a setting of interest? (2) Which cutoffs best discriminate between correctly specified and misspecified models in that setting? In this regard, our approach generates tailored cutoffs for well-performing fit indices. We illustrate this approach with empirical examples and provide a shiny app that facilitates its application.

The Logic Behind Generating Tailored Cutoffs for Fit Indices

In recent years, methodologists have called for a move away from using the same fixed cutoffs across diverse empirical settings and proposed several approaches to generate cutoffs tailored to the setting of interest (e.g., McNeish & Wolf, 2021; Millsap, 2013; Pornprasertmanit, 2014). Before introducing any of these approaches to tailored cutoffs, we need to differentiate between two core situations: In an empirical testing setting (i.e., when fitting the analysis model of interest to empirical data), one never knows whether the analysis model is correctly specified or misspecified because the true population model is unknown. In a hypothetical scenario (e.g., when simulating data), one knows that the analysis model is either correctly specified or misspecified because one can define the population model.

Before testing the analysis model with tailored cutoffs for fit indices, we need to encode different hypotheses about how the empirical data might have come about. More specifically, we follow the Neyman-Pearson approach to hypothesis testing here (Neyman & Pearson, 1928, 1933; see Biau et al., 2010, Moshagen & Erdfelder, 2016; Perezgonzalez, 2015). The Neyman-Pearson approach requires specifying a null hypothesis (H_0) and an alternative hypothesis (H_1). H_0 states that the analysis model is identical to the population model. The analysis model is correctly specified. H_1 states that the analysis model is not identical to the population model to a specific degree of intolerable misspecification. The analysis model is misspecified.

All approaches to tailored cutoffs (e.g., McNeish & Wolf, 2021) operationalize the two hypotheses within hypothetical scenarios (i.e., with known population models). They provide different ways to operationalize how fit index values might be distributed when the analysis model is either correctly specified or misspecified. In particular, approaches to tailored cutoffs do not use any hypothetical scenario (like approaches to fixed cutoffs would do), but they explicitly operationalize the fit index distribution within the setting of interest (e.g., using a sample size equal to the empirical one and the analysis model of interest). A cutoff is then generated from these fit index distributions (e.g., corresponding to a certain percentile). A cutoff is selected in a way that it classifies a large share of correctly specified models as correctly specified and a large share of misspecified models as misspecified. It should classify only a small share of correctly specified models as misspecified (i.e., Type I error rate) and only a small share of misspecified models as correctly specified (i.e., Type II error rate).

After deriving a tailored cutoff by utilizing fit index distributions under hypothetical scenarios through any of the various approaches, one tests with this cutoff which hypothesis—

H_0 or H_1—is more plausible given the empirical data generated from an unknown population model (Neyman & Pearson, 1928).[1] If the empirical fit index value passes the cutoff, one accepts the analysis model. Accepting the analysis model means finding empirical evidence that favors the H_0 instead of the H_1. However, as the population model that has generated the empirical data remains unknown, one can never exclude that there is a better analysis model among the myriad possible models that one has not considered (Groeben & Westmeyer, 1981). If the empirical fit index value fails the cutoff, one rejects the analysis model. Rejecting the analysis model means finding empirical evidence that favors the H_1 instead of the H_0. Empirical evidence suggests a model superior to the analysis model—however, that analysis model remains unknown, like the population model that has generated the empirical data (Groeben & Westmeyer, 1981).

[1] Another way to test whether empirical evidence favors the H_0 (i.e., the model-implied variance-covariance matrix equals the one in the population) is to look at confidence intervals for fit indices. If those confidence intervals include (or are very close to) 0 indicating perfect fit (for RMSEA and SRMR, alternatively 1 for CFI), empirical evidence favors the H_0 (e.g., Schermelleh-Engel et al., 2003; or at least one is not able to find evidence against it, Yuan et al., 2016). Confidence intervals have been suggested for several widely used fit indices such as CFI (Cheng & Wu, 2017; Lai, 2019; Yuan et al., 2016; Zhang & Savalei, 2016), RMSEA (Brosseau-Liard et al., 2012; Browne & Cudeck, 1992; Cheng & Wu, 2017; Zhang & Savalei, 2016), and SRMR (Cheng & Wu, 2017; Maydeu-Olivares, 2017; Maydeu-Olivares et al., 2018).

A Review of Existing Approaches to Generating Tailored Cutoffs

Currently, there are four principal approaches allowing to generate tailored cutoffs in hypothetical scenarios (Table 1)[1] that fall on a continuum from parametric to non-parametric approaches:

(1) The χ^2 distribution-based approach generates cutoffs by relying on statistical assumptions of the χ^2 distribution without and with misspecification (Moshagen & Erdfelder, 2016).

(2) The regression-based approach generates cutoffs based on meta-regression results from a prior simulation study (Nye & Drasgow, 2011; see also Groskurth et al., 2022).

(3) The simulation-based approach generates cutoffs based on fit index distributions from an analysis model fit to multiple samples from a known population model (McNeish & Wolf, 2021, 2022; Millsap, 2007, 2013; Mai et al., 2021; Niemand & Mai, 2018; Pornprasertmanit, 2014).

(4) The bootstrap approach generates cutoffs based on fit index distributions from an analysis model fit to multiple samples based on transformed empirical data (i.e., as if the model has generated the data, Bollen & Stine, 1992; Kim & Millsap, 2014).

[1] Some would also include the table-based approach to generating tailored cutoffs (e.g., Groskurth et al., 2022). Reminiscent of looking up critical values for *z*-scores, one reads out scenario-specific cutoffs from large tables originating from simulation studies. However, as this approach is still very inflexible (as it only allows to read out cutoffs for those scenarios covered in the initial simulation study), we dismiss the approach in our review.

Table 1: *Existing Approaches to Generate Tailored Cutoffs*

Principal approach	Author(s)	Type I error?	Type II error?	Performance of fit indices?	Tailored to …	Helpful resources
χ^2 Distribution: Generating cutoffs based on distributional assumptions	Moshagen & Erdfelder (2016)	✓	✓	✗	sample size, degrees of freedom, and number of items for fit indices whose distributions can be derived from χ^2	Shiny app: https://sempower.shinyapps.io/sempower, https://sjak.shinyapps.io/power4SEM/ (Jak et al., 2021) R package: semPower Tutorial: Jobst et al. (2021)
Regression: Generating cutoffs based on meta-regressions	Nye & Drasgow (2011)	✓	✗	✗	sample size and response distribution for RMSEA and SRMR	Regression formulae: included in the paper
	Groskurth et al. (2022)	✓	✗	✗	estimator, number of items, number of response options, response distribution, loading magnitude, sample size, and factor correlation for χ^2, χ^2/degrees of freedom, CFI, RMSEA, SRMR	Regression formulae: included in the paper R code: included in the paper
Simulation: Generating cutoffs based on simulated fit index distributions	Niemand & Mai (2018), Mai et al. (2021)	✓	✗	✗	number of items, number of factors, loading magnitude, degrees of freedom, sample size, and response distribution for multiple fit indices	R package: FCO
	Millsap (2007, 2013)	✓	✗	✗	all model, estimation, and data characteristics for available fit indices	R package: simsem (Pornprasertmanit et al., 2021), ezCutoffs (Schmalbach et al., 2019)
	McNeish & Wolf (2021, 2022)	✓	✓	✗	all model, estimation, and data characteristics for available fit indices	Shiny app: https://dynamicfit.app/__landing__/ R package: dynamic Mplus code: included in the paper
	Pornprasertmanit (2014)	✓	✓	✗	all model, estimation, and data characteristics for available fit indices	

Principal approach	Author(s)	Type I error?	Type II error?	Performance of fit indices?	Tailored to …	Helpful resources
Bootstrap: Generating cutoffs based on bootstrapped fit index distributions	Bollen & Stine (1992), Kim & Millsap (2014)	✓	✗	✗	all model, estimation, and data characteristics for available fit indices	R package: simsem (Pornprasertmanit et al., 2021), lavaan (Rosseel, 2012) R code: included in the paper
	Yuan & Hayashi (2003), Yuan et al. (2004, 2007)	✓	✓	✗	all model, estimation, and data characteristics for available fit indices	R package: lavaan (Rosseel, 2012)
Simulation + ROC analysis	Present study (simulation-cum-ROC)	✓	✓	✓	all model, estimation, and data characteristics for available fit indices	Shiny app: https://kg11.shinyapps.io/tailoredcutoffs/ R code: included in the paper

Note. Mai et al. (2021) provided general recommendations on the performance of fit indices depending on the purpose of the research question (testing an established versus a novel model), the focus of estimation (testing a measurement model or structural model), and sample size (below or above $N = 200$) derived from an extensive simulation study. As those recommendations were based on a prior simulation study and cannot be specifically derived within settings of interest, we did not highlight them in this table.

χ^2 Distribution-Based Approach

One option to generate tailored cutoffs is via the parametric χ^2 distribution-based approach as outlined by Moshagen and Erdfelder (2016; see also Jak et al., 2021; Jobst et al., 2021). Both for the χ^2 test statistic itself and for fit indices that incorporate the χ^2 test statistic (e.g., RMSEA), one can infer the distributions based on correctly specified and misspecified models from the χ^2 distribution. Those distributions can be harnessed to generate cutoffs.

More specifically, the χ^2 test statistic follows a central χ^2 distribution if the analysis model is correctly specified (i.e., the model-implied variance-covariance matrix is exactly identical to the one in the population, H_0). Contrariwise, the χ^2 test statistic follows a non-central χ^2 distribution if the analysis model is misspecified (i.e., the model-implied matrix is not exactly identical to the one in the population, H_1). The expected value of the central χ^2 distribution equals the model's degrees of freedom. The expected value of the non-central χ^2 distribution equals the degrees of freedom plus a so-called non-centrality parameter. The non-centrality parameter depends on the misspecification and sample size (for a detailed description, see Bollen, 1989; Chun & Shapiro, 2009; Moshagen & Erdfelder, 2016). If one defines a to-be-detected effect size difference (based on the non-centrality parameter) between the central and non-central χ^2 distribution, one can obtain a cutoff for the χ^2 test statistic at a specific ratio of Type I and Type II error rates. Typically, the Type I and Type II error rates are balanced (i.e., equally small).

The χ^2 distribution-based approach has the advantage of computational speed. Statistical tools such as R rapidly solve the equations needed to generate cutoffs. However, a disadvantage of this procedure is the limited extent of tailoring. The approach can only generate cutoffs for fit indices that are transformations of the χ^2 test statistic (e.g., RMSEA). It is not applicable to fit indices that are based, for example, on standardized residuals (e.g., SRMR) and, thus, do not follow a known distribution. Moreover, one calculates tailored cutoffs from Moshagen and Erdfelder's (2016) χ^2 distribution-based approach under specific assumptions such that items follow a multivariate normal distribution, in which case the distribution of the χ^2 test statistic is known. Non-normal multivariate distributions of the items (e.g., Fouladi, 2000; Yuan & Bentler, 1999, 2000b; Yuan et al., 2004) or large models with many items (Moshagen, 2012) violate the distributional assumptions of the χ^2 test statistic. Then, different test statistics (e.g., Yuan & Bentler, 2007) are necessary to generate valid cutoffs that are not always straightforward to handle. In sum, the χ^2 distribution-based approach limits the extent

to which users can tailor cutoffs to their specific setting of interest and the range of fit indices for which users can generate the cutoffs (see Table 1).

Regression-Based Approach

The basic idea of the parametric regression-based approach is to predict tailored cutoffs through a regression formula. The formula originates from a single, though ideally extensive, simulation study. The formula comprises predictors with associated regression coefficients that contain information about how various model, estimation, and data characteristics (e.g., number of items, type of estimator, and distribution of responses) influence cutoffs for a fit index. Users plug the model, estimation, and data characteristics of the setting of interest into the formula to obtain a cutoff.

For example, Nye and Drasgow (2011) simulated data based on multiple predefined characteristics. They evaluated Type I and Type II error rates for a range of possible cutoffs (derived by informed guesses). To arrive at a tailored cutoff, they regressed the Type I error rate on influential characteristics and cutoff values. Rearranging the formula leads to an appropriate cutoff value for the user-defined characteristics and a predefined Type I error rate. By following the same approach, Groskurth et al. (2022) also provided formulae that allow for the prediction of tailored cutoffs, though for more fit indices and across a wider range of characteristics than Nye and Drasgow's (2011) formulae.

Like the χ^2 distribution-based approach, the regression-based approach has the advantage of speed. Users merely have to plug their characteristics into the formula, commonly solved by a statistical tool such as R. However, each formula is only as inclusive as the simulation study from which it was derived. For instance, Nye and Drasgow's (2011) formulae covered only models estimated with diagonally weighted least squares, while Groskurth et al.'s (2022) formulae covered only CFA models. The formula may not be valid for settings beyond those covered in the original simulation study. Further, one can only obtain cutoffs for those fit indices that were considered in the simulation study from which the formulae hail. Groskurth et al. (2022) developed formulae for χ^2, χ^2/degrees of freedom, CFI, RMSEA, and SRMR; Nye and Drasgow (2011) developed formulae for RMSEA and SRMR. Akin to the χ^2 distribution-based approach, the regression-based approach limits the extent to which users can tailor cutoffs to their specific setting of interest and the range of fit indices for which users can generate the cutoffs (see Table 1).

Simulation-Based Approach

As a third parametric approach, one may use a Monte Carlo simulation to generate tailored cutoffs (McNeish & Wolf, 2021, 2022; Millsap, 2007, 2013; Mai et al., 2021; Niemand & Mai, 2018; Pornprasertmanit, 2014; for nested models, see Pornprasertmanit et al., 2013). Before initializing the simulation, one defines a population model. Then, one repeatedly simulates data from that population model, fits the analysis model to each data set, and records the fit index values. One can then set cutoffs based on a certain percentile of the fit index distribution.

In more detail, the general procedure is to specify a population model from which to draw multiple samples. Drawing multiple samples from a known population model characterizes a so-called Monte Carlo simulation (for an overview and detailed description, see Boomsma, 2013). After conducting the Monte Carlo simulation, the next step is to fit the analysis model to each simulated data set (i.e., drawn sample). The analysis model is identical (or nearly identical) to the population model; it captures all relevant features of the population model and is, thus, correctly specified (H_0). After fitting the analysis model to the data, one records the fit index values of each fitted model. A cutoff can then be set based on a specific percentile, commonly the 95th, of the resulting fit index distribution (or, equivalently, the 5th percentile for fit indices where higher values indicate better fit). At this percentile, the cutoff categorizes 95% of correctly specified models as correctly specified and 5% of correctly specified models as misspecified (i.e., the Type I error rate).

One may repeat the procedure with the same analysis model but a population model with more (or different) parameters than the analysis model. For instance, one fixes non-zero parameters in the population model to zero in the analysis model (Hu & Bentler, 1998). The analysis model is, thus, underspecified (i.e., misspecified) relative to the population model (H_1). Misspecification implies that the analysis model fails to capture relevant features of the population model. Including a misspecified scenario allows for evaluating how many misspecified models a cutoff categorizes as correctly specified (i.e., the Type II error rate).

The simulation-based approach is computationally intensive but also very flexible. It allows tailoring cutoffs for *all* fit indices available in a given statistical program to the specific model, estimation, and data characteristics. Combined with the computers' continuously increasing statistical power, this is one of the reasons why this approach has recently gained traction (McNeish & Wolf, 2021; 2022).

Bootstrap Approach

Tailored cutoffs can be generated not only via a Monte Carlo simulation, which is essentially a parametric bootstrap approach (simulating, i.e., sampling, data based on model parameters), but also via a non-parametric bootstrap approach (i.e., sampling data based on transformed empirical data). Thus, the fourth approach uses non-parametric bootstrapping to generate tailored cutoffs in hypothetical scenarios (Bollen & Stine, 1992; Kim & Millsap, 2014; Yuan & Hayashi, 2003; Yuan et al., 2004, 2007). The algorithm transforms the empirical data such that the analysis model fits it. By repeatedly sampling the transformed data and fitting the analysis model, one can record fit index values, obtain a distribution for each fit index, and generate cutoffs.

In the following, we explain the bootstrap approach according to Bollen and Stine (1992) and Kim and Millsap (2014). Their bootstrap approach transforms each observation in the empirical data using the data-based and model-implied covariance and mean structure (see also Yung & Bentler, 1996). After the transformation, one obtains data that behaves as if the analysis model has generated it. The algorithm repeatedly samples the transformed data (with replacement), fits the analysis model to each resampled data set, and records the values of fit indices for each fitted model. The bootstrap method outlined above allows evaluating Type I error rates (i.e., incorrectly rejecting a correctly specified model) for cutoffs that correspond to a certain percentile of the resulting fit index distribution—like in the simulation-based approach. Yuan and Hayashi (2003), as well as Yuan et al. (2004; 2007), developed an extended bootstrap approach that also allows investigating power (i.e., correctly rejecting a misspecified model—the complement of the Type II error rate).

The bootstrap approach is highly flexible, similar to the simulation-based approach. Through repeated sampling, users can generate cutoffs for all available fit indices tailored to all choice characteristics. Akin to the simulation-based approach, this comes at the expense of greater computational intensity than required for the χ^2 distribution- and regression-based approaches.

Limitations of the Existing Approaches

All four approaches to tailored cutoffs have their merits and constitute a clear advancement over fixed cutoffs. Some approaches have an advantage in terms of computational speed in arriving at tailored cutoffs (i.e., the χ^2 distribution-based and regression-based, both parametric). Other approaches stand out as they are very general and generate cutoffs for a wide range of fit indices across a wide range of characteristics (i.e., the parametric simulation-based and non-parametric bootstrap).

However, these approaches also have specific limitations (see Table 1). One limitation they share is that they do not assess which fit index (among several fit indices a researcher may consider) is best able to discriminate between correctly specified and misspecified models in the setting of interest. The existing approaches do not guide researchers on which fit indices they should rely on for judging model fit. Such guidance on how much weight to assign to each fit index is especially needed when fit index decisions on model fit disagree, which often occurs in practice (e.g., Lai & Green, 2016; Moshagen & Auerswald, 2018).

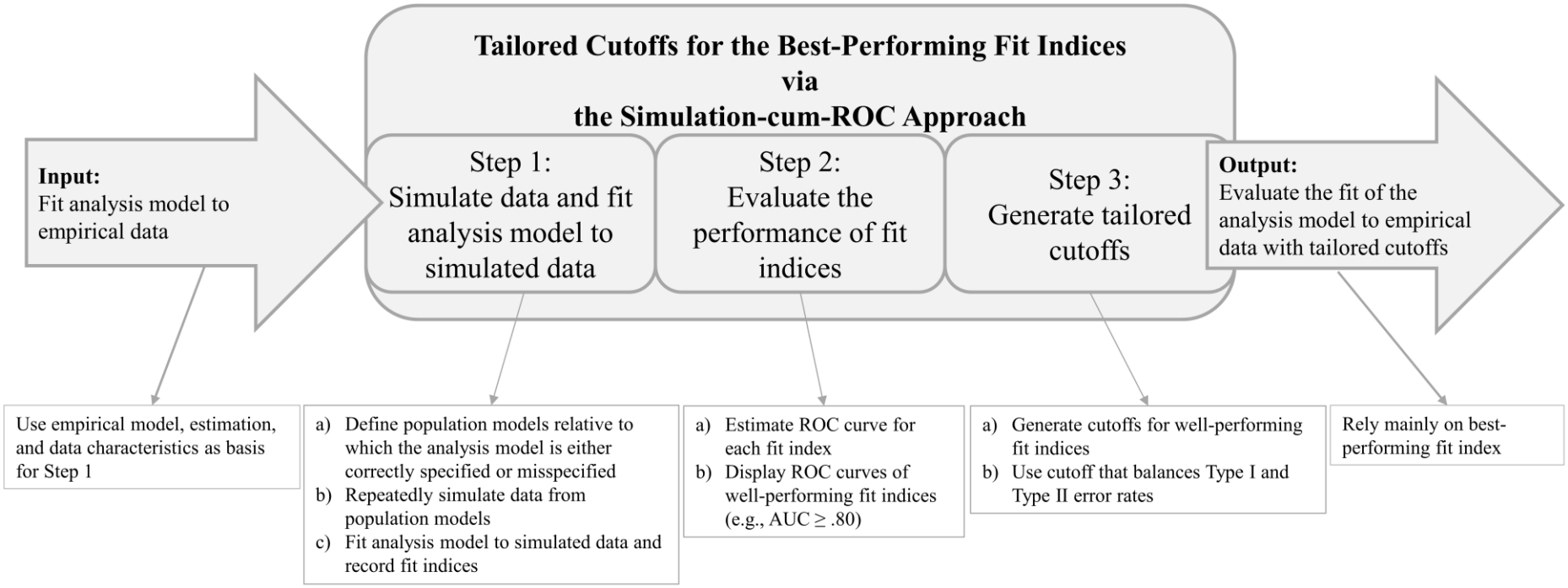
We, therefore, introduce a novel approach that builds on previous approaches and extends them by (1) identifying well-performing fit indices in a specific setting of interest while (2) generating tailored cutoffs that control both Type I and Type II error rates. This new approach is both general and adaptable enough to support valid judgments of model fit across the various settings that researchers may encounter.

A Novel Approach to Tailored Cutoffs: The Simulation-cum-ROC Approach

Our novel approach augments the simulation-based approach (e.g., McNeish & Wolf, 2021; Millsap, 2013; Pornprasertmanit, 2014) that is currently most well-known among applied researchers and has a long tradition for generating cutoffs (dating back to the initial Hu & Bentler, 1999, article). The unique contribution of our approach is to combine the simulation-based approach with a receiver operating characteristic (ROC) analysis. The so-called simulation-cum-ROC approach enables us to (1) rank the performance of any fit index in the setting of interest, including—but not limited to—the canonical fit indices on which we focus in this study (i.e., CFI, RMSEA, SRMR, but also χ^2). Further, it enables us to (2) generate tailored cutoffs at balanced Type I and Type II error rates for well-performing fit indices. Our approach thus allows for a more informative and rigorous evaluation of model fit.

In a nutshell, our approach works as follows. First, we use a Monte Carlo simulation to generate data from two population models that encode different assumptions about what models may have generated the data. One population model is structurally identical to the analysis model (such that the analysis model is correctly specified relative to the population model; H_0). The other population model diverges from that analysis model (such that the analysis model is misspecified relative to the population model; H_1). We fit the analysis model to data simulated from the two population models and record the fit index values. Second, we analyze the fit index distributions with ROC analysis. ROC analysis equips researchers with a tool to rank fit indices in terms of their ability to discriminate between correctly specified and misspecified models. Third, we generate cutoffs for well-performing fit indices. These cutoffs balance Type I and Type II error rates. We visualized the three steps to arrive at tailored cutoffs for well-performing fit indices in Figure 1.

Figure 1: *The Simulation-cum-ROC Approach*



Fundamentals of ROC Analysis

Before outlining the details of our approach, we briefly introduce ROC analysis. We base the introduction of ROC analysis on Flach (2016) and Padgett and Morgan (2021). Flach (2016) provided a general description of ROC analysis, and Padgett and Morgan (2021) connected ROC analysis to model fit evaluation.

ROC analysis originated within the context of signal detection theory in communication technology (for a detailed overview of the history of ROC analysis and signal detection theory, see Wixted, 2020). It provides a tool to evaluate the ability of a binary classifier to make correct diagnostic decisions in diverse scenarios, such as hypothesis testing. ROC analysis finds the optimal value for a classifier in making a diagnostic decision, such as classifying an analysis model as correctly specified or misspecified. It has supported decision-making in medicine for many decades and gained wide popularity in machine learning (for an overview, see Majnik & Bosnić, 2013).

Fit indices are, in essence, continuous classifiers that generally suggest better fit for correctly specified and worse fit for misspecified models. Cutoffs for these fit indices act as decision thresholds. These cutoffs should be chosen so that a large share of analysis models is correctly classified as correctly specified or misspecified.

Cutoffs for fit indices have a high sensitivity (i.e., true positive rate) if they classify a high share of misspecified models as misspecified (i.e., true positive) and only a small share of misspecified models as correctly specified (i.e., false negative, Type II error). In turn, cutoffs for fit indices have a high specificity (i.e., true negative rate) if they classify a high share of correctly specified models as correctly specified (i.e., true negative) and only a small share of correctly specified models as misspecified (i.e., false positive, Type I error). The formulae to calculate sensitivity and specificity read as

$$\textit{Sensitivity (or True Positive Rate)} = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives + Number\ of\ False\ Negatives}; \quad (1)$$

$$\textit{Specificity (or True Negative Rate)} = \frac{Number\ of\ True\ Negatives}{Number\ of\ True\ Negatives + Number\ of\ False\ Positives}. \quad (2)$$

The goal is to find a cutoff for each fit index that provides an optimal balance between sensitivity and specificity (i.e., which maximizes the sum of sensitivity and specificity – 1, the so-called Youden index). Such an optimal cutoff has a high accuracy, which means that the share of true positives and true negatives is large among all classified cases:

$$\textit{Accuracy} = \frac{Number\ of\ True\ Positives + Number\ of\ True\ Negatives}{Number\ of\ True\ Positives + Number\ of\ True\ Negatives + Number\ of\ False\ Positives + Number\ of\ False\ Negatives} \quad (3)$$

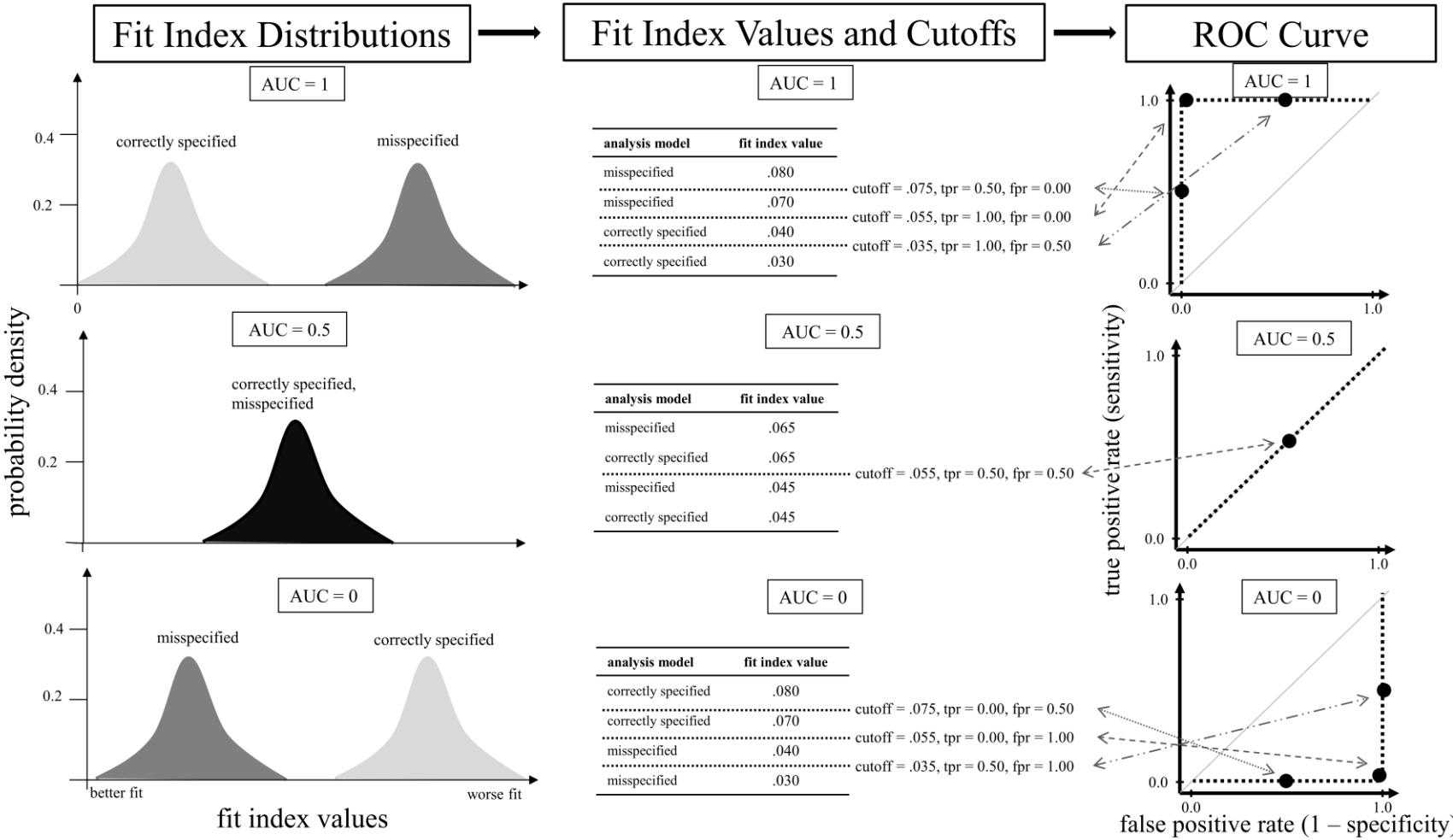
A so-called ROC curve visualizes the sensitivity and specificity at different cutoffs. These cutoffs may be generated arbitrarily (within the range of fit index values, e.g., Flach, 2016), or the actual fit index values are taken as cutoffs (as we do here, following Thiele & Hirschfeld, 2021). The graph visualizing the ROC curve has the sensitivity (or true positive rate) on its Y-axis and 1 − specificity (or false positive rate) on its X-axis. The area under the curve (AUC) quantifies the information of the ROC curve. We visualized the relationship between the distributions of a fit index, true and false positive rates of cutoffs, the ROC curve, and AUC values in Figure 2.

The AUC ranges between 0 and 1. It indicates the discrimination ability of a fit index at different cutoffs. An AUC of 1 is most favorable; it implies that all cutoffs have a true positive rate of 1 or a false negative rate of 0. Thus, 100% of the time, the fit index correctly discriminates between correctly specified and misspecified models (e.g., D'Agostino et al., 2013). The optimal cutoff, with the optimal balance between sensitivity and specificity, has a true positive rate of 1 and a false negative rate of 0. The ROC curve peaks in the upper left of the graph. Fit index distributions from correctly specified and misspecified models do not overlap and behave as expected (see Figure 2).

An AUC of 0.5 can imply different things, but most importantly, it can imply that all cutoffs have equal true and false positive rates. The discrimination ability of the fit index at different cutoffs is no better than a guess (e.g., D'Agostino et al., 2013). No optimal cutoff can be identified. In this case, the ROC curve is an ascending diagonal. Fit index distributions from correctly specified and misspecified models completely overlap; no distinction is possible (see Figure 2).

An AUC of 0 implies that all cutoffs have a true positive rate of 0 or a false positive rate of 1. The fit index has no discrimination ability at all at different cutoffs. An optimal cutoff cannot be identified. The ROC curve peaks in the lower right of the graph. Fit index distributions do not overlap; however, fit index values from correctly specified models behave unexpectedly and indicate worse fit than those from misspecified models (see Figure 2).

Figure 2: *Relation of Fit Index Distributions, Cutoffs, and the ROC Curve for Different AUC Values*



Note. The figure shows fit index distributions and a sample of fit index values from those distributions. Higher fit index values indicate worse fit here. We further estimated true and false positive rates of cutoffs based on the sample of fit index values. The ROC curve visualizes the true and false positive rates of cutoffs. The interplay of fit index distributions, true and false positive rates, and the ROC curve differ across AUC values. tpr = true positive rate; fpr = false positive rate.

Overall, the outlined relations indicate that the AUC quantifies what the ROC curve visualizes, namely, the performance of fit indices in terms of true and false positive rates at different cutoffs. The optimal cutoff is the one that has the highest sum of sensitivity (i.e., true positive rate) and specificity (i.e., 1 − false positive rate) across all evaluated cutoffs. Thus, the optimal cutoff shows up as a peak in the upper left of the graph (i.e., highest true positive rate and lowest false positive rate).

Combining Monte Carlo Simulation with ROC Analysis to Generate Tailored Cutoffs for Fit Indices

Having reviewed the basics of ROC analysis, we now detail our novel approach to evaluating the performance of fit indices and generating tailored cutoffs. We walk the reader through each step of the procedure shown in Figure 1.

Input: Fit Analysis Model to Empirical Data

Suppose we want to test whether a six-item scale measures a single underlying factor as its theory proposes. Survey data, including 500 participants' responses to the six items of the scale, form the basis for our empirical test of the model. We fit our to-be-tested analysis model of interest—a one-factor CFA model—to these data using robust maximum likelihood (MLR). We aim to test two hypotheses. The H_0 states that a population model identical to the analysis model (i.e., a one-factor model) has generated the data; if empirical evidence favors the H_0, we want to accept this analysis model. The H_1 states that an alternative population model different from the analysis model has generated the data (i.e., the population model has more parameters than the analysis model); if empirical evidence favors the H_1, we want to reject the analysis model. Thus, we define two diverging states of the world that describe how the data may hypothetically have come about (i.e., H_0 and H_1), and we can find evidence in favor of one or the other. To test the two hypotheses, we want to compare the empirical values of fit indices, obtained when fitting the analysis model to empirical data, against cutoffs tailored to the specific characteristics of our empirical setting. We obtain these cutoffs through the following three steps.

Step 1: Simulate Data and Fit Analysis Model to Simulated Data

In the first step, we conduct a Monte Carlo simulation closely designed to mimic the real empirical setting in terms of the model of interest (e.g., number of items, the magnitude of

factor loadings), the estimator (e.g., MLR), and the data characteristics (e.g., N = 500, multivariate distribution).

In keeping with the Neyman-Pearson approach, we operationalize the two competing hypotheses, H_0 versus H_1, about population models that may have generated the data. We simulate distributions of fit indices for our analysis model (i.e., the model of interest that we seek to test) through a Monte Carlo simulation under these two hypotheses. We encourage researchers to provide a strong rationale for their hypotheses and models. Providing a strong rationale is in line with recent calls for more rigorous theory testing in psychology, formalized theories, and preregistration (e.g., Borsboom et al., 2021; Fried, 2020; Guest & Martin, 2021).

We then simulate data from an H_0 population model structurally identical to the analysis model (i.e., a one-factor CFA model). We also simulate data from an H_1 population model that diverges substantially from the analysis model. For example, an H_1 population model could have two factors, whereas the H_0 population model (and the analysis model of interest) have one factor. No prior assumptions need to be made about the strength of misspecification, that is, how strongly the analysis model diverges from the H_1 population model. Notably, neither model needs to be nested (i.e., analysis and population models alike do not need to represent a subspace of each other), meaning that our approach is very flexible regarding model definition.[1]

After repeatedly simulating data from the H_0 and H_1 population models (e.g., 500 times each), we fit the one-factor analysis model to all simulated data and record the values of the fit indices. We obtain *distributions* of fit index values for correctly specified models (under the H_0) and misspecified models (under the H_1).

Step 2: Evaluate the Performance of Fit Indices

After simulating data and obtaining fit index distributions, we evaluate the performance of fit indices on the simulated data via the ROC curve and the AUC in particular. Both reflect the balance of a fit index between the true positive rate, or sensitivity, and the false positive rate, or 1 – specificity, at certain cutoffs. Fit indices with an AUC closer to 1 perform better; they have higher sensitivity and specificity across cutoffs. An AUC closer to 1 demonstrates a good ability of a fit index to discriminate between correctly specified and misspecified models.

[1] The simulation-cum-ROC requires two different population models representing hypothetical scenarios on how the data might have come about, encoded in H_0 and H_1 (following the Neyman-Pearson approach). But these population models are not being compared in the same way that researchers would compare competing analysis models in their empirical data; instead, these populations models are just a "crutch" needed for generating cutoffs. These cutoffs are, in turn, used to make a statistical decision between the H_0 and H_1 for the analysis model tested in empirical data.

In the following, we consider only those fit indices that reach an AUC of at least .80 or higher, which aligns with earlier work (Padgett & Morgan, 2021). An AUC of .80 implies that 80% of the time, the fit index correctly discriminates between correctly specified and misspecified models at the different potential cutoffs (e.g., D'Agostino et al., 2013). Notably, this AUC threshold of .80 is not a universally valid one. We use it for illustrative purposes here. Depending on the specific application, a researcher may choose higher (stricter) or lower (more lenient) AUC values—especially as Type I and Type II error rates of the corresponding cutoffs can exceed conventional levels of 5% at such an AUC threshold. We return to this point in the Discussion. The key point to understand here is that ROC analysis, particularly the AUC, offers a highly informative tool to evaluate the performance of fit indices in the given setting. Hence, it provides guidance regarding which fit index (or indices) are best to judge the model's fit.

Although we recommend focusing on high-performing fit indices with an AUC above a certain threshold (e.g., .80), it can be informative to inspect the distributions of low-performing fit indices as well and, if desired, also consider low-performing fit indices in judging the model fit. This is because different fit indices quantify different model, estimation, and data aspects (for an overview, see Schermelleh-Engel et al., 2003). For example, the χ^2 test statistic (e.g., Bollen, 1989) quantifies the discrepancy between model-implied and sample-based variance-covariance matrix (with RMSEA being a transformation of it; Steiger, 1990). CFI indicates how well the model reproduces the sample-based variance-covariance matrix compared to a model where all items are uncorrelated (Bentler, 1990). SRMR quantifies the average residuals between model-implied and sample-based covariance matrices (Bentler, 1995). The distributions' shape and overlap for each fit index help diagnose models further—as fit indices characterize models differently (see Browne et al., 2002; Lai & Green, 2016; Moshagen & Auerswald, 2018).

Thus, even strongly overlapping distributions (i.e., AUC around .50) may provide important insights. Such strongly overlapping fit index distributions imply that a fit index cannot distinguish between correctly specified and misspecified models. The misspecification of the analysis model relative to the H_1 population model as quantified through the fit index might not be strong enough. Alternatively, the analysis model can flexibly account for data from both population models. Flexible (i.e., more complex) models are weaker than inflexible (i.e., less complex) ones, as flexible models fit a wide range of data (e.g., MacCallum, 2003).

Step 3: Generate Tailored Cutoffs

After identifying well-performing fit indices (e.g., AUC ≥ .80) and screening out the others, we can identify optimal cutoffs. For each fit index, ROC analysis selects an optimal cutoff at the highest sum of sensitivity and specificity and, thus, the highest accuracy. At those cutoffs, the fit indices can best classify correctly specified models as correctly specified and misspecified ones as misspecified.

We provide cutoffs along with their accuracy, Type I error rate (i.e., 1 – specificity), and Type II error rate (i.e., 1 – sensitivity). Generally, the cutoff with the highest accuracy across fit indices belongs to the best-performing fit index (i.e., the one with the highest AUC).[1] In addition, we visualize the fit index distributions from correctly specified and misspecified models. An essential strength of the simulation-cum-ROC approach is that it returns the error probabilities associated with applying a set of cutoffs. It draws researchers' attention to how well cutoffs can discriminate between correctly specified and misspecified models in the context of interest (quantified through Type I and Type II error rates).

Output: Evaluate the Fit of the Analysis Model to Empirical Data with Tailored Cutoffs

Having generated tailored cutoffs for well-performing fit indices, we can evaluate how well our analysis model (i.e., a one-factor model in our example) fits the empirical data by comparing the empirical values of the fit indices against the tailored cutoffs. In doing so, three scenarios may occur: (a) all fit indices point to good model fit, (b) all fit indices point to bad model fit, or (c) some fit indices point to good and some to bad model fit.

If all empirical values of fit indices pass the proposed tailored cutoffs, then the analysis model has a good fit. The evidence unequivocally favors the H_0 instead of the H_1, and we can accept the analysis model. If all empirical values of fit indices fail the proposed tailored cutoffs, the analysis model has poor fit. The evidence favors the H_1 instead of the H_0, and we need to reject the analysis model.

There could be less straightforward empirical settings where the fit indices disagree (i.e., some pass, but others fail their respective cutoffs). In such cases, we can leverage the information from the ROC curve about the performance of fit indices. If there is a best-performing fit index and its empirical value suggests that the analysis model fits (i.e., passes

[1] Exceptions may occur where the fit index with the highest AUC does not have the cutoff with the highest accuracy. For instance, the fit index with the highest AUC does not need to have the cutoff with the highest accuracy if AUC values of different fit indices are only marginally different from each other.

its tailored cutoff), the evidence favors the H_0 instead of the H_1. We accept the analysis model. If the best-performing fit index suggests that the analysis model does not fit, the evidence favors the H_1 instead of the H_0. We reject the analysis model. Thus, in those less-straightforward scenarios, we prioritize the best-performing fit index and its corresponding cutoff for our decision on model fit.

If we reject the analysis model, we might want to modify it to find a better-fitting alternative. Modification indices help identify local misfit, though theory should also guide model modification (Fried, 2020). If theoretical and empirical indications lead to alterations of the analysis model, we need to test the modified model again. We need to repeat the above procedure (Steps 1 to 3) once we state a new H_0 and H_1. We always state a new H_0 and H_1 when we modify a model and test it again.

Application of the Simulation-cum-ROC Approach

In the following, we provide two examples that illustrate the simulation-cum-ROC approach. The main aim of the first example is to walk the reader through the three steps to generate and apply tailored cutoffs. For that reason, we chose a simple example without further complications. In this example, an alternative model, which can serve as an H_1 population model, was already proposed in the literature. All fit indices performed equally well in this example. Neither is always guaranteed to be the case in empirical applications.

Thus, the main aim of the second example is to showcase the potential of the simulation-cum-ROC approach in ranking the performance of fit indices. In this example, an alternative model, which can serve as an H_1 population model, has not already been proposed in the literature. Further, the fit indices of interest differed in their performance in this example; not all fit indices performed well enough to be useful for model evaluation.

We used publicly available secondary data for both examples (Nießen et al., 2018, 2020). We conducted all analyses with R (version 4.1.1; R Core Team, 2021). We employed the R package lavaan to fit the models (version 0.6.9; Rosseel, 2012), simsem to simulate the data (version 0.5.16; Pornprasertmanit et al., 2021), pROC to plot the ROC curves (version 1.18.0; Robin et al., 2011), and cutpointr to obtain cutoffs for fit indices (version 1.1.1; Thiele & Hirschfeld, 2021). We documented all other used packages in the R code. Additional File 1 of the Supplementary Material includes the computational code. We did not preregister the study.

We also programmed a shiny app which is available under https://kg11.shinyapps.io/tailoredcutoffs/. Specifically, one needs to plug in their analysis model, population models, marginal skewness and excess kurtosis of each item's response distribution (used to obtain multivariate non-normal data with Vale and Maurelli's method, 1983[1]), estimator, sample size, number of simulation runs, fit indices one is interested in, and AUC cutpoint. The shiny app internally runs through Steps 1 to 3 of the simulation-cum-ROC approach. It allows downloading the ROC curves from Step 2 as well as the fit index distributions and tailored cutoffs from Step 3. Users do not need to execute any statistical program locally; the shiny app does all the computational work to arrive at tailored cutoffs within the simulation-cum-ROC approach.

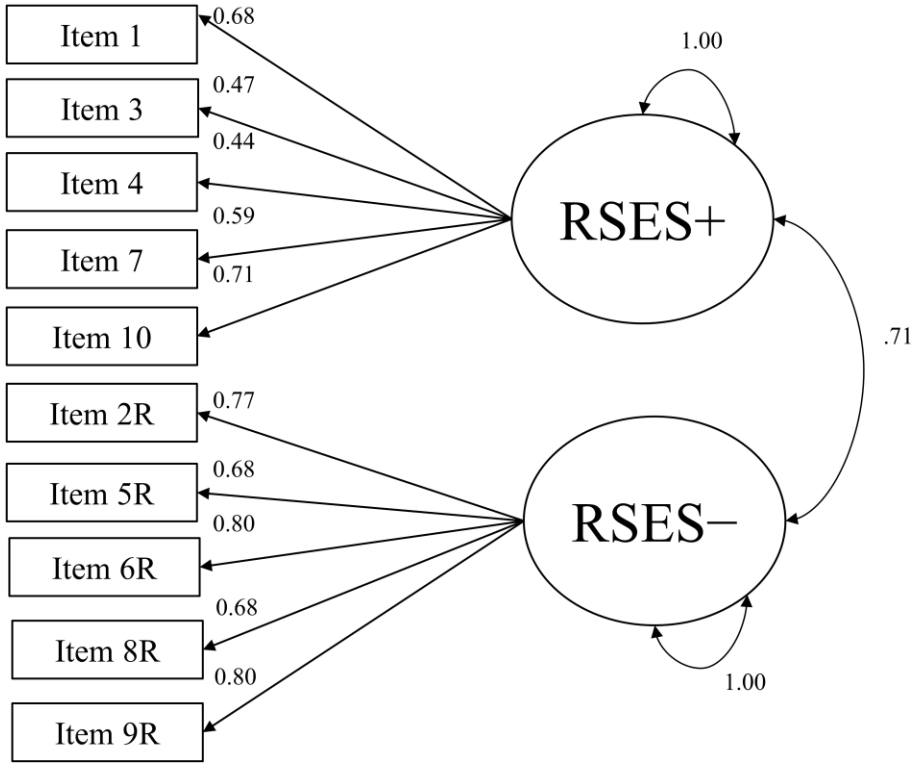
Example 1: The Rosenberg Self-Esteem Scale

We chose the Rosenberg Self-Esteem Scale for the first example of generating tailored cutoffs via the simulation-cum-ROC approach (Rosenberg, 1965). This scale measures global self-esteem with ten items (five referring to positive feelings about the self and five to negative ones) rated on a four-point rating scale. Initially constructed with a single factor, later studies found evidence for a two-factor structure (or even more complex structures, see Supple et al., 2013, for an overview). We used publicly available data (N = 468; Nießen et al., 2020) that contains the Rosenberg Self-Esteem Scale applied to a quota sample of adults aged 18 to 69 from the United Kingdom.

Input: Fit Analysis Model to Empirical Data. We fit the two-factor model to the empirical data using MLR. Figure 3 depicts the two-factor model and the empirical fit index values. Here, we evaluated whether empirical evidence favored the H_0 or H_1 for the two-factor model using tailored cutoffs. We would accept the two-factor model if empirical evidence favored the H_0 stating that the two-factor model was identical to the population model. We would reject the two-factor model if empirical evidence favored the H_1 stating that the two-factor model was not identical to the population model to an intolerable degree of misspecification.

[1] Olvera Astivia and Zumbo (2015) have shown that estimates of skewness and kurtosis are downward-biased using Vale and Maurelli's method, especially in small samples. Because we employed the simsem package (Pornprasertmanit et al., 2021) and no alternative method was implemented there, we relied on Vale and Maurelli's method to obtain multivariate non-normal data.

Figure 3: *Empirical Two-Factor Rosenberg Self-Esteem Scale Model*



Fit indices	χ^2	*df*	CFI	RMSEA	SRMR
Empirical values	119.05***	34	.947	.073	.051

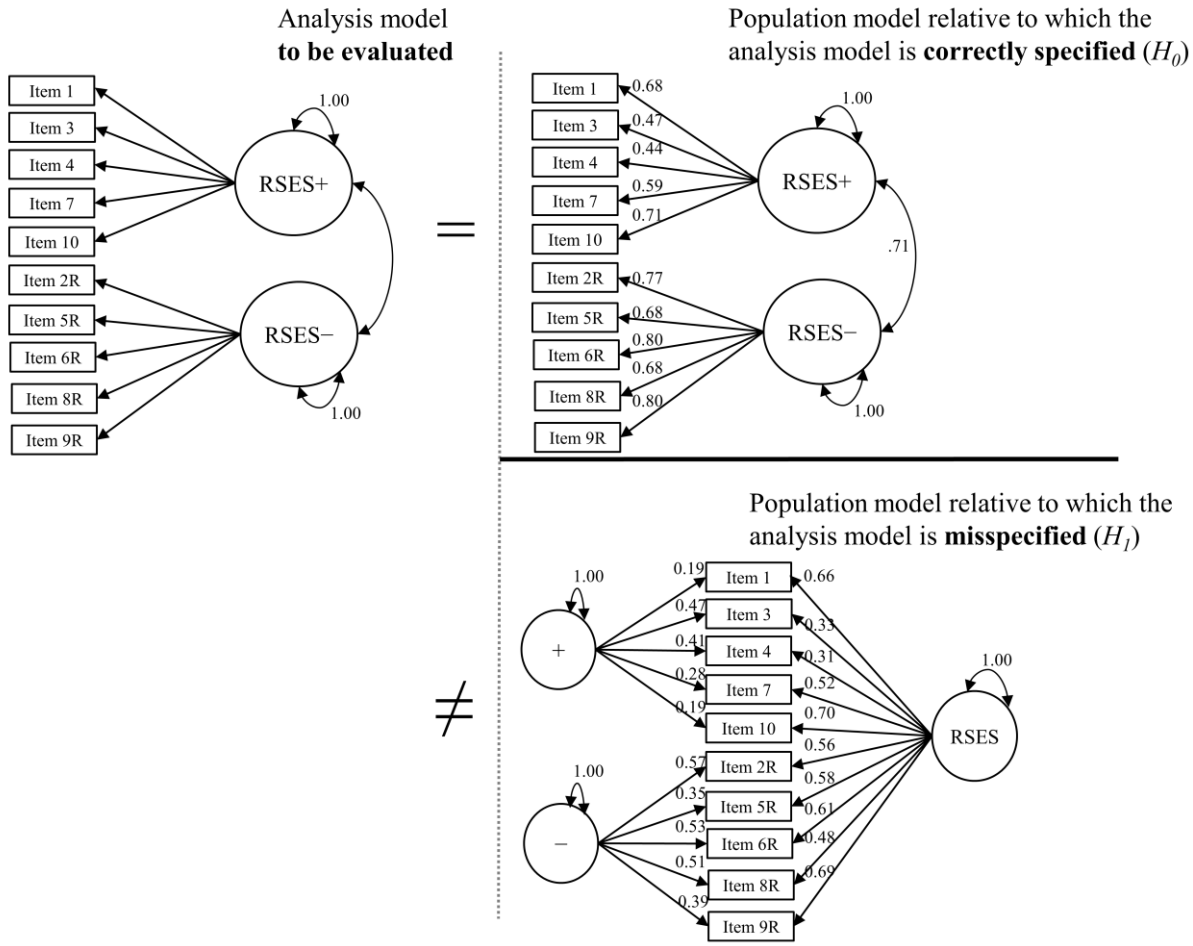
Note. Unstandardized coefficients. RSES = Rosenberg Self-Esteem Scale. We recoded the items so that higher values imply higher self-esteem. We omitted the residual variances and the mean structure for clarity. $N = 468$. *** $p < .001$.

Step 1: Simulate Data and Fit Analysis Model to Simulated Data. After fitting the two-factor model to empirical data, we operationalized H_0 and H_1 as a basis for the Monte Carlo simulation. The two-factor model served as an analysis model in the simulation. The structure and parameter estimates of the two-factor model fit to empirical data served as the H_0 population model. Relative to the two-factor population model, the two-factor analysis model was correctly specified.

As an H_1 population model, we chose a bi-factor model proposed in the literature on the Rosenberg Self-Esteem Scale (for an overview, see Supple et al., 2013). The structure and parameter estimates of a bi-factor model fit to empirical data served as the H_1 population model. Relative to the bi-factor population model, the two-factor analysis model was severely underspecified (i.e., misspecified). Figure 4 shows the population and analysis models.

We simulated data from the H_0 and H_1 population models, fit the two-factor analysis model to that data, and recorded the fit index values. The Monte Carlo simulation closely resembled the empirical setting regarding the sample size (i.e., $N = 468$), the estimator of choice (i.e., MLR), and the multivariate response distribution. We simulated 500 data sets from each population model. Simulating the data, fitting the analysis model, and recording the fit indices took four to five minutes on a standard computer using R (single-threaded).

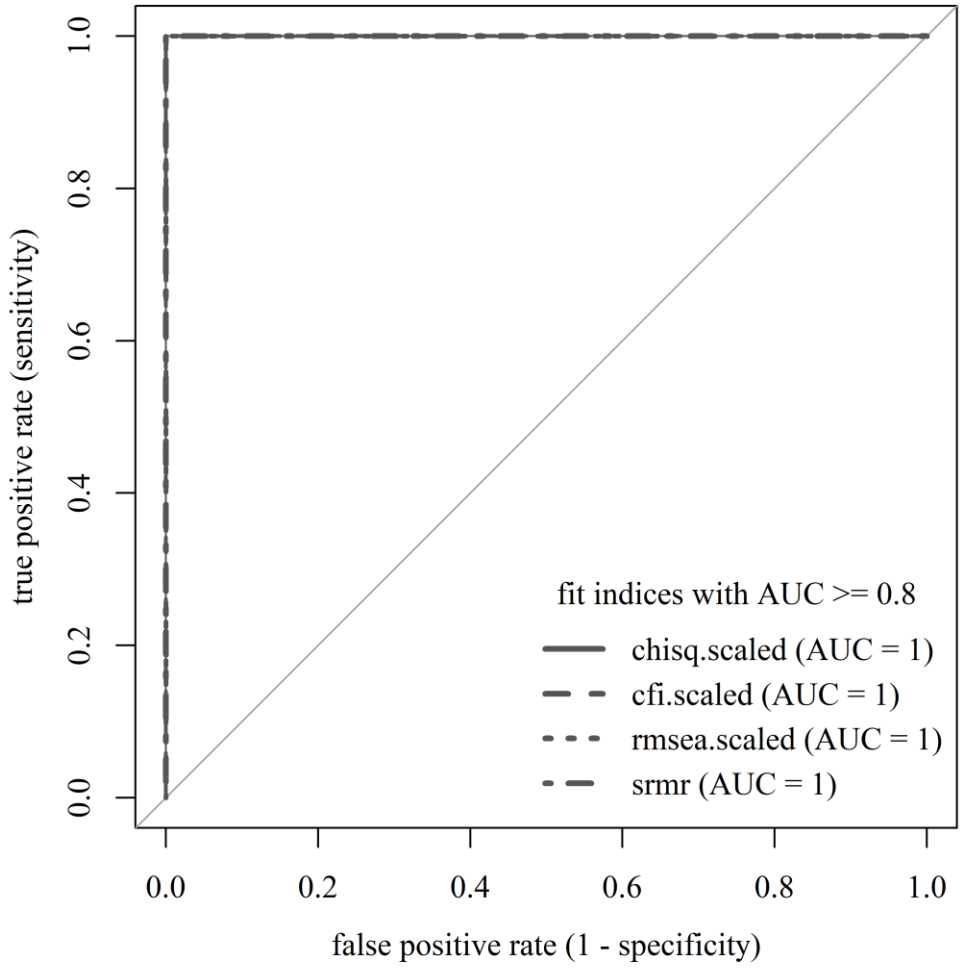
Figure 4: *Proposed Analysis and Population Models of the Rosenberg Self-Esteem Scale*



Note. We simulated data from both population models and fit the same analysis model to the data. Because the analysis model was structurally identical to the H_0 population model, it was correctly specified when fit to data generated from that population model. Because the analysis model differed from the H_1 population model, it was misspecified when fit to data generated from that population model. Unstandardized coefficients. RSES = Rosenberg Self-Esteem Scale. We recoded the items so that higher values imply higher self-esteem. We omitted the residual variances and the mean structure for clarity.

Step 2: Evaluate the Performance of Fit Indices. After simulating the data, we evaluated the performance of fit indices as quantified through the AUC. We stipulated to only consider fit indices with an AUC of .80 or higher (Padgett & Morgan, 2021) and disregarded all others. Figure 5 displays the ROC curves of the fit indices (in different line shapes). *All* fit indices had an AUC equal to or higher than .80, namely an AUC of 1. Therefore, all fit indices discriminated equally well between correctly specified and misspecified models.

Figure 5: *ROC Curves for Fit Indices with AUC ≥ .80 of the Rosenberg Self-Esteem Scale Model*

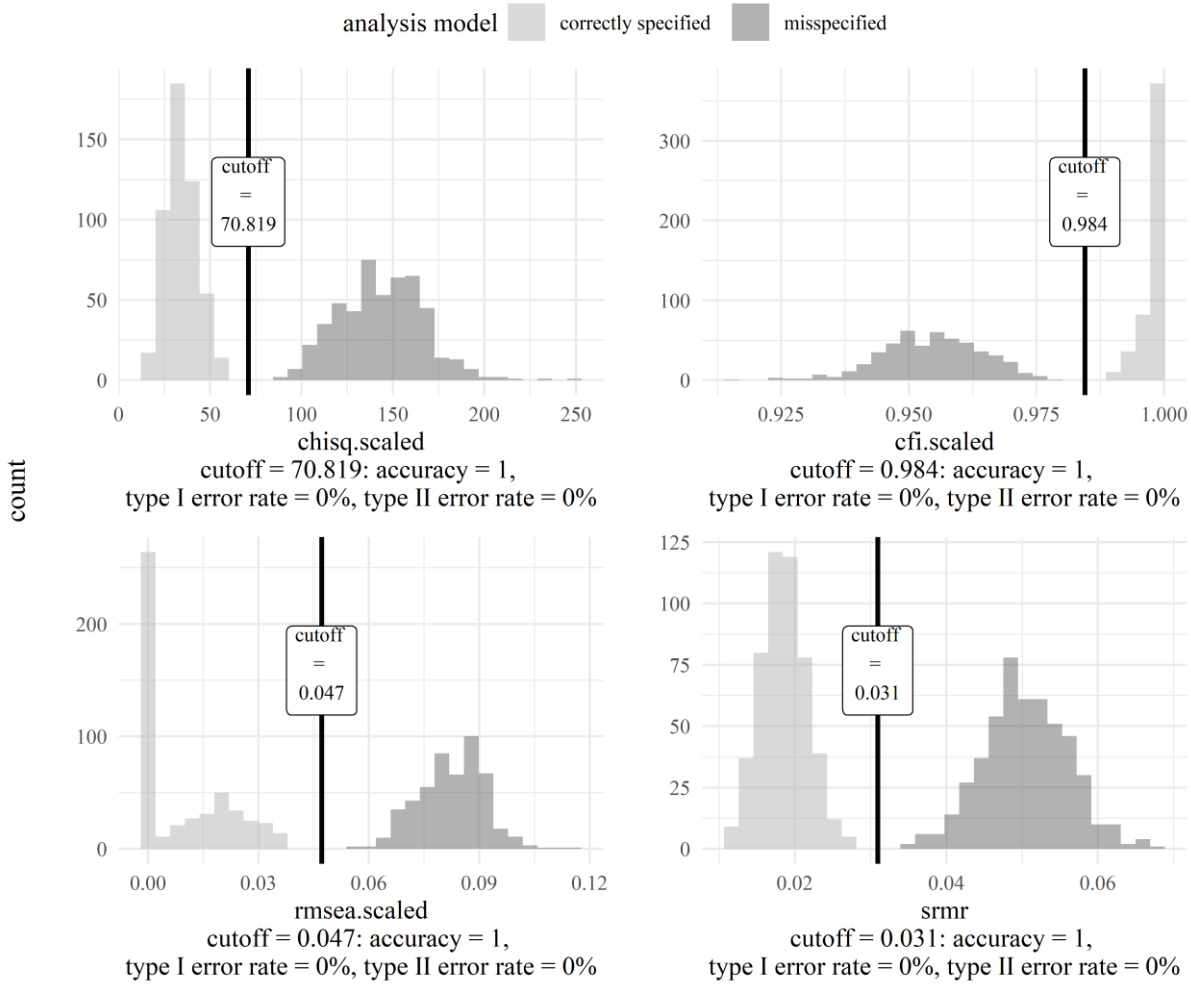


Note. Chisq.scaled is a χ^2 test statistic asymptotically equivalent to the robust Yuan-Bentler test statistic (Yuan & Bentler, 2000a) to account for non-normality. Cfi.scaled is the CFI version and rmsea.scaled is the RMSEA version calculated with this test statistic.

Step 3: Generate Tailored Cutoffs. In Step 3, we generated cutoffs for well-performing fit indices. All fit indices performed equally well (as quantified through the AUC). Thus, we generated tailored cutoffs for all fit indices. Figure 6 depicts the fit index distributions for the simulated data. The distribution colored in lighter gray is the one for fit index values from correctly specified models. The distribution colored in darker gray is the one for fit index values from misspecified models. The vertical dash corresponds to the cutoff (maximizing the sum of sensitivity and specificity $-$ 1).[1] The cutoffs were the following: $\chi^2(34) \leq 70.82$, CFI $\geq .984$, RMSEA $\leq .047$, SRMR $\leq .031$. All cutoffs across fit indices had an accuracy of 1. Type I and Type II error rates were zero for all cutoffs. Thus, all cutoffs perfectly discriminated between correctly specified and misspecified models in this setting.

[1] As evident from Figure 6, we let the algorithm take the mean of the optimal cutoff and the next highest fit index value as a revised optimal cutoff (or the next lowest fit index value when lower values imply worse fit, such as for CFI; Thiele & Hirschfeld, 2021). To find an optimal cutoff, the algorithm first uses each fit index value as a potential cutoff starting with those indicating good (e.g., CFI = 1.00, RMSEA = 0.00) to bad model fit (e.g., CFI = 0.00, RMSEA = 1.00) and evaluates sensitivity and specificity. Then it selects the fit index value with the highest sum of sensitivity and specificity as an optimal cutoff. For non-overlapping distributions, both the worst fit index value from correctly specified models and the best fit index value from misspecified models have the highest sum of sensitivity and specificity. The algorithm would, thus, choose the worst fit index value from correctly specified models as an optimal cutoff. It is the first value with the highest sum of sensitivity and specificity (because the algorithm starts from good to bad model fit). We let the algorithm take the mean between the optimal cutoff (i.e., the worst fit index value from correctly specified models in non-overlapping distributions) and the next value (i.e., the best fit index value from misspecified models in non-overlapping distributions) as a revised optimal cutoff to avoid bias in favor of correctly specified models.

Figure 6*: Cutoffs for Fit Indices with AUC ≥ .80 of the Rosenberg Self-Esteem Scale Model*



Note. Chisq.scaled is a χ^2 test statistic asymptotically equivalent to the robust Yuan-Bentler test statistic (Yuan & Bentler, 2000a) to account for non-normality. Cfi.scaled is the CFI version and rmsea.scaled is the RMSEA version calculated with this test statistic. The distribution colored in lighter gray originates from correctly specified models. The distribution colored in darker gray originates from misspecified models. Overlapping (parts of) distributions have an even darker gray color than the distribution from misspecified models. The vertical dash corresponds to the cutoff for each fit index (at the highest sum of sensitivity and specificity – 1).

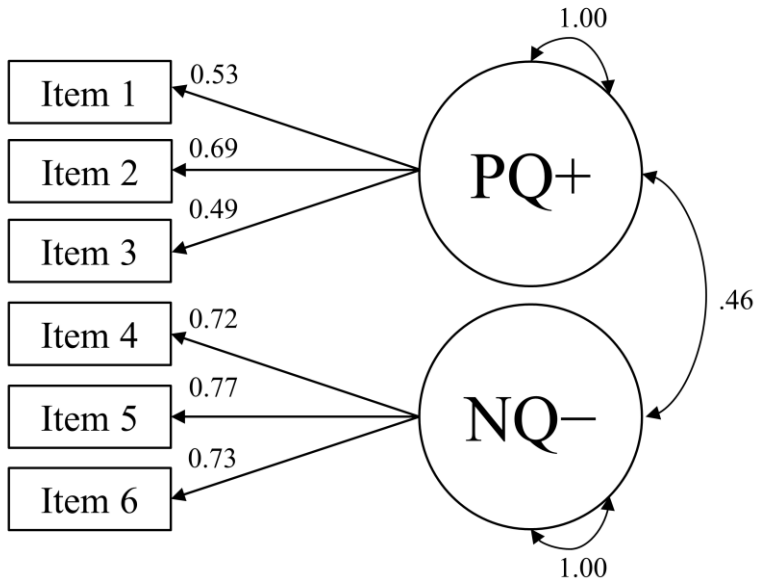
Output: Evaluate the Fit of the Analysis Model to Empirical Data with Tailored Cutoffs. Judged against the tailored cutoffs, we rejected the two-factor model for the Rosenberg Self-Esteem Scale fit to empirical data. None of the empirical fit index values for the two-factor model ($\chi^2(34) = 119.05$; CFI = .947; RMSEA = .073; SRMR = .051) passed the tailored cutoffs (i.e., $\chi^2(34) \leq 70.82$, CFI ≥ .984, RMSEA ≤ .047, SRMR ≤ .031). Evidence favored the H_1, stating that another (less restrictive) population model had generated the data. Interestingly, traditional fixed cutoffs of CFI around .950, RMSEA around .060, and SRMR around .080 (Hu & Bentler, 1999) would wrongly lead to accepting the two-factor model.

Example 2: The Social Desirability-Gamma Short Scale

To illustrate the potential of the simulation-cum-ROC approach, we took the Social Desirability-Gamma Short Scale (Kemper et al., 2014; Nießen et al., 2019) as a second example. Paulhus's (2002) theoretical model of socially desirable responding was the basis for this scale. Socially desirable responding refers to deliberate attempts to present oneself as a nice person or good citizen. The Social Desirability-Gamma Short Scale measures the two aspects of the Gamma factor of socially desirable responding: exaggerating one's positive qualities (PQ+) and minimizing one's negative qualities (NQ−) with three items each. Respondents rate these items on a five-point rating scale. Publicly available data (N = 474; Nießen et al., 2018) contains the German version of the scale applied to a quota sample of adults aged 18 to 69 years in Germany.

Input: Fit Analysis Model to Empirical Data. We fit the two-factor model of the Social Desirability-Gamma Short Scale to the empirical data using MLR (following Nießen et al., 2019). Figure 7 depicts the two-factor model and its empirical values of fit indices. Here, we evaluated whether empirical evidence favored the H_0 or H_1 for the two-factor model using tailored cutoffs. We would accept the two-factor model if evidence favored the H_0 stating that the two-factor model was identical to the population model. We would reject the two-factor model if empirical evidence favored the H_1 stating that the two-factor model was not identical to population model to an intolerable degree of misspecification.

Figure 7: *Empirical Two-Factor Social Desirability-Gamma Short Scale Model*



Fit indices	χ^2	*df*	CFI	RMSEA	SRMR
Empirical values	32.06***	8	.947	.080	.048

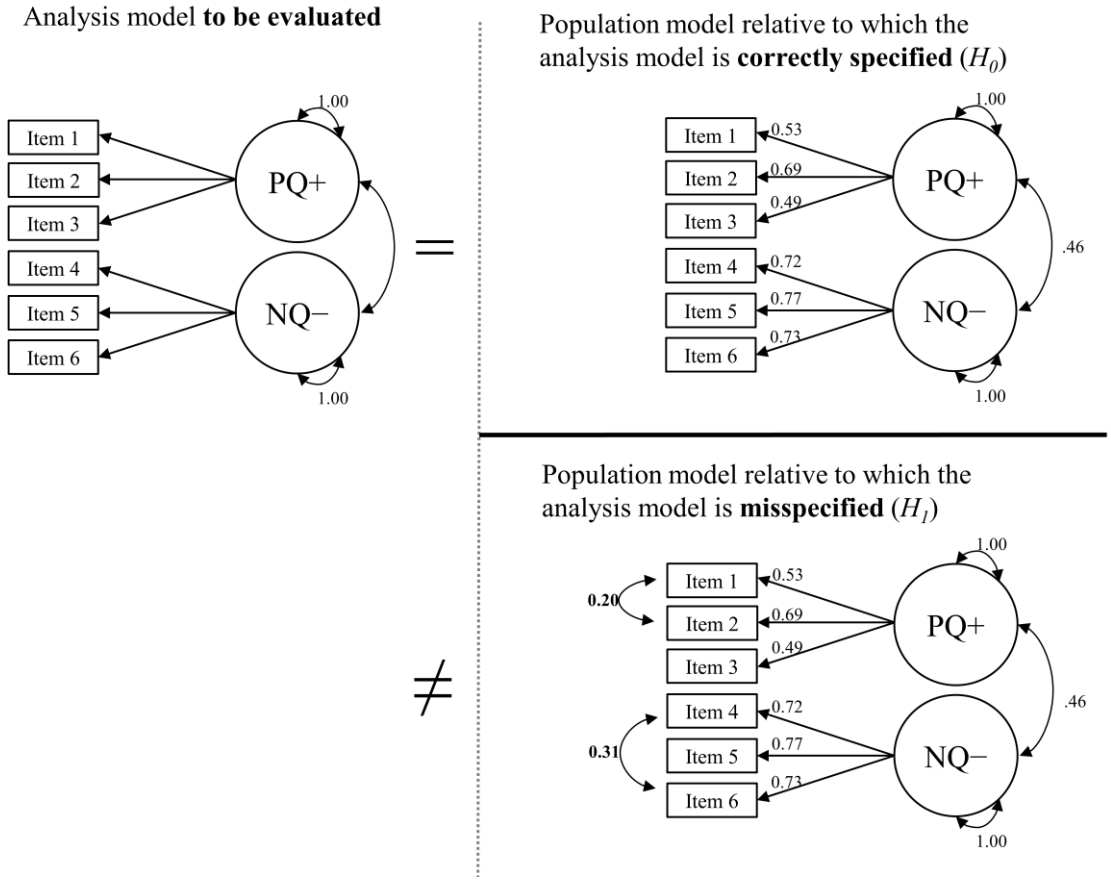
Note. Unstandardized coefficients. PQ+ = exaggerating positive qualities; NQ− = minimizing negative qualities. We recoded NQ− so that higher values imply more socially desirable responses. We omitted the residual variances and the mean structure for clarity. $N = 474$. *** $p < .001$.

Step 1: Simulate Data and Fit Analysis Model to Simulated Data. After fitting the two-factor model to empirical data, we operationalized H_0 and H_1 as a basis for the Monte Carlo simulation. The two-factor model served as an analysis model in the simulation. The structure and parameter estimates of the two-factor model fit to empirical data served as the H_0 population model. Relative to the H_0 population model, the two-factor analysis model was correctly specified.

Next, we needed to define a theoretically justifiable H_1 population model—which has not been suggested in the literature yet and was, thus, not as immediately apparent as in the previous example. A good candidate for an H_1 population model could be a two-factor model that contains additional residual covariances to capture shared wording effects. The question of whether additional residual covariances are needed to account for the covariances among items fully is one with which applied researchers frequently grapple (e.g., Bluemke et al., 2016; Podsakoff et al., 2003). Correlations of $r = .50$ have been considered large (Cohen, 1992). Two unmodeled residual correlations have been considered moderate misspecification for six-item models (McNeish & Wolf, 2021). We chose an H_1 population model that was identical to the

H_0 population model (and, thus, the analysis model) in the latent-variable part but comprised two residual correlations of r = .50 each. We modeled one residual correlation between the first and second item of the PQ+ factor (resulting in a residual covariance of 0.20), both of which ask for emotional control. We modeled another residual correlation between the first and third item of the NQ− factor (resulting in a residual covariance of 0.31), both of which refer to behavior in interactions. Relative to this H_1 population model, the two-factor analysis model was severely underspecified (i.e., misspecified). Figure 8 shows the population and analysis models for examining H_0 and H_1.

Figure 8: *Proposed Analysis and Population Models of the Social Desirability-Gamma Short Scale*

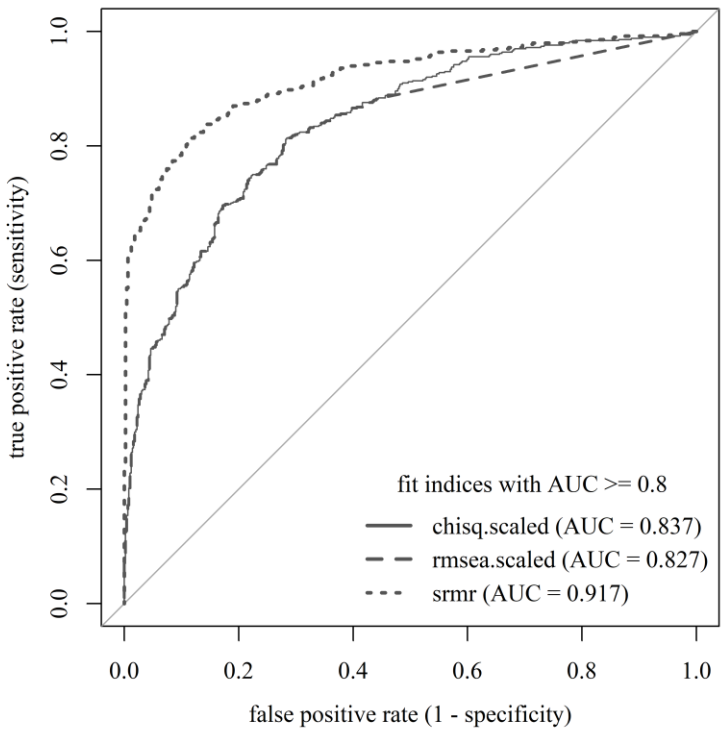


Note. We simulated data from both population models and fit the same analysis model to the data. Because the analysis model was structurally identical to the H_0 population model, it was correctly specified when fit to data generated from that population model. Because the analysis model differed from the H_1 population model, it was misspecified when fit to data generated from that population model. Unstandardized coefficients. PQ+ = exaggerating positive qualities; NQ− = minimizing negative qualities. We recoded NQ− so that higher values imply more socially desirable responses. We omitted the residual variances and the mean structure for clarity.

We simulated data from the population models, fit the analysis model to each simulated data set, and recorded the fit indices. Essential features of the simulation mimicked the empirical setting in terms of the sample size (i.e., N = 474), estimator (i.e., MLR), the multivariate response distribution, and the number of simulation runs (i.e., 500, which we recommend as a minimum). Simulating the data, fitting the analysis model, and recording the fit indices took two to three minutes on a standard computer using R (single-threaded).

Step 2: Evaluate the Performance of Fit Indices. In the following, we took a closer look at the performance of the fit indices. In contrast to the previous example, not all fit indices passed the AUC ≥ .80 benchmark, and the AUCs were generally lower. Figure 9 visualizes the ROC curves of three fit indices with an AUC of .80 or higher: χ^2, RMSEA, and SRMR. We disregarded CFI because, with an AUC below .80, it did not perform adequately in this setting. Among the three well-performing fit indices with AUC ≥ .80 (i.e., χ^2, RMSEA, and SRMR but not CFI), SRMR had the highest AUC (= .92) and was, thus, the best-performing one in the setting of interest.

Figure 9: *ROC Curves for Fit Indices with AUC ≥ .80 of the Social Desirability-Gamma Short Scale Model*



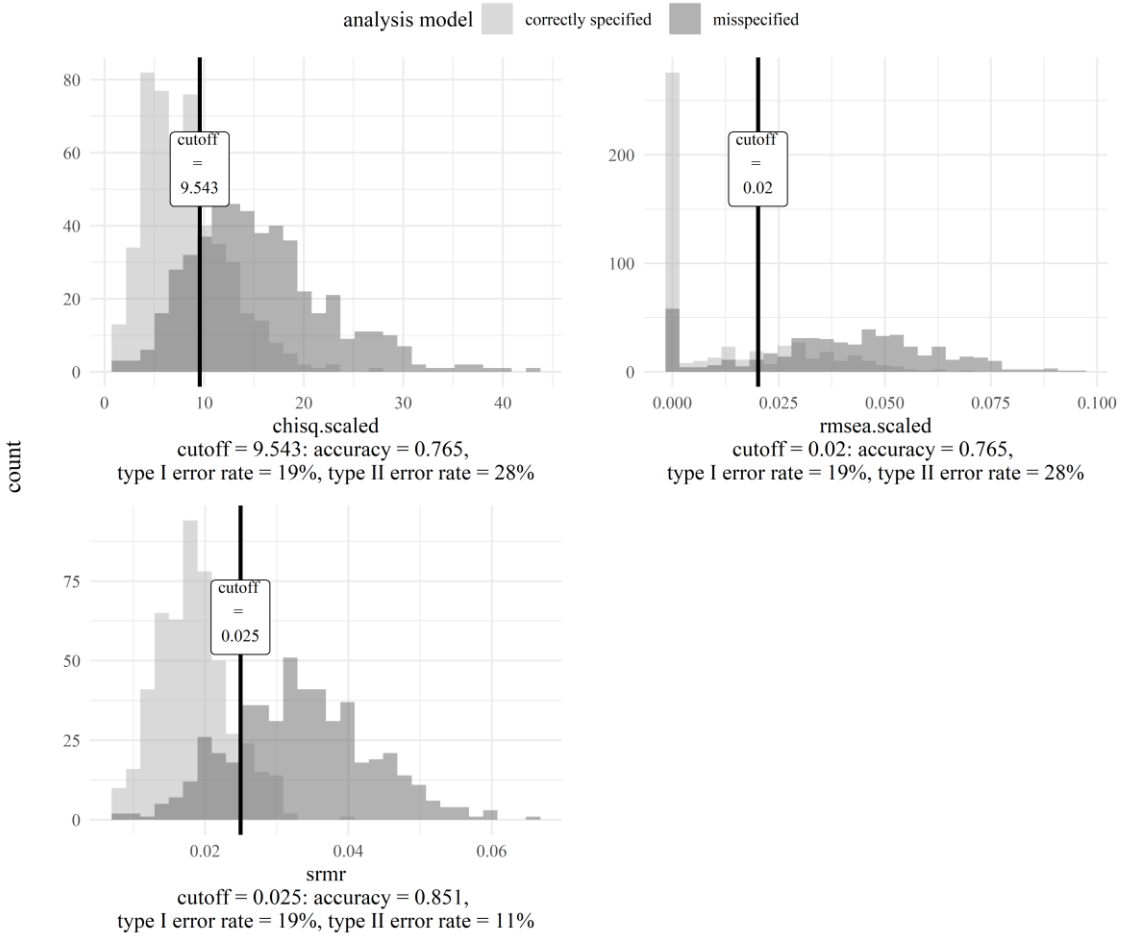
Note. Chisq.scaled is a χ^2 test statistic asymptotically equivalent to the robust Yuan-Bentler test statistic (Yuan & Bentler, 2000a) to account for non-normality. Rmsea.scaled is the RMSEA version calculated with this test statistic.

Step 3: Generate Tailored Cutoffs. We generated cutoffs only for the three well-performing fit indices in the following. The recommended cutoff for χ^2 was 9.54, for RMSEA .020, and for SRMR .025 (Figure 10). In line with the AUC, the cutoff for SRMR had the highest accuracy (= .851) and the lowest Type II error rate (= 11%). It better categorized misspecified models as misspecified than cutoffs for other fit indices. The Type I error rate was the same for all cutoffs (= 19%). Thus, the SRMR, with its corresponding cutoff, had the best ability to demarcate between correctly specified and misspecified models in the setting at hand. The greatest difference between correctly specified and misspecified models in the specific setting was due to average standardized residuals.

The reader may have noticed that these cutoffs' Type I and Type II error rates are above conventional levels of 5%. The strength of the simulation-cum-ROC approach is the thorough analysis of fit index distributions and the deliberate use of cutoffs. It makes us aware of the error probabilities involved. If we deem the error rates too high, we can redefine the H_1 population model. To redefine the H_1 population model, we need to repeat Steps 1 through 3 of the simulation-cum-ROC approach: In Step 1, we need to define a new H_1 population model, from which the analysis model is "further" away than the initial H_1 population model. For instance, the new H_1 population model contains more or higher non-zero parameter values than the initial H_1 population model, which the analysis model wrongly fixes to zero.

Alternatively, we can use the cutoffs while accepting their given error probability. In this example of the Social Desirability-Gamma Short Scale, we deemed the error probabilities acceptable (especially the ones of SRMR) because we explicitly wanted to retain the definitions of population models as outlined and justified. Imposing stronger misspecification through redefining the H_1 population model would lead to more lenient cutoffs than the current ones. This would imply that those cutoffs could lead to accepting an empirical model that contains misfit of a size that we initially deemed unacceptable (i.e., through the initial definition of the H_1 population model relative to which the analysis model is misspecified).

Figure 10: *Cutoffs for Fit Indices with AUC ≥ .80 of the Social Desirability-Gamma Short Scale Model*



Note. Chisq.scaled is a χ^2 test statistic asymptotically equivalent to the robust Yuan-Bentler test statistic (Yuan & Bentler, 2000a) to account for non-normality. Rmsea.scaled is the RMSEA version calculated with this test statistic. The distribution colored in lighter gray originates from correctly specified models. The distribution colored in darker gray originates from misspecified models. Overlapping (parts of) distributions have an even darker gray color than the distribution from misspecified models. The vertical dash corresponds to the cutoff for each fit index (at the highest sum of sensitivity and specificity – 1).

Output: Evaluate the Fit of the Model to Empirical Data with Tailored Cutoffs. When comparing the empirical fit index values to the cutoffs tailored to the setting of interest, we needed to reject the two-factor model of the Social Desirability-Gamma Short Scale. The empirical values of fit indices ($\chi^2(8) = 32.06$, $p < .001$; CFI = .947; RMSEA = .080; SRMR = .048) clearly failed the tailored cutoffs ($\chi^2(8) \leq 9.54$; CFI = should not be considered; RMSEA ≤ .020; SRMR ≤ .025). Thus, we found empirical evidence favoring the H_1 instead of the H_0, concluding that a model different from a two-factor one is likely to have generated the data. Notably, traditional fixed cutoffs of CFI around .950, RMSEA around .060, and SRMR around .080 (Hu & Bentler, 1999) would wrongly lead to accepting the two-factor model.

Discussion

Fixed cutoffs for fit indices are far more problematic than many researchers realize (e.g., Groskurth et al., 2022; Marsh et al., 2004; Lai & Green, 2016). Fixed cutoffs have low external validity and do not generalize well to settings not covered in simulation studies from which these cutoffs originate. This is because fit indices are susceptible to various influences other than model misspecification they should detect (for an overview, see Groskurth et al., 2022; Niemand & Mai, 2018; McNeish & Wolf, 2021, 2022; Pornprasertmanit, 2014). Cutoffs tailored to the setting of interest are generally more appropriate than fixed cutoffs whenever the setting falls outside the limited range of simulation scenarios from which these cutoffs were derived (such as those by Hu and Bentler, 1999). Therefore, methodologists are increasingly urging that fixed cutoffs should be abandoned and replaced by tailored (or "dynamic") cutoffs (e.g., Markland, 2007; Marsh et al., 2004; McNeish & Wolf, 2021; Niemand & Mai, 2018; Nye & Drasgow, 2011).

We reviewed four principal approaches to generating tailored cutoffs in the current study. Ours is the first study to review and systematize the approaches to tailored cutoffs comprehensively. While we have outlined their strengths and limitations on a conceptual level, future research may additionally want to compare their performance statistically. For example, simulation studies comparing the Type I and Type II error rates of cutoffs generated from the various approaches in different contexts have yet to be conducted.

We then introduced a novel approach, the simulation-cum-ROC approach, that augments the simulation-based approach to tailored cutoffs that has gained traction in the recent literature (e.g., McNeish & Wolf, 2021, 2022; Millsap, 2013; Niemand & Mai, 2018). By applying ROC analysis to distributions of fit indices from a Monte Carlo simulation, the simulation-cum-ROC approach provides a highly informative way to evaluate model fit. Like several other approaches outlined in our review, the simulation-cum-ROC approach generates (1) tailored cutoffs at balanced Type I and Type II error rates for several fit indices across various settings. However, it conceptually advances previous approaches by (2) ranking the performance of fit indices in the specific setting of interest. Thus, the unique strength of the simulation-cum-ROC approach is that it provides guidance on which fit index to rely (or at least assign the greatest weight) when evaluating model fit in the specific setting of interest.

To illustrate how our proposed simulation-cum-ROC approach works, we tested models of the Rosenberg Self-Esteem Scale and the Social Desirability-Gamma Short Scale. We wish

to emphasize that we intend these examples as proof of principle. In presenting these examples, we made several choices on the selection of fit indices, the definition of population models, and the relative importance of Type I and Type II error rates in generating tailored cutoffs. Researchers can modify most of these choices when applying the proposed simulation-cum-ROC approach to other empirical problems. We highlight some of these choices in the following to underscore our approach's generality and identify areas in which future research may progress.

To begin with, researchers may consider additional variants of fit indices or different fit indices altogether. In our examples, we focused on the three widely used fit indices, CFI, RMSEA, and SRMR (Jackson et al., 2009), to keep these examples simple. Additionally, as is routine in applied research, we considered χ^2 in much the same way (and not as a strict formal test; see Jöreskog & Sörbom, 1993).[1] We relied on a χ^2 test statistic approximately equivalent to the Yuan-Bentler one (Yuan & Bentler, 2000a; called chisq.scaled in lavaan, see also Savalei & Rosseel, 2022). Following standard practice (e.g., Muthén & Muthén, 1998-2017), we relied on the CFI and RMSEA versions calculated with this χ^2 test statistic (called cfi.scaled and rmsea.scaled in lavaan). The standard formulations of fit indices (and test statistics) are not without critics. Several authors (Brosseau-Liard et al., 2012; Brosseau-Liard & Savalei, 2014; Gomer et al., 2019; Yuan & Marshall, 2004; Yuan, 2005; Zhang, 2008) have pointed out problems and suggested improved formulations. Therefore, researchers may prefer not to go with the fit indices and their conventional formulations we used in the examples. Notably, the simulation-cum-ROC approach can be generalized to include any other fit index (and test statistic), including variants of the canonical fit indices (e.g., Yuan, 2005) but also other, less widely used fit indices (e.g., McDonald's measure of centrality, McDonald, 1989, or the adjusted goodness of fit index, Jöreskog & Sörbom, 1986).

Moreover, in our examples of the simulation-cum-ROC approach, we chose an AUC value of .80 as a threshold. Researchers may choose higher AUC thresholds to obtain lower Type I and Type II error rates. Moreover, we selected that cutoff as the optimal one that had the highest sum of sensitivity + specificity – 1 (i.e., the Youden index balancing Type I and Type II error rates). Alternatively, researchers may maximize sensitivity given a minimal specificity value to obtain optimal cutoffs (or vice versa).

[1] As one reviewer correctly pointed out, RMSEA is just a transformation of χ^2 (e.g., Moshagen & Erdfelder, 2016). RMSEA can therefore be considered redundant because its performance in terms of the AUC will be the same as that of the χ^2. Nonetheless, we decided to generate cutoffs for χ^2 and RMSEA in the examples because both are regularly used for model evaluation (Jackson et al., 2009).

Further, we rejected the empirical model based on tailored cutoffs in both examples. In both examples, we did not go through the steps of modifying the model and testing that modified model again. However, we demonstrate how to employ the simulation-cum-ROC approach to test a modified Social Desirability-Gamma Short Scale model for interested readers in Additional File 2 of the Supplementary Material.

It is essential to realize that tailored cutoffs derived from the simulation-cum-ROC approach are the most accurate decision thresholds for the setting from which they originate. That said, one should not make the same mistake as with traditional cutoffs and generalize tailored cutoffs to any different combination of model, estimation, and data characteristics. Different combinations affect the performance of fit indices and their cutoffs in unexpected and non-traceable ways (for an overview, see Niemand & Mai, 2018; Pornprasertmanit, 2014), and erroneous conclusions may result. We instead underline that no general cutoff or general statement on the performance of those commonly used fit indices exists (see also, e.g., Marsh et al., 2004; McNeish & Wolf, 2021; Nye & Drasgow, 2011).

Advanced Definitions of Population Models

A challenge in applying the simulation-cum-ROC approach—one that it shares with similar simulation-based approaches (e.g., Pornprasertmanit, 2014)—concerns the definition of the H_0 and H_1 population models. More advanced definitions of population models can be easily integrated into the simulation-cum-ROC approach. For example, one could define an H_0 population model relative to which the analysis model is negligibly underspecified to test for approximate fit, as suggested by Millsap (2007, 2013) and Pornprasertmanit (2014). We indeed believe that alternative definitions of the population model can be fruitful, which is why we briefly review possible extensions of our approach (and similar approaches) that have been proposed in prior work. We further identify areas in which future work on generating tailored cutoffs could make further progress.

Approximate Fit

In our examples illustrating the simulation-cum-ROC approach, the to-be-tested analysis models were always identical to the H_0 population models. In other words, we generated cutoffs based on an analysis model that exactly fit the data generated by (i.e., simulated from) an H_0 population model. Only fluctuations through sampling influenced the resulting fit index distributions and, accordingly, the cutoffs (Cudeck & Henly, 1991; MacCallum, 2003;

MacCallum & Tucker, 1991). Testing this assumption of exact fit has guided model evaluation for years; the entire distributional assumptions of the χ^2 test statistic rely on exact fit testing (e.g., Bollen, 1989). Testing exact fit is legitimate if the aim is to find a model that perfectly describes the specific population. This model should perfectly reproduce all major and minor common factors in the specific data.

In empirical applications, researchers commonly want to find models that do not solely reproduce a specific population but are generalizable to different populations (Cudeck & Henly, 1991; Millsap, 2007). In other words, researchers do not want to find an overfitting model. Toward that end, it can be advantageous to consider not only sampling fluctuations but also model error when generating cutoffs (Cudeck & Henly, 1991; MacCallum, 2003; MacCallum & Tucker, 1991). Model error, in this context, means choosing an H_0 population model relative to which the analysis model already contains minor misspecification, such as small unmodeled residual correlations (e.g., Millsap, 2007, 2013). The analysis model is underspecified (i.e., misspecified) to a certain degree relative to the H_0 population model. Researchers still consider the analysis model correctly specified, barring minor misspecification they deem acceptable. It is within their realistic expectations of how well a model can capture the complexities of a real population while still being plausible in other populations (for an overview and in-depth discussion, see MacCallum, 2003). Including model error (in addition to sampling fluctuations) in the derivation of cutoffs is known as testing approximate fit and has already been implemented in several approaches (e.g., Kim & Millsap, 2014; McNeish & Wolf, 2021; Millsap, 2013; Yuan & Hayashi, 2003; Yuan et al., 2004, 2007).

We opted against testing approximate fit in our two examples for didactic reasons (i.e., to keep the exposition simple). However, for interested readers, we included an additional example that illustrates how to select the H_0 population model to test approximate (instead of exact) fit in Additional File 2 of the Supplementary Material. As the example demonstrates, testing approximate fit via the simulation-cum-ROC approach works in much the same way as testing exact fit and poses no additional hurdle.

Multiple Population Models

So far, we have always defined a single H_1 population model to test the fit of an analysis model of interest. In the words of Pornprasertmanit (2014; see also Pornprasertmanit et al., 2013), we followed the *fixed method* (see also Millsap, 2013). By following the fixed method (i.e., defining only a single H_1 population model relative to an analysis model), we take only one

form and one size of misspecification (e.g., omitted residual correlation of $r = .50$) out of all possible ones in the space of conceivable models into account.

Thus, Pornprasertmanit (2014; see also Pornprasertmanit et al., 2013) proposed new methods that take a wider variety of misspecification forms and sizes into account (e.g., omitted residual correlations of $r \geq .50$; omitted cross-loadings $\geq .20$). The methods apply to both H_0 population models (to test approximate fit) and H_1 population models. The only difference in defining the population models is that the H_0 population model implies trivial, acceptable misspecification and the H_1 population model implies severe, unacceptable misspecification of the analysis model.

In the *random method*, one defines several H_0 / H_1 population models relative to an underspecified analysis model. (The analysis model is trivially underspecified relative to the H_0 population models and severely underspecified relative to the H_1 population models.) The algorithm randomly picks a new H_0 / H_1 population model from the initially defined ones each time it starts simulating data. This approach considers multiple H_0 / H_1 population models relative to an underspecified analysis model. Thus, H_0 / H_1 population models are the same for different fit indices but differ across simulation runs.

In the *maximal method* (for defining H_0 population models) or the *minimal method* (for defining H_1 population models), one again defines several H_0 / H_1 population models relative to an underspecified analysis model. Then, one draws data from all those population models and fits the analysis model to the data. When selecting an H_0 population model, one picks the population model that generates data with the largest trivial misfit of the analysis model (quantified through the fit index of interest). When selecting an H_1 population model, one picks the population model that generates data with the smallest severe misfit of the analysis model. Thus, H_0 / H_1 population models can differ for different fit indices but are the same across simulation runs.

Although we only applied the fixed method in our examples (again, to keep the exposition simple and help readers understand the basic mechanics of our simulation-cum-ROC approach), we encourage researchers to consider the random and maximal/minimal methods in future work on the simulation-cum-ROC approach. So far, neither the random nor the maximal/minimal methods are features of the shiny app and the supplementary R code. We plan to implement these features in later versions. Further, a tutorial on the simulation-cum-ROC approach, including exemplary R code containing the random and maximal/minimal methods, will surely aid the application.

Forms and Sizes of Misspecification

So far, we have followed the traditional way of defining population models: The analysis model is (either trivially or severely) misspecified relative to the population model, as it either omits specific parameters, fixes them to a wrong value, or proposes a different model structure altogether (e.g., Curran et al., 1996; Hu & Benter, 1998, 1999; McNeish & Wolf, 2021; Millsap, 2013; Satorra & Saris, 1985; Yuan & Bentler, 1997). This traditional way of defining population models might be considered too static and context-dependent (i.e., dependent on a specific form of misspecification). A further step could be to define population models relative to which analysis models are either trivially (H_0) or severely misspecified (H_1) in an effect size logic that does not require defining a specific form of misspecification (e.g., omitted parameters). For example, some authors (Cudeck & Browne, 1992; Moshagen & Auerswald, 2018; Yuan et al., 2007) proposed defining (or approximating) the population variance-covariance matrix at a given distance from the analysis model structure. This is reminiscent of defining the difference in χ^2 distributions through a predefined difference in the non-centrality parameter (e.g., Moshagen & Erdfelder, 2016; see Jak et al., 2021, for a tool to quantify omitted parameters in terms of the non-centrality parameter). This approach does not require defining misspecification in terms of omitted or wrongly fixed parameters—but rather in an effect size logic.

A more general challenge of defining reasonable population models for generating fit index distributions is that the literature does not provide specific guidelines on the appropriate effect size of trivial or severe misspecification. Although such guidelines would undoubtedly be helpful, providing universally applicable guidelines across settings may not be possible. What constitutes a reasonable population model, and a trivial or severe misspecification of the analysis model relative to that population model, depends on many characteristics of the study, such as the research question, empirical setting, study design, and the data. Researchers need to justify their definition of a population model based on those characteristics. By requiring that population models need to be made explicit, editors, reviewers, and readers of the study can judge the appropriateness of the underlying assumptions.

Conclusion

Tailored cutoffs are ideally suited to the setting of interest because they account for the many model, estimation, and data characteristics that can influence fit indices and render fixed cutoffs questionable. This study reviewed four principal approaches that researchers can employ to generate tailored cutoffs. We then presented a novel approach, the simulation-cum-ROC approach, that extends previous approaches by introducing ROC analysis. Introducing ROC analysis to model fit evaluation is a contribution that uniquely characterizes our approach. It allows evaluating the performance of fit indices in a given setting, enabling researchers to make informed decisions about which fit indices to rely on (or which to assign the greatest weight to). Our approach then derives the most accurate cutoffs for the setting of interest. To the best of our knowledge, the proposed procedure is the only one that allows basing cutoff decisions on balanced Type I and Type II error rates combined with a performance index for fit indices. Our procedure comprises three steps (plus fitting and testing the empirical analysis model). We provide a shiny app and R code to enable researchers to easily generate tailored cutoffs for their own empirical problems. We hope to encourage applied researchers to abandon the traditional fixed cutoffs in favor of tailored ones. This will allow them to make more valid judgments about model fit and ultimately increase the replicability of research findings. By reviewing possible extensions of our approach, we also hope to encourage methodologists to expand further—and help disseminate—the current approaches to generating tailored cutoffs (including our simulation-cum-ROC approach).

References

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M. (1995). *EQS structural equations program manual.* Multivariate Software.

Biau, D. J., Jolles, B. M., & Porcher, R. (2010). P value and the theory of hypothesis testing: an explanation for new researchers. *Clinical Orthopaedics and Related Research, 468*(3), 885–892. https://doi.org/10.1007/s11999-009-1164-4

Bluemke, M., Jong, J., Grevenstein, D., Mikloušić, I., & Halberstadt, J. (2016). Measuring cross-cultural supernatural beliefs with self- and peer-reports. *PLoS ONE, 11*(10), Article e0164291. https://doi.org/10.1371/journal.pone.0164291

Bollen, K. A. (1989). *Structural equations with latent variables.* Wiley.

Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research, 21*(2), 205–229. https://doi.org/10.1177%2F0049124192021002004

Boomsma, A. (2013). Reporting Monte Carlo studies in structural equation modeling. *Structural Equation Modeling, 20*(3), 518–540. https://doi.org/10.1080/10705511.2013.797839

Borsboom, D., van der Maas, H. L., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science, 16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Brosseau-Liard, P. E., & Savalei, V. (2014). Adjusting incremental fit indices for nonnormality. *Multivariate Behavioral Research, 49*(5), 460–470. https://doi.org/10.1080/00273171.2014.933697

Brosseau-Liard, P. E., Savalei, V., & Li, L. (2012). An investigation of the sample performance of two non-normality corrections for RMSEA. *Multivariate Behavioral Research, 47*(6), 904–930. https://doi.org/10.1080/00273171.2012.715252

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

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. https://doi.org/10.1037/1082-989X.7.4.403

Cheng, C., & Wu, H. (2017). Confidence intervals of fit indexes by inverting a bootstrap test. *Structural Equation Modeling, 24*(6), 870–880. https://doi.org/10.1080/10705511.2017.1333432

Chun, S. Y., & Shapiro, A. (2009). Normal versus noncentral chi-square asymptotics of misspecified models. *Multivariate Behavioral Research, 44*(6), 803–827. https://doi.org/10.1080/00273170903352186

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cudeck, R., & Browne, M. W. (1992). Constructing a covariance matrix that yields a specified minimizer and a specified minimum discrepancy function value. *Psychometrika, 57*(3), 357–369. https://doi.org/10.1007/BF02295424

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. https://doi.org/10.1037/0033-2909.109.3.512

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(1), 16–29. https://doi.org/10.1037/1082-989X.1.1.16

D'Agostino Sr, R. B., Pencina, M. J., Massaro, J. M., & Coady, S. (2013). Cardiovascular disease risk assessment: Insights from Framingham. *Global Heart, 8*(1), 11–23. http://doi.org/10.1016/j.gheart.2013.01.001

Flach, P. A. (2016). ROC analysis. In C. Sammut, & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 1–8). Springer. https://doi.org/10.1007/978-1-4899-7502-7_739-1

Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Structural Equation Modeling, 7*(3), 356–410. https://doi.org/10.1207/S15328007SEM0703_2

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry, 31*(4), 271–288. https://doi.org/10.1080/1047840X.2020.1853461

Gomer, B., Jiang, G., & Yuan, K.-H. (2019). New effect size measures for structural equation modeling. *Structural Equation Modeling, 26*(3), 371–389. https://doi.org/10.1080/10705511.2018.1545231

Groeben, N. & Westmeyer, H. (1981). *Kriterien psychologischer Forschung* [Criteria of psychological research]. Juventa.

Groskurth, K., Bluemke, M., & Lechner, C. M. (2022). *Why we need to abandon fixed cutoffs for goodness-of-fit indices: A thorough simulation and possible solutions*. PsyArXiv. https://doi.org/10.31234/osf.io/5qag3

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science.* Advance online publication. https://doi.org/10.1177/1745691620970585

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. https://doi.org/10.1037/a0024917

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

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

Jackson, D. L., Gillaspy Jr, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: an overview and some recommendations. *Psychological Methods, 14*(1), 6–23. https://doi.org/10.1037/a0014694

Jak, S., Jorgensen, T. D., Verdam, M. G., Oort, F. J., & Elffers, L. (2021). Analytical power calculations for structural equation modeling: A tutorial and Shiny app. *Behavior Research Methods, 53*(4), 1385–1406. https://doi.org/10.3758/s13428-020-01479-0

Jobst, L. J., Bader, M., & Moshagen, M. (2021). A tutorial on assessing statistical power and determining sample size for structural equation models. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000423

Jöreskog, K. G., & Sörbom, D. (1986). *LISREL VI: Analysis of linear structural relationships by maximum likelihood and least squares methods.* Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Scientific Software International.

Kemper, C. J., Beierlein, C., Bensch, D., Kovaleva, A., & Rammstedt, B. (2014). Soziale Erwünschtheit-Gamma (KSE-G). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. https://doi.org/10.6102/zis186

Kim, H., & Millsap, R. (2014). Using the Bollen-Stine bootstrapping method for evaluating approximate fit indices. *Multivariate Behavioral Research, 49*(6), 581–596. https://doi.org/10.1080/00273171.2014.947352

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

Lai, K. (2019). A simple analytic confidence interval for CFI given nonnormal data. *Structural Equation Modeling, 26*(5), 757–777. https://doi.org/10.1080/10705511.2018.1562351

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research, 51*(2-3), 220–239. https://doi.org/10.1080/00273171.2015.1134306

MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research, 38*(1), 113–139. https://doi.org/10.1207/S15327906MBR3801_5

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

Mai, R., Niemand, T., & Kraus, S. (2021). A tailored-fit model evaluation strategy for better decisions about structural equation models. *Technological Forecasting and Social Change, 173*, Article 121142. https://doi.org/10.1016/j.techfore.2021.121142

Majnik, M., & Bosnić, Z. (2013). ROC analysis of classifiers in machine learning: A survey. *Intelligent Data Analysis, 17*(3), 531–558. https://doi.org/10.3233/IDA-130592

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. https://doi.org/10.1016/j.paid.2006.09.023

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. https://doi.org/10.1207/s15328007sem1103_2

Maydeu-Olivares, A. (2017). Assessing the size of model misfit in structural equation models. *Psychometrika, 82*(3), 533–558. https://doi.org/10.1007/s11336-016-9552-7

Maydeu-Olivares, A., Shi, D., & Rosseel, Y. (2018). Assessing fit in structural equation models: A Monte-Carlo evaluation of RMSEA versus SRMR confidence intervals and tests of close fit. *Structural Equation Modeling, 25*(3), 389–402. https://doi.org/10.1080/10705511.2017.1389611

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification, 6*(1), 97–103. https://doi.org/10.1007/BF01908590

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000425

McNeish, D., & Wolf, M. G. (2022). Dynamic fit index cutoffs for one-factor models. *Behavior Research Methods.* Advance online publication. https://doi.org/10.3758/s13428-022-01847-y

Millsap, R. E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences, 42*(5), 875–881. https://doi.org/10.1016/j.paid.2006.09.021

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

Moshagen, M. (2012). The model size effect in structural equation modeling: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling, 19*(1), 86–98. https://doi.org/10.1080/10705511.2012.634724

Moshagen, M., & Auerswald, M. (2018). On congruence and incongruence of measures of fit in structural equation modeling. *Psychological Methods, 23*(2), 318–336. https://doi.org/10.1037/met0000122

Moshagen, M., & Erdfelder, E. (2016). A new strategy for testing structural equation models. *Structural Equation Modeling, 2*3(1), 54–60. https://doi.org/10.1080/10705511.2014.950896

Muthén, L.K., & Muthén, B.O. (1998-2017). *Mplus user's guide* (8th ed). Muthén & Muthén.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika, 20A*(1), 175–240. https://doi.org/10.2307/2331945

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231*(694–706), 289–337. https://www.jstor.org/stable/91247

Niemand, T., & Mai, R. (2018). Flexible cutoff values for fit indices in the evaluation of structural equation models. *Journal of the Academy of Marketing Science, 46*, 1148–1172. https://doi.org/10.1007/s11747-018-0602-9

Nießen, D., Partsch, M. V., Kemper, C. J., & Rammstedt, B. (2019). An English-language adaptation of the Social Desirability-Gamma Short Scale (KSE-G). *Measurement Instruments for the Social Sciences, 1*, Article 2. https://doi.org/10.1186/s42409-018-0005-1

Nießen, D., Partsch, M., Groskurth, K. (2020). Data for: An English-language adaptation of the Risk Proneness Short Scale (R-1) (Version: 1.0.0). https://doi.org/10.7802/2080

Nießen, D., Partsch, M., Rammstedt, B. (2018). Data for: An English-language adaptation of the Social Desirability-Gamma Short Scale (KSE-G) (Version: 1.0.0). https://doi.org/10.7802/1752

Nye, C. D., & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods, 14*(3), 548–570. https://doi.org/10.1177/1094428110368562

Olvera Astivia, O. L., & Zumbo, B. D. (2015). A cautionary note on the use of the Vale and Maurelli method to generate multivariate, nonnormal data for simulation purposes. *Educational and Psychological Measurement, 75*(4), 541–567. https://doi.org/10.1177/0013164414548895

Padgett, R. N., & Morgan, G. B. (2021). Multilevel CFA with ordered categorical data: A simulation study comparing fit indices across robust estimation methods. *Structural Equation Modeling, 28*(1), 51–68. https://doi.org/10.1080/10705511.2020.1759426

Paulhus, D. L. (2002). Social desirable responding. The evolution of a construct. In H. I. Braun & D. N. Jackson (Eds.), *The role of constructs in psychological and educational measurement.* Erlbaum.

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology, 6*, Article 223. https://doi.org/10.3389/fpsyg.2015.00223

Pornprasertmanit, S. (2014). *The unified approach for model evaluation in structural equation modeling* [Unpublished doctoral dissertation]. University of Kansas. http://hdl.handle.net/1808/16828

Pornprasertmanit, S., Miller, P., Schoemann, A., & Jorgensen, T. D. (2021). *simsem: SIMulated Structural Equation Modeling*. R package version 0.5.16. https://CRAN.R-project.org/package=simsem

Pornprasertmanit, S., Wu, W., & Little, T. D. (2013). 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). Springer. https://doi.org/10.1007/978-1-4614-9348-8_12

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Reußner, M. (2019). *Die Güte der Gütemaße: Zur Bewertung von Strukturgleichungsmodellen* [The fit of fit indices: The evaluation of model fit for structural equation models]. Walter de Gruyter.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12,* Article 77. https://doi.org/10.1186/1471-2105-12-77

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton University Press.

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

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50*(1), 83–90. https://doi.org/10.1007/BF02294150

Savalei, V., & Rosseel, Y. (2022). Computational options for standard errors and test statistics with incomplete normal and nonnormal data in SEM. *Structural Equation Modeling, 29*(2), 163–181. https://doi.org/10.1080/10705511.2021.1877548

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23–74.

Schmalbach, B., Irmer, J. P., & Schultze, M. (2019). *ezCutoffs: Fit Measure Cutoffs in SEM.* R package version 1.0.1. https://CRAN.R-project.org/package=ezCutoffs

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. https://doi.org/10.1177/0013164418783530

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*(2), 173–180. https://doi.org/10.1207/s15327906mbr2502_4

Supple, A. J., Su, J., Plunkett, S. W., Peterson, G. W., & Bush, K. R. (2013). Factor structure of the Rosenberg Self-Esteem Scale. *Journal of Cross-Cultural Psychology, 44*(5), 748–764. https://doi.org/10.1177/0022022112468942

Thiele, C., & Hirschfeld, G. (2021). cutpointr: Improved estimation and validation of optimal cutpoints in R. *Journal of Statistical Software, 98*(11), 1–27. https://doi.org/10.18637/jss.v098.i11

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika, 48*(3), 465–471. https://doi.org/10.1007/BF02293687

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology, 46*(2), 201–233. https://doi.org/10.1037/xlm0000732

Xia, Y., & Yang, Y. (2018). The influence of number of categories and threshold values on fit indices in structural equation modeling with ordered categorical data. *Multivariate Behavioral Research*, *53*(5), 731–755. https://doi.org/10.1080/00273171.2018.1480346

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods, 51*(1), 409–428. https://doi.org/10.3758/s13428-018-1055-2

Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research, 40*(1), 115–148. https://doi.org/10.1207/s15327906mbr4001_5

Yuan, K.-H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association, 92*(438), 767–774. https://doi.org/10.1080/01621459.1997.10474029

Yuan, K.-H., & Bentler, P. M. (1999). On normal theory and associated test statistics in covariance structure analysis under two classes of nonnormal distributions. *Statistica Sinica, 9*(3), 831–853. https://www.jstor.org/stable/24306618

Yuan, K.-H., & Bentler, P. M. (2000a). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 30*(1), 165–200. https://doi.org/10.1111/0081-1750.00078

Yuan, K.-H., & Bentler, P. M. (2000b). Inferences on correlation coefficients in some classes of nonnormal distributions. *Journal of Multivariate Analysis*, *72*(2), 230–248. https://doi.org/10.1006/jmva.1999.1858

Yuan, K.-H., & Bentler, P. M. (2007). Robust procedures in structural equation modeling. In S.-Y. Li (Ed.), *Handbook of latent variable and related models* (pp. 367–397). North-Holland.

Yuan, K.-H., & Hayashi, K. (2003). Bootstrap approach to inference and power analysis based on three test statistics for covariance structure models. *British Journal of Mathematical and Statistical Psychology*, *56*(1), 93–110. https://doi.org/10.1348/000711003321645368

Yuan, K.-H., & Marshall, L. L. (2004). A new measure of misfit for covariance structure models. *Behaviormetrika, 31*(1), 67–90. https://doi.org/10.2333/bhmk.31.67

Yuan, K.-H., Bentler, P. M., & Chan, W. (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika, 69*(3), 421–436. https://doi.org/10.1007/BF02295644

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. https://doi.org/10.1080/10705511.2015.1065414

Yuan, K.-H., Hayashi, K., & Yanagihara, H. (2007). A class of population covariance matrices in the bootstrap approach to covariance structure analysis. *Multivariate Behavioral Research, 42*(2), 261–281. https://doi.org/10.1080/00273170701360662

Yung, Y. F., & Bentler, P. M. (1996). Bootstrap techniques in analysis of mean and covariance structures. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 195–226). Erlbaum.

Zhang, W. (2008). A comparison of four estimators of a population measure of model fit in covariance structure analysis. *Structural Equation Modeling*, *15*(2), 301–326. https://doi.org/10.1080/10705510801922555

Zhang, X., & Savalei, V. (2016). Bootstrapping confidence intervals for fit indices in structural equation modeling. *Structural Equation Modeling, 23*(3), 392–408. https://doi.org/10.1080/10705511.2015.1118692

Additional File 1: R Code for the Examples

R Code for the Self-Written Functions Needed for the Examples

```
################################
#Functions for: TAILORED CUTOFFS FOR FIT INDICES VIA THE SIMULATION-CUM-ROC APPROACH
################################

#NOTE:  If the goal is NOT to put equal weight on sensitivity and specificity when deriving
cutoffs (metric = youden)
#       but to put either more weight on sensitivity (metric = sens_constrain) or specificity
(metric = spec_constrain),
#       the "metric"-argument of cutpointr needs to be changed accordingly.


####
#Function to simulate data and fit analysis model to it
####
#Default: CFA model identified via latent variance
sim.Dat <- function(runs=runs, analysisModel=analysisModel, sample=sample,
populationModel.cor=populationModel.cor,
                    populationModel.mis=populationModel.mis, std.lv=TRUE, modelKind = "cfa",
estimator = estimator,
                    dist = dist, seed=seed){

          #Load required package
          if (!require(simsem)) { install.packages("simsem"); require(simsem) }

          #Simulate data from population model relative to which the analysis model is
correctly specified, fit analysis model to simulated data
          Output.cor <- simsem::sim(runs, model=analysisModel, n=sample,
generate=populationModel.cor, std.lv=std.lv, lavaanfun = modelKind, estimator = estimator,
indDist = dist, seed=seed, silent=TRUE)

          #Simulate data from population model relative to which the analysis model is
misspecified, fit analysis model to simulated data
          Output.mis <- simsem::sim(runs, model=analysisModel, n=sample,
generate=populationModel.mis, std.lv=std.lv, lavaanfun = modelKind, estimator = estimator,
indDist = dist, seed=seed,  silent=TRUE)

          #Extract fit indices
          fit.cor <- inspect(Output.cor, "fit")
          fit.mis <- inspect(Output.mis, "fit")

          #Add variable indicating whether the analysis model is correctly specified or
misspecified
          fit.cor$specification <- factor(0, 0:1, c("correctly specified", "misspecified"))
          fit.mis$specification <- factor(1, 0:1, c("correctly specified", "misspecified"))

          #Combine data relative to which the analysis model is either correctly specified and
misspecified
          fit <- rbind(fit.cor, fit.mis)
          return(fit)

}


####
#Function to convert strings to quosures
####
#Credits: https://github.com/r-lib/rlang/issues/116
stringToQuoser <- function(varName) {
  if (!require(wrapr)) { install.packages("wrapr"); require(wrapr) }
  if (!require(rlang)) { install.packages("rlang"); require(rlang) }
  wrapr::let(c(VARNAME = varName), rlang::quo(VARNAME))

}


####
#Function to display ROC curve
####
```

```
plotROC <- function(GOFs = GOFs, GOFs.reverse = GOFs.reverse, AUC.cut = AUC.cut, data = fit) {

  #Load required packages
  if (!require(cutpointr)) { install.packages("cutpointr"); require(cutpointr) }
  if (!require(pROC)) { install.packages("pROC"); require(pROC) }
  if (!require(ggrepel)) { install.packages("ggrepel"); require(simsemggrepel) }
  if (!require(dplyr)) { install.packages("dplyr"); require(dplyr) }

  #Select only the best demarcating fit indices (i.e., those with AUC >= than AUC cutpoint)
and save them in one object
  GOFs.updated <- c()
  for (g in 1:length(GOFs)){
    GOF <- stringToQuoser(GOFs[[g]])

    if (is.na(match(GOFs[[g]], GOFs.reverse)) == TRUE){
      test.GOF <- cutpointr::cutpointr(fit, x = !!GOF, class = specification, pos_class =
"correctly specified",
                                       neg_class = "misspecified",
                                       direction = "<=",
                                       method = maximize_metric, metric = youden, na.rm=TRUE,
use_midpoints=TRUE)

    }else{
      test.GOF <- cutpointr::cutpointr(fit, x = !!GOF, class = specification, pos_class =
"correctly specified",
                                       neg_class = "misspecified",
                                       direction = ">=",
                                       method = maximize_metric, metric = youden, na.rm=TRUE,
use_midpoints=TRUE)
    }

    if (test.GOF$AUC>=AUC.cut){
      GOFs.updated <- c(GOFs.updated, GOFs[[g]])
    }
  }

  #Here: only for those fit indices with AUC >= than AUC cutpoint.
  if (length(GOFs.updated)>0 & length(GOFs.updated)<7){

    test.GOF <- list()
    models <- list()
    preds <- list()
    rocs <- list()

    #Save fit indices with AUC >= than AUC cutpoint.
    for (gu in 1:length(GOFs.updated)){
      GOF <- stringToQuoser(GOFs.updated[[gu]])

      if (is.na(match(GOFs.updated[[gu]], GOFs.reverse)) == TRUE){
        test.GOF[[gu]] <- cutpointr::cutpointr(fit, x = !!GOF, class = specification,
pos_class = "correctly specified",
                                               neg_class = "misspecified",
                                               direction = "<=",
                                               method = maximize_metric, metric = youden,
na.rm=TRUE, use_midpoints=TRUE)

      }else{
        test.GOF[[gu]] <- cutpointr::cutpointr(fit, x = !!GOF, class = specification,
pos_class = "correctly specified",
                                               neg_class = "misspecified",
                                               direction = ">=",
                                               method = maximize_metric, metric = youden,
na.rm=TRUE, use_midpoints=TRUE)
      }



      #Predict fitted values for those with AUC >= than AUC cutpoint.
      models[[gu]] <- glm(as.formula(paste0("specification ~ ", GOFs.updated[[gu]])), data =
fit, family = binomial(link = "logit"))
      preds[[gu]] <- predict(models[[gu]], type = "response")

      if (is.na(match(GOFs.updated[[gu]], GOFs.reverse)) == TRUE){
      rocs[[gu]] <- pROC::roc(as.formula(paste0("specification ~ ", GOFs.updated[[gu]])), fit,
direction = "<")
      }else{
```

```
      rocs[[gu]] <- pROC::roc(as.formula(paste0("specification ~ ", GOFs.updated[[gu]])), fit,
direction = ">")
      }
        }


    #Plot ROC curves
    windowsFonts(Times = windowsFont("Times New Roman"))
    for (gu in 1:length(GOFs.updated)){
      if(gu == 1){
        plot(rocs[[1]], col = "#555555", lty = 1, lwd=1, xlab="false positive rate (1 -
specificity)", ylab="true positive rate (sensitivity)", legacy.axes = TRUE, family = "Times")
        legend.full <- c(paste0(GOFs.updated[[1]], " (AUC = ", round(test.GOF[[1]]$AUC,3),
")"))
      } else {
        plot(rocs[[gu]], col = "#555555", lty = gu, lwd=1+gu/10*5, add = TRUE, legacy.axes =
TRUE, family = "Times")
        legend.full <- c(legend.full,paste0(GOFs.updated[[gu]], " (AUC = ",
round(test.GOF[[gu]]$AUC,3), ")"))
      }
    }

    op <- par(family = "serif")

    legend("bottomright",
           title= paste0("fit indices with AUC >= ", AUC.cut),
           legend = legend.full,
           col = "#555555",
           bty="n",
           lty = 1:gu,
           lwd=1:1+gu/10*5,
           text.col = "black",
           horiz = F)

    par(op)

    #Save plot externally
    tiff("ROCcurve.tiff", units="cm", width=14, height=14, res=600)

    windowsFonts(Times = windowsFont("Times New Roman"))
    for (gu in 1:length(GOFs.updated)){
      if(gu == 1){
        plot(rocs[[1]], col = "#555555", lty = 1, lwd=1, xlab="false positive rate (1 -
specificity)", ylab="true positive rate (sensitivity)", legacy.axes = TRUE, family = "Times")
        legend.full <- c(paste0(GOFs.updated[[1]], " (AUC = ", round(test.GOF[[1]]$AUC,3),
")"))
      } else {
        plot(rocs[[gu]], col = "#555555", lty = gu, lwd=1+gu/10*5, add = TRUE, legacy.axes =
TRUE, family = "Times")
        legend.full <- c(legend.full,paste0(GOFs.updated[[gu]], " (AUC = ",
round(test.GOF[[gu]]$AUC,3), ")"))
      }
    }

    op <- par(family = "serif")

    legend("bottomright",
           title= paste0("fit indices with AUC >= ", AUC.cut),
           legend = legend.full,
           col = "#555555",
           bty="n",
           lty = 1:gu,
           lwd=1:1+gu/10*5,
           text.col = "black",
           horiz = F)
    par(op)

    dev.off()

  } else if (length(GOFs.updated)>0 & length(GOFs.updated)<7){

    print("ROC curves are only displayed when less than seven fit indices perform well.")

  } else {print("There is no fit index meeting the AUC criteria.")}

}
```

```
####
#Function to display fit index distribution and cutoff (including its accuracy)
####

cutoffPerform <- function(GOFs = GOFs, GOFs.reverse = GOFs.reverse, AUC.cut = AUC.cut, data =
fit) {


  #Load required packages
  if (!require(cutpointr)) { install.packages("cutpointr"); require(cutpointr) }
  if (!require(ggpubr)) { install.packages("ggpubr"); require(ggpubr) }

  #Select only the best demarcating fit indices (i.e., those with AUC >= than AUC cutpoint)
and save them in one object
  return.list <- list()
  GOFs.updated <- c()
  for (g in 1:length(GOFs)){
    GOF <- stringToQuoser(GOFs[[g]])

    if (is.na(match(GOFs[[g]], GOFs.reverse)) == TRUE){
      test.GOF <- cutpointr::cutpointr(fit, x = !!GOF, class = specification, pos_class =
"correctly specified",
                                       neg_class = "misspecified",
                                       direction = "<=",
                                       method = maximize_metric, metric = youden, na.rm=TRUE,
use_midpoints=TRUE)

    }else{
      test.GOF <- cutpointr::cutpointr(fit, x = !!GOF, class = specification, pos_class =
"correctly specified",
                                       neg_class = "misspecified",
                                       direction = ">=",
                                       method = maximize_metric, metric = youden, na.rm=TRUE,
use_midpoints=TRUE)
    }

    if (test.GOF$AUC>=AUC.cut){
      GOFs.updated <- c(GOFs.updated, GOFs[[g]])
    }
  }

  #Here: only for those fit indices with AUC >= than AUC cutpoint.
  if (length(GOFs.updated)>0){

    test.GOF <- list()

    #Save those GOFs with AUC >= than AUC cutpoint.
    for (gu in 1:length(GOFs.updated)){
      GOF <- stringToQuoser(GOFs.updated[[gu]])

      if (is.na(match(GOFs.updated[[gu]], GOFs.reverse)) == TRUE){
        test.GOF[[gu]] <- cutpointr::cutpointr(fit, x = !!GOF, class = specification,
pos_class = "correctly specified",
                                               neg_class = "misspecified",
                                               direction = "<=",
                                               method = maximize_metric, metric = youden,
na.rm=TRUE, use_midpoints=TRUE)

      }else{
        test.GOF[[gu]] <- cutpointr::cutpointr(fit, x = !!GOF, class = specification,
pos_class = "correctly specified",
                                               neg_class = "misspecified",
                                               direction = ">=",
                                               method = maximize_metric, metric = youden,
na.rm=TRUE, use_midpoints=TRUE)
      }

    }


    #Plot fit index distribution
    distr <- list()
    for (gu in 1:length(GOFs.updated)){
      names(test.GOF[[gu]][["data"]][[1]])[1] <- "value"
```

```
      windowsFonts(Times = windowsFont("Times New Roman"))
      distr[[gu]] <- test.GOF[[gu]][["data"]][[1]] %>% ggplot( aes(x=value,
fill=specification)) +
                      geom_histogram(alpha=0.5, position = 'identity', bins=30) +
                      xlab(paste0(GOFs.updated[[gu]], "\ncutoff =
",round(test.GOF[[gu]]$optimal_cutpoint,3),": accuracy = ", round(test.GOF[[gu]]$acc,3),
                                                          ", \ntype I error rate = ", round((1-
test.GOF[[gu]]$specificity)*100,0),"%",
                                                          ", type II error rate = ", round((1-
test.GOF[[gu]]$sensitivity)*100,0),"%")) +
                      ylab("")+ aes(fill=specification) + theme_minimal() +
scale_fill_manual(values=c("gray70", "gray40")) +
                      geom_vline(xintercept = test.GOF[[gu]]$optimal_cutpoint,
linetype="solid", color = "black", size=1)+
                      annotate(x=test.GOF[[gu]]$optimal_cutpoint,y=+Inf, label= paste0("cutoff
\n= \n",round(test.GOF[[gu]]$optimal_cutpoint,3)),vjust=2, geom="label", size=3, family =
"Times")+
                      guides(fill=guide_legend(title="analysis model"))+
                      theme(text=element_text(family="Times"))
      names(test.GOF[[gu]][["data"]][[1]])[1] <- paste0(GOFs.updated[[gu]])
       }

    full.distr <- ggpubr::ggarrange(plotlist=distr,  common.legend=TRUE)
    full.distr <- ggpubr::annotate_figure(full.distr,
                                          left = text_grob(paste0("count", collapse=""), color
= "black", size = 12, rot = 90, family = "Times"))


    return.list[[1]] <- full.distr

    #Save plots externally
    ggplot2::ggsave("Cutoff.tiff", full.distr, height = 17, width = 20, dpi = 600, units =
"cm")

    #Display (accuracy of) cutoffs for fit indices
    for (gu in 1:length(GOFs.updated)){
      return.list[[gu+1]] <- summary(test.GOF[[gu]])
    }

    return(return.list)

  } else {print("There is no fit index meeting the AUC criteria.")}

}


####
#Function to tabulate tailored cutoffs against empirical values of fit indices
####

modelfit <- function(GOFs = GOFs, GOFs.reverse = GOFs.reverse, AUC.cut = AUC.cut,
empirical.model = empirical.model, simulated.data = fit) {

  #Load required packages
  if (!require(cutpointr)) { install.packages("cutpointr"); require(cutpointr) }
  if (!require(ggpubr)) { install.packages("ggpubr"); require(ggpubr) }
  if (!require(lavaan)) { install.packages("lavaan"); require(lavaan) }
  if (!require(dplyr)) { install.packages("dplyr"); require(dplyr) }
  if (!require(tibble)) { install.packages("tibble"); require(tibble) }
  if (!require(tableHTML)) { install.packages("tableHTML"); require(tableHTML) }

  #Select only the best demarcating fit indices (i.e., those with AUC >= than AUC cutpoint)
and save them in one object
  GOFs.updated <- c()
  for (g in 1:length(GOFs)){
    GOF <- stringToQuoser(GOFs[[g]])

    if (is.na(match(GOFs[[g]], GOFs.reverse)) == TRUE){
      test.GOF <- cutpointr::cutpointr(fit, x = !!GOF, class = specification, pos_class =
"correctly specified",
                                       neg_class = "misspecified",
                                       direction = "<=",
                                       method = maximize_metric, metric = youden, na.rm=TRUE,
use_midpoints=TRUE)

    }else{
```

```
      test.GOF <- cutpointr::cutpointr(fit, x = !!GOF, class = specification, pos_class =
"correctly specified",
                                       neg_class = "misspecified",
                                       direction = ">=",
                                       method = maximize_metric, metric = youden, na.rm=TRUE,
use_midpoints=TRUE)
    }

    if (test.GOF$AUC>=AUC.cut){
      GOFs.updated <- c(GOFs.updated, GOFs[[g]])
    }
  }

  #Here: only for those fit indices with AUC >= than AUC cutpoint.
  if (length(GOFs.updated)>0){

    test.GOF <- list()

    #Save those fit indices with AUC >= than AUC cutpoint.
    for (gu in 1:length(GOFs.updated)){
      GOF <- stringToQuoser(GOFs.updated[[gu]])

      if (is.na(match(GOFs.updated[[gu]], GOFs.reverse)) == TRUE){
        test.GOF[[gu]] <- cutpointr::cutpointr(fit, x = !!GOF, class = specification,
pos_class = "correctly specified",
                                               neg_class = "misspecified",
                                               direction = "<=",
                                               method = maximize_metric, metric = youden,
na.rm=TRUE, use_midpoints=TRUE)

      }else{
        test.GOF[[gu]] <- cutpointr::cutpointr(fit, x = !!GOF, class = specification,
pos_class = "correctly specified",
                                               neg_class = "misspecified",
                                               direction = ">=",
                                               method = maximize_metric, metric = youden,
na.rm=TRUE, use_midpoints=TRUE)
      }

    }


    #Save tailored cutoffs and compare them in a table to the empirical values of fit indices
    ROC.fit <- c()
    for (gu in 1:length(GOFs.updated)){
      ROC.fit <- c(ROC.fit, round(test.GOF[[gu]]$optimal_cutpoint,3))
    }
    names(ROC.fit) <- GOFs.updated

    fit.tbl <-
cbind(as.data.frame(ROC.fit),as.data.frame(round(c(lavaan::fitMeasures(empirical.model,
fit.measures=GOFs.updated)), digits = 3)))
    colnames(fit.tbl) <- c("tailored.cutoffs", "empirical.values")
    rownames(fit.tbl) <- GOFs.updated

    for (gu in 1:length(GOFs.updated)){
      if (is.na(match(GOFs.updated[[gu]], GOFs.reverse)) == FALSE){
        fit.tbl[GOFs.updated[[gu]],] <- fit.tbl[GOFs.updated[[gu]],]*(-1)
      }}

    fit.tbl <- data.frame(fit.tbl) %>%
      dplyr::mutate(empirical.values = ifelse(empirical.values > tailored.cutoffs,
paste0('<font color="red">', empirical.values, '</font>'), paste0('<font color="green">',
empirical.values, '</font>')))
    rownames(fit.tbl) <- GOFs.updated

    for (gu in 1:length(GOFs.updated)){
      if (is.na(match(GOFs.updated[[gu]], GOFs.reverse)) == FALSE){
        fit.tbl[GOFs.updated[[gu]], "tailored.cutoffs"] <- fit.tbl[GOFs.updated[[gu]],
"tailored.cutoffs"]*(-1)
        fit.tbl[GOFs.updated[[gu]], "empirical.values"] <- gsub(">-
","<",fit.tbl[GOFs.updated[[gu]], "empirical.values"])
      }}

    overview <- data.frame(fit.tbl) %>%
      tableHTML::tableHTML(escape = FALSE, rownames=TRUE, widths = rep(100,3)) %>%
```

```
      add_theme('scientific')

    return(overview)

  } else {print("There is no fit index meeting the AUC criteria.")}

}
```

R Code for the Examples

R Code for the Rosenberg Self-Esteem Scale Example

```
############################
#Generate Tailored Cutoffs via the Simulation-cum-ROC Approach
#for the Two-Factor Model of the Rosenberg Self-Esteem Scale
############################

#Background:  - The Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965) measures global self-
esteem
#             - 10 items (5 referring to positive feelings, 5 to negative ones)
#             - Answers must be given on a four-point Likert scale

#Goal:  Although initially constructed as a one-factor model,
#       empirical studies instead found evidence for a two-factor model
#       (for an overview see Supple et al., 2013).
#       The goal is to test the two-factor RSES analysis models.


#Data:  Publicly available data set including the RSES in Germany and the UK
#       (Niessen, et al., 2020).
#       --> here we only look at the UK data


###################
# Preparation
###################

#Clear working space
rm(list = ls())

#Set working directory
#setwd()

#Load data
load("R-1.Rda")

#Select RSES data for UK respondents
UK <- R1[ which(R1$COUN==2), ]
RSES.data <- UK[c("RSES1", "RSES2R", "RSES3", "RSES4", "RSES5R", "RSES6R", "RSES7", "RSES8R",
"RSES9R", "RSES10")]
rm(R1, UK)

#RSES in UK must to be recoded
#so that higher values imply more self-esteem
RSES.data$RSES1 <- 5-RSES.data$RSES1
RSES.data$RSES2R <- 5-RSES.data$RSES2R
RSES.data$RSES3 <- 5-RSES.data$RSES3
RSES.data$RSES4 <- 5-RSES.data$RSES4
RSES.data$RSES5R <- 5-RSES.data$RSES5R
RSES.data$RSES6R <- 5-RSES.data$RSES6R
RSES.data$RSES7 <- 5-RSES.data$RSES7
RSES.data$RSES8R <- 5-RSES.data$RSES8R
RSES.data$RSES9R <- 5-RSES.data$RSES9R
RSES.data$RSES10 <- 5-RSES.data$RSES10

#Load required package
if (!require(lavaan)) { install.packages("lavaan"); require(lavaan) }
```

```
if (!require(psych)) { install.packages("psych"); require(psych) }
if (!require(simsem)) { install.packages("simsem"); require(simsem) }

#Read in functions for generating cutoffs via the simulation-cum-ROC approach
source("FUNCTIONS_cutoffsROC.R")

#####################
# Input: Fit analysis model to empirical data
#####################

#################
# Two-factor
#################
cfa1 <- "
LV_RSES =~ NA*RSES1 + RSES3 + RSES4 + RSES7 + RSES10
LV_RSESR =~ NA*RSES2R + RSES5R + RSES6R  + RSES8R + RSES9R

LV_RSES + LV_RSESR ~ 0*1
"

cfa1.fit <- lavaan::cfa(cfa1, data = RSES.data, estimator = "mlr", missing = "fiml",
std.lv=TRUE)
summary(cfa1.fit, standardized = T, fit.measures = T)

#Extract model parameters for the simulation-cum-ROC approach
model1 <- as.data.frame(lavaan::parameterEstimates(cfa1.fit))[c("lhs", "op", "est", "rhs")]
model1$rhs[model1$op == "~1"] <- "1"
model1$op[model1$op == "~1"] <- "~"
model1 <- paste(paste0(model1$lhs, model1$op, round(model1$est, digits=3), "*", model1$rhs),
collapse = "\n ")
model1


#########################################################################################
##############

#########
#Step 1: Simulate data and fit analysis model to simulated data
#########


#Population model relative to which the analysis model is correctly specified
populationModel.cor <- model1

#Population model (bi-factor with positive and negative method factor) relative to which the
analysis model is misspecified
cfa2 <- "
LV_RSES =~ NA*RSES1 + RSES3 + RSES4 + RSES7 + RSES10 + RSES2R + RSES5R + RSES6R  + RSES8R +
RSES9R

LV_RSESPO =~ NA*RSES1 + RSES3 + RSES4 + RSES7 + RSES10
LV_RSESNE =~ NA*RSES2R + RSES5R + RSES6R  + RSES8R + RSES9R

LV_RSES + LV_RSESPO + LV_RSESNE ~ 0*1
LV_RSES ~~ 0*LV_RSESPO
LV_RSES ~~ 0*LV_RSESNE
LV_RSESPO ~~ 0*LV_RSESNE
"

cfa2.fit <- lavaan::cfa(cfa2, data = RSES.data, estimator = "mlr", missing = "fiml",
std.lv=TRUE)
summary(cfa2.fit, standardized = T, fit.measures = T)

model2 <- as.data.frame(lavaan::parameterEstimates(cfa2.fit))[c("lhs", "op", "est", "rhs")]
model2$rhs[model2$op == "~1"] <- "1"
model2$op[model2$op == "~1"] <- "~"
model2 <- paste(paste0(model2$lhs, model2$op, round(model2$est, digits=3), "*", model2$rhs),
collapse = "\n ")
model2

populationModel.mis <- model2

#Analysis model
analysisModel <- cfa1

######
```

```
#Response distribution per item (skewness and excessive kurtosis)
dist <- simsem::bindDist(skewness = psych::describe(RSES.data)$skew,
                         kurtosis = psych::describe(RSES.data)$kurtosis)

#Sample size
sample <- c(468)

#Simulation runs
runs <- 500

#Estimator
estimator <- c("MLR")

#Select seed to make the simulation reproducible
seed <- c(12345)

#Track starting time
start_time <- Sys.time()

#Execute simulation and save fit indices per run
fit <- sim.Dat(runs=runs, analysisModel=analysisModel, sample=sample,
populationModel.cor=populationModel.cor,
               populationModel.mis=populationModel.mis, std.lv=TRUE, modelKind = "cfa",
estimator = estimator,
               dist = dist, seed=seed)

#Track end time and estimate needed time
end_time <- Sys.time()
end_time - start_time

#Clean cache to free up memory space
invisible(gc())

#########
#Step 2: Evaluate the performance of fit indices
#########

#Choose fit indices of interest
GOFs <- c("chisq.scaled","cfi.scaled", "rmsea.scaled", "srmr")

#Which of the fit indices of interest are reverse coded, that is higher values imply better
fit (e.g., CFI)?
#if no reversed fit indices are included please write GOFs.reverse <- c()
GOFs.reverse <- c("cfi.scaled")

#Set AUC cutpoint --> you may play around with it
AUC.cut <- 0.800

#Plot ROC curve(s)
plotROC(GOFs = GOFs, GOFs.reverse = GOFs.reverse, AUC.cut = AUC.cut, data = fit)

#########
#Step 3: Generate tailored cutoffs
#########
cutoffPerform(GOFs = GOFs, GOFs.reverse = GOFs.reverse, AUC.cut = AUC.cut, data = fit)

##############################################################################################
#########
#Output: Evaluate the fit of the  analysis model to empirical data with tailored cutoffs
########

#Tabulate tailored cutoffs via the simulation-cum-ROC approach against empirical fit indices
modelfit(GOFs = GOFs, GOFs.reverse = GOFs.reverse, AUC.cut = AUC.cut, empirical.model =
cfa1.fit, simulated.data = fit)


############################
#REFERENCES
#Niessen, D., Partsch, M., Groskurth, K. (2020). Data for: An English-Language Adaptation of
the Risk Proneness Short Scale (R-1) (Version: 1.0.0). https://doi.org/10.7802/2080
#Rosenberg, M. (1965). Society and the adolescent self-image. Princeton University Press.
#Supple, A. J., Su, J., Plunkett, S. W., Peterson, G. W., & Bush, K. R. (2013). Factor
structure of the Rosenberg self-esteem scale. Journal of Cross-Cultural Psychology, 44(5),
748-764. https://doi.org/10.1177/0022022112468942
```

R Code for the Social Desirability-Gamma Short Scale Example

```
############################
#Generate Tailored Cutoffs via the Simulation-cum-ROC Approach
#for the Model of the Social Desirability-Gamma Short Scale
############################

#Background:  - The Social Desirability-Gamma Short Scale (KSE-G; Kemper et al., 2014; Niessen
et al., 2019) measures two aspects of the Gamma factor
#              of socially desirable responding (SDR) with three items each:
#               exaggerating positive qualities (PQ+) and minimizing negative qualities (NQ-).
#             - The scale is constructed as a two-factor scale. Both factors are allowed to
correlate.

#Data:  Freely available data set including the KSE-G in Germany and the UK
#       (Niessen, et al., 2018).
#       For the sample description, see Niessen, et al. (2019) --> here we only look at the
German data


###################
# Preparation
###################

#Clear working space
rm(list = ls())

#Set working directory
#setwd()

#Load data
load("KSE-G.Rda")

#Select KSE-G data for German respondents
GER <- KSEG[ which(KSEG$COUN==1), ]
KSEG.data <- GER[c("SDPQ1", "SDPQ2", "SDPQ3", "SDNQ1", "SDNQ2", "SDNQ3")]
rm(KSEG, GER)

#Recode NQ- so that higher values imply more socially desirable responding
KSEG.data$SDNQ1 <- 6-KSEG.data$SDNQ1
KSEG.data$SDNQ2 <- 6-KSEG.data$SDNQ2
KSEG.data$SDNQ3 <- 6-KSEG.data$SDNQ3

#Load required package
if (!require(lavaan)) { install.packages("lavaan"); require(lavaan) }
if (!require(psych)) { install.packages("psych"); require(psych) }
if (!require(simsem)) { install.packages("simsem"); require(simsem) }

#Read in functions for generating cutoffs via the simulation-cum-ROC approach
source("FUNCTIONS_cutoffsROC.R")

#####################
# Input: Fit analysis model to empirical data
#####################

#################
# Two-factor (exact/approximate fit)
#################
cfa1 <- "
LV_SDPQ =~ NA*SDPQ1 + SDPQ2 + SDPQ3
LV_SDNQ =~ NA*SDNQ1 + SDNQ2 + SDNQ3

LV_SDPQ+LV_SDNQ ~ 0*1
"

cfa1.fit <- lavaan::cfa(cfa1, data = KSEG.data, estimator = "mlr", missing = "fiml",
std.lv=TRUE)
summary(cfa1.fit, standardized = T, fit.measures = T)

#Extract model parameters for the simulation-cum-ROC approach
model1 <- as.data.frame(lavaan::parameterEstimates(cfa1.fit))[c("lhs", "op", "est", "rhs")]
model1$rhs[model1$op == "~1"] <- "1"
model1$op[model1$op == "~1"] <- "~"
model1 <- paste(paste0(model1$lhs, model1$op, round(model1$est, digits=3), "*", model1$rhs),
collapse = " \n")
```

```
model1

#################
# Select for: Modification of the two-factor analysis model (exact fit)
#################
#modindices(cfa1.fit) #Highest modification index: SDPQ1 ~~ SDPQ2: MI = 25.437

#Define two-factor analysis model including additional residual covariance
#cfa2 <- "
#LV_SDPQ =~ NA*SDPQ1 + SDPQ2 + SDPQ3
#LV_SDNQ =~ NA*SDNQ1 + SDNQ2 + SDNQ3

#LV_SDPQ+LV_SDNQ ~ 0*1
#SDPQ1 ~~ SDPQ2
#"

#cfa2.fit <- lavaan::cfa(cfa2, data = KSEG.data, estimator = "mlr", missing = "fiml",
std.lv=TRUE)
#summary(cfa2.fit, standardized = T, fit.measures = T)

#Extract model parameters for the simulation-cum-ROC approach
#model2 <- as.data.frame(lavaan::parameterEstimates(cfa2.fit))[c("lhs", "op", "est", "rhs")]
#model2$rhs[model2$op == "~1"] <- "1"
#model2$op[model2$op == "~1"] <- "~"
#model2 <- paste(paste0(model2$lhs, model2$op, round(model2$est, digits=3), "*", model2$rhs),
collapse = " \n")
#model2


##############################################################################################
###############

#########
#Step 1: Simulate data and fit analysis model to simulated data
#########

#Population model relative to which the analysis model is correctly specified
populationModel.cor <- model1

#Population model relative to which the analysis model is misspecified
print(paste0("The residual correlation of .500 between SDPQ1 and SDPQ2 corresponds to a
residual covariance of ", round(0.50*sqrt(0.505*0.321),3))) # 0.201
print(paste0("The residual correlation of .500 between SDNQ1 and SDNQ3 corresponds to a
residual covariance of ", round(0.50*sqrt(0.673*0.556),3))) # 0.306
populationModel.mis <- paste(model1, " \nSDPQ1~~0.201*SDPQ2 \nSDNQ1~~0.306*SDNQ3")

#Analysis model
analysisModel <- cfa1

#################
# Select for: Modification of the two-factor analysis model (exact fit)
#################
#Population model relative to which the analysis model is correctly specified
#populationModel.cor <- model2

#Population model relative to which the analysis model is misspecified
#print(paste0("The residual correlation of .500 between SDPQ1 and SDPQ3 corresponds to a
residual covariance of ", round(0.50*sqrt(0.653*0.224),3))) # 0.191
#print(paste0("The residual correlation of .500 between SDNQ1 and SDNQ3 corresponds to a
residual covariance of ", round(0.50*sqrt(0.684*0.554),3))) # 0.308
#populationModel.mis <- paste(model2, "\nSDPQ1~~0.191*SDPQ3 \nSDNQ1~~0.308*SDNQ3")

#Analysis model
#analysisModel <- cfa2

#################
# Select for: Two-factor analysis model (approximate fit)
#################
#Population model relative to which the analysis model is correctly specified
#print(paste0("The residual correlation of .200 between SDPQ1 and SDPQ2 corresponds to a
residual covariance of ", round(0.20*sqrt(0.505*0.321),3))) # 0.081
#print(paste0("The residual correlation of .200 between SDNQ1 and SDNQ3 corresponds to a
residual covariance of ", round(0.20*sqrt(0.673*0.556),3))) # 0.122
#populationModel.cor <- paste(model1, "\nSDPQ1~~0.081*SDPQ2 \nSDNQ1~~0.122*SDNQ3")

#Population model relative to which the analysis model is misspecified
```

```
#print(paste0("The residual correlation of .500 between SDPQ1 and SDPQ2 corresponds to a
residual covariance of ", round(0.50*sqrt(0.505*0.321),3))) # 0.201
#print(paste0("The residual correlation of .500 between SDNQ1 and SDNQ3 corresponds to a
residual covariance of ", round(0.50*sqrt(0.673*0.556),3))) # 0.306
#populationModel.mis <- paste(model1, " \nSDPQ1~~0.201*SDPQ2 \nSDNQ1~~0.306*SDNQ3")

#Analysis model
#analysisModel <- cfa1

######

#Response distribution per item (skewness and excessive kurtosis)
dist <- simsem::bindDist(skewness = psych::describe(KSEG.data)$skew,
                         kurtosis = psych::describe(KSEG.data)$kurtosis)

#Sample size
sample <- c(474)

#Simulation runs
runs <- 500

#Estimator
estimator <- c("MLR")

#Select seed to make the simulation reproducible
seed <- c(12345)

#Track starting time
start_time <- Sys.time()

#Execute simulation and save fit indices per run
fit <- sim.Dat(runs=runs, analysisModel=analysisModel, sample=sample,
populationModel.cor=populationModel.cor,
               populationModel.mis=populationModel.mis, std.lv=TRUE, modelKind = "cfa",
estimator = estimator,
               dist = dist, seed=seed)

#Track end time and estimate needed time
end_time <- Sys.time()
end_time - start_time

#Clean cache to free up memory space
invisible(gc())

#########
#Step 2: Evaluate the performance of fit indices
#########

#Choose fit indices of interest
GOFs <- c("chisq.scaled","cfi.scaled", "rmsea.scaled", "srmr")

#Which of the fit indices of interest are reverse coded, that is higher values imply better
fit (e.g., CFI)?
#if no reversed fit indices are included please write GOFs.reverse <- c()
GOFs.reverse <- c("cfi.scaled")

#Set AUC cutpoint --> you may play around with it
AUC.cut <- 0.800
#AUC.cut <- 0.700 # to reproduce example of approximate fit

#Plot ROC curve(s)
plotROC(GOFs = GOFs, GOFs.reverse = GOFs.reverse, AUC.cut = AUC.cut, data = fit)

#########
#Step 3: Generate tailored cutoffs
#########
cutoffPerform(GOFs = GOFs, GOFs.reverse = GOFs.reverse, AUC.cut = AUC.cut, data = fit)

##############################################################################################
#########
#Output: Evaluate the fit of the analysis model to empirical data with tailored cutoffs
########

#################
#Test of two-factor analysis model (exact/approximate fit)
#################
```

```
#Tabulate tailored cutoffs via the simulation-cum-ROC approach against empirical fit indices
modelfit(GOFs = GOFs, GOFs.reverse = GOFs.reverse, AUC.cut = AUC.cut, empirical.model =
cfa1.fit, simulated.data = fit)

#################
# Select for: Modification of the two-factor analysis model (exact fit)
#################
#Tabulate tailored cutoffs via the simulation-cum-ROC approach against empirical fit indices
#modelfit(GOFs = GOFs, GOFs.reverse = GOFs.reverse, AUC.cut = AUC.cut, empirical.model =
cfa2.fit, simulated.data = fit)


############################
#REFERENCES
#Kemper, C. J., Beierlein, C., Bensch, D., Kovaleva, A., & Rammstedt, B. (2014). Soziale
Erwuenschtheit-Gamma (KSE-G). Zusammenstellung sozialwissenschaftlicher Items und Skalen
(ZIS). https://doi.org/10.6102/zis186
#Niessen, D., Partsch, M. V., Kemper, C. J., & Rammstedt, B. (2019). An English-language
adaptation of the Social Desirability-Gamma Short Scale (KSE-G). Measurement Instruments for
the Social Sciences, 1, Article 2. https://doi.org/10.1186/s42409-018-0005-1
#Niessen, D., Partsch, M., Rammstedt, B. (2018). Data for: An English-Language Adaptation of
the Social Desirability-Gamma Short Scale (KSE-G) (Version: 1.0.0).
https://doi.org/10.7802/1752
```

Additional File 2: Generate Tailored Cutoffs for Additional Models of the Social Desirability-Gamma Short Scale

Modified Two-Factor Model

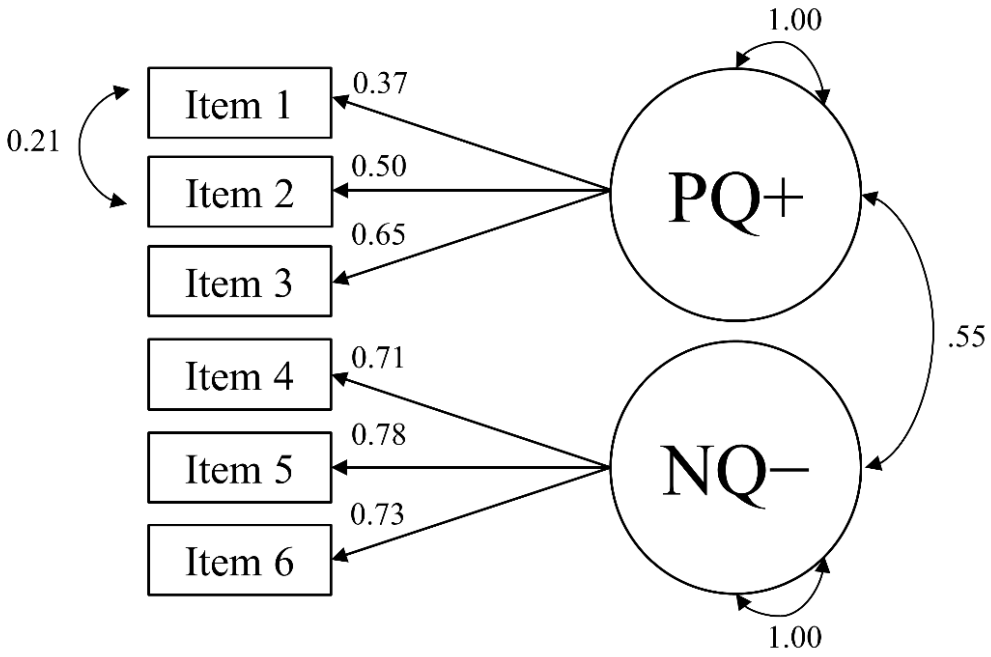
Tailored cutoffs showed that the two-factor Social Desirability-Gamma Short Scale model did not exactly fit the data. The modification indices (MI) suggested including a residual covariance between the first and second item of the PQ+ factor (exaggerating one's positive qualities; MI = 25.44). Including such a residual covariance was not only methodologically but also theoretically justified as both items ask for emotional control (Item 1, PQ+, English-language version: "In an argument, I always remain objective and stick to the facts." Item 2, PQ+, English-language version: "Even if I am feeling stressed, I am always friendly and polite to others."). In the following, we derived tailored cutoffs via the simulation-cum-ROC approach to testing the modified two-factor model (i.e., the two-factor model with a residual covariance between Items 1 and 2 of PQ+) in empirical data (see Figure S1).

We took the structure and parameter estimates from the empirical modified two-factor model as the H_0 population model. A plausible H_1 population model might be like the H_0 population model but with two additional residual correlations of r = .50. We chose an additional residual correlation between the first and last item of the NQ− factor (resulting in a covariance of 0.31), both referring to behavior in interactions. We chose another residual correlation between the first and last item of the PQ+ factor (resulting in a covariance of 0.19), both referring to conversations. We simulated data from the H_0 and H_1 population models, aligning with the characteristics of the empirical setting. We used the modified two-factor model as an analysis model and fit it to all simulated data (Figure S2). Figure S3 shows that χ^2, CFI, RMSEA, and SRMR had an AUC of .80 or higher. SRMR had the highest AUC (= .98). The recommended cutoff for χ^2 was 12.59, for CFI .994, for RMSEA .041, and for SRMR .025. As suggested by the AUC, the SRMR cutoff had the highest accuracy (= .94), as well as the lowest Type I error rate (= 8%) and Type II error rate (= 4%), as visualized in Figure S4. The SRMR had the best ability to demarcate between correctly specified and misspecified models.

The empirical values of χ^2, RMSEA, and SRMR passed the corresponding cutoffs, and the empirical value of CFI failed its cutoff (although it was very close to it; compare Figures S1 and S4). We accepted the modified two-factor model because χ^2, RMSEA, and SRMR had a better discrimination ability (i.e., higher AUCs) than CFI in the setting of interest. A two-

factor population model with a residual covariance between Items 1 and 2 of PQ+ seemed to have generated the data.

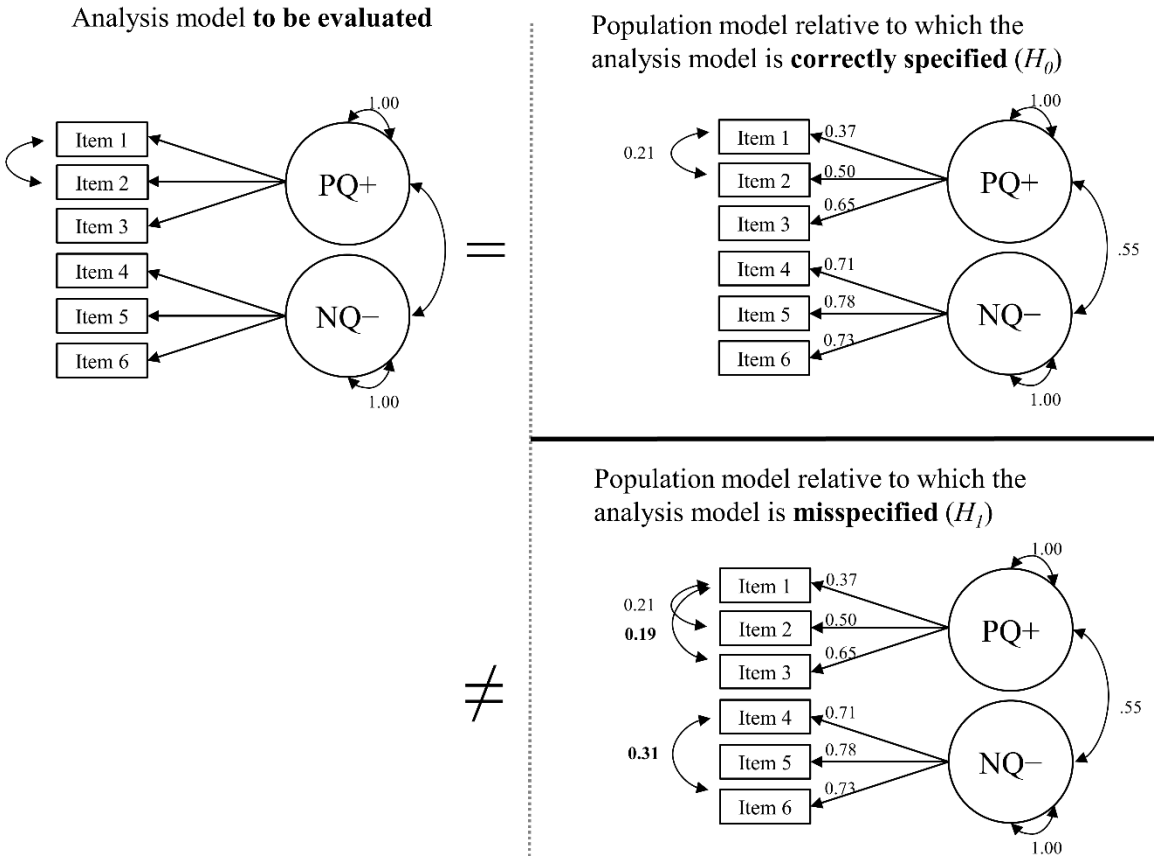
Figure S1: *Empirical Modified Two-Factor Social Desirability-Gamma Short Scale Model*



Fit indices	χ^2	*df*	CFI	RMSEA	SRMR
Empirical values	10.42	7	.992	.032	.023

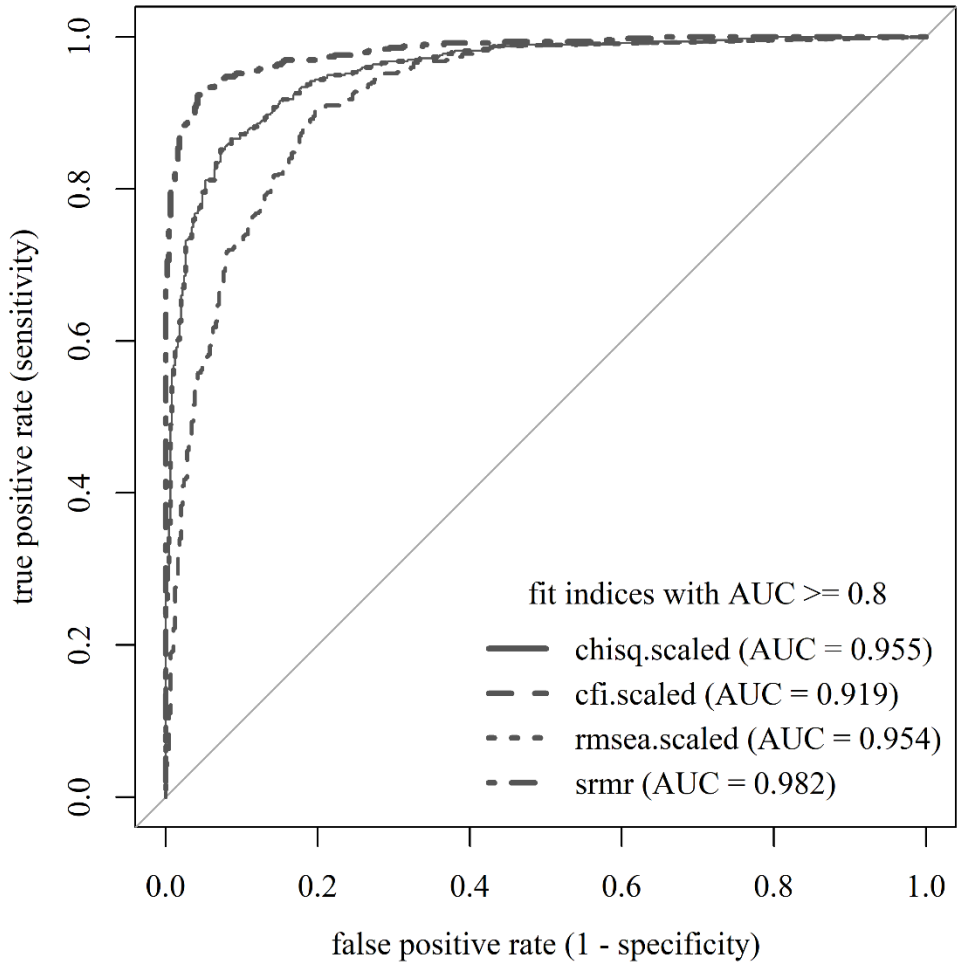
Note. Unstandardized coefficients. PQ+ = exaggerating positive qualities; NQ− = minimizing negative qualities. We recoded NQ− so that higher values imply more socially desirable responses. We omitted the residual variances and the mean structure for clarity. $N = 474$. *** $p < .001$.

Figure S2: *Proposed Modified Analysis and Population Models of the Social Desirability-Gamma Short Scale*



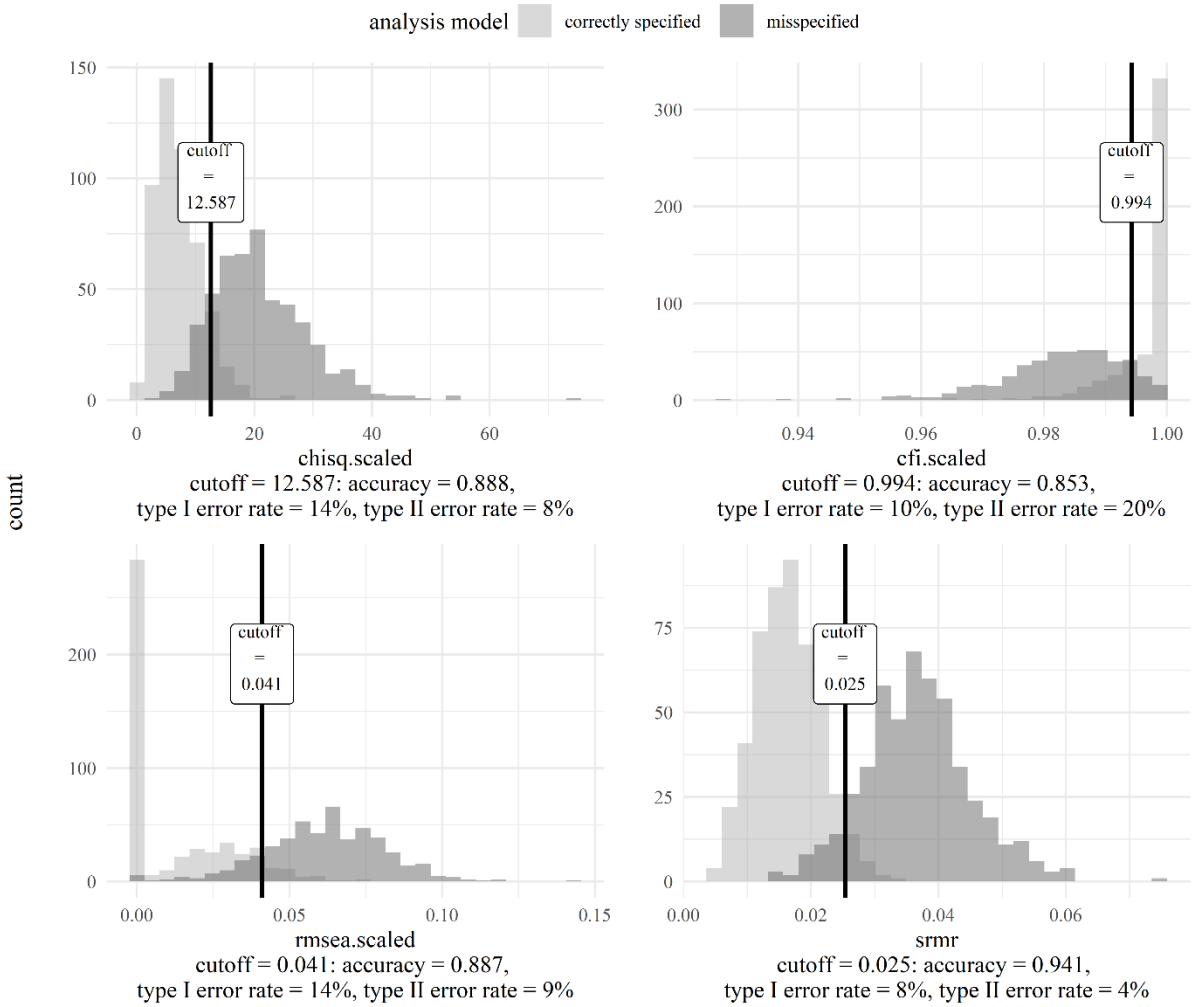
Note. We simulated data from both population models and fit the same analysis model to the data. Because the analysis model was structurally identical to the H_0 population model, it was correctly specified when fit to data generated from that population model. Because the analysis model differed from the H_1 population model, it was misspecified when fit to data generated from that population model. Unstandardized coefficients. PQ+ = exaggerating positive qualities; NQ− = minimizing negative qualities. We recoded NQ− so that higher values imply more socially desirable responses. We omitted the residual variances and the mean structure for clarity.

Figure S3: *ROC Curves for Fit Indices with AUC ≥ .80 of the Modified Social Desirability-Gamma Short Scale Model*



Note. Chisq.scaled is a χ^2 test statistic asymptotically equivalent to the robust Yuan-Bentler test statistic (Yuan & Bentler, 2000a) to account for non-normality. Cfi.scaled is the CFI version and rmsea.scaled is the RMSEA version calculated with this test statistic.

Figure S4*: Cutoffs for Fit Indices with AUC ≥ .80 of the Modified Social Desirability-Gamma Short Scale Model*



Note. Chisq.scaled is a χ^2 test statistic asymptotically equivalent to the robust Yuan-Bentler test statistic (Yuan & Bentler, 2000a) to account for non-normality. Cfi.scaled is the CFI version and rmsea.scaled is the RMSEA version calculated with this test statistic. The distribution colored in lighter gray originates from correctly specified models. The distribution colored in darker gray originates from misspecified models. Overlapping (parts of) distributions have an even darker gray color than the distribution from misspecified models. The vertical dash corresponds to the cutoff for each fit index (at the highest sum of sensitivity and specificity – 1).

Approximately Fitting Model

Here, we again generated tailored cutoffs for the initial two-factor model of the Social Desirability-Gamma Short Scale (see Figure S5). Unlike the example in the paper, we generated cutoffs for an approximately (instead of exactly) correctly specified model. Generating tailored cutoffs for approximately (instead of exactly) correctly specified models allowed us to incorporate the assumption of imperfect but acceptable model fit.

Like in the paper, we fit the two-factor model of the Social Desirability-Gamma Short Scale (Kemper et al., 2014; Nießen et al., 2019) to data from Germany (Nießen et al., 2018). Then, we defined the H_0 and H_1 population models (Figure S6, see Figure 8 of the paper). The structure and parameter estimates of the H_0 population model were identical to those from the two-factor empirical model. Additionally, the H_0 population model included two residual correlations of $r = .20$ that we considered small. We modeled a residual correlation between the first and second item of the PQ+ factor (resulting in a residual covariance of 0.08), which both ask for emotional control. Further, we modeled an additional residual correlation between the first and third item of the NQ− factor (resulting in a residual covariance of 0.12), which both refer to behavior in interactions. We used the two-factor model as an analysis model. When fitting the two-factor analysis model to data sampled from that H_0 population model, the analysis model was underspecified (or misspecified). We still labeled the two-factor analysis model fit to data sampled from that H_0 population model as (approximately) correctly specified. We were explicitly willing to accept minor misspecification (such as small residual correlations), for instance, to prevent overfitting.

We defined the H_1 population model in the same way as in the paper to compare cutoffs across examples for exact and approximate fit. The H_1 population model was identical to the H_0 population model except for two substantial residual correlations with $r = .50$. If the true population model were a two-factor one with (at most) two minor residual correlations ($r = .20$), we would accept the two-factor analysis model. If the true population model were different from a two-factor one, such as a two-factor model with two additional residual correlations of (at least) $r = .50$, we would reject the two-factor analysis model.

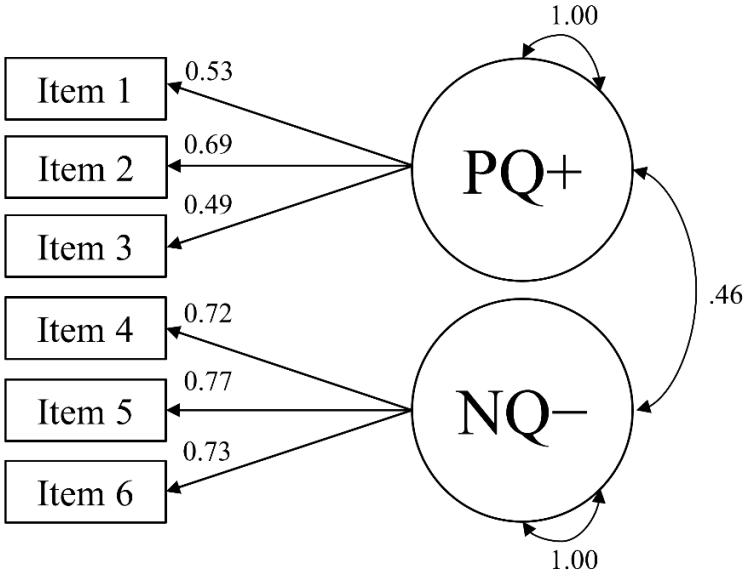
The discrimination ability of χ^2, RMSEA, and SRMR was worse for fit indices based on approximately instead of exactly correctly specified models (Figure S7, Figure 9). The AUCs ranged from .73 to .82 for the former and from .83 to .92 for the latter. Remember that the fit index distribution from misspecified models stayed the same. Thus, the fit index distribution from approximately correctly specified models overlapped stronger with the one

from misspecified models than the fit index distribution from exactly correctly specified models. The fit index distribution from approximately correctly specified models contained an (acceptable) form of misspecification. It was shifted towards the fit index distribution from (severely) misspecified models. Alike (Figure S8, Figure 10), cutoffs that balance Type I and Type II errors were (a bit) more lenient with approximately ($\chi^2(8) \leq 11.25$; RMSEA $\leq .029$; SRMR $\leq .028$) than exactly correctly specified models ($\chi^2(8) \leq 9.54$; RMSEA $\leq .020$; SRMR $\leq .025$). The accuracy, as well as the Type I and Type II error rates, were worse for the former (e.g., accuracy = 751; Type I error rate = 28%; Type II error rate = 21% for SRMR) than the latter (e.g., accuracy = .851; Type I error rate = 19%; Type II error rate = 11% for SRMR).

In this example, the conclusion was the same: The two-factor model of the Social Desirability-Gamma Short Scale must be rejected because the empirical values ($\chi^2(8) = 32.06$, $p < .001$; RMSEA = .080; SRMR = .048) failed all cutoffs.

So, what is the essential difference between using approximately instead of exactly correctly specified models? The crucial difference is that we explicitly model different assumptions of the population model. When using approximately correctly specified models, we explicitly model the assumption (and generate cutoffs accordingly) that the analysis model does not need to be perfectly identical to the population model to be correctly specified. When using exactly correctly specified models, we explicitly model the assumption (and generate cutoffs accordingly) that the analysis model needs to be perfectly identical to the population model to be correctly specified. Thus, the crucial difference of cutoffs based on exactly and approximately correctly specified models concerns the definition of the H_0 population model—we either incorporate the assumption of perfect or imperfect fit. In turn, the definition of the H_0 population model influences the resulting cutoffs, accuracy, Type I and Type error rates, and AUCs, and, thus, how we look at the empirical data.

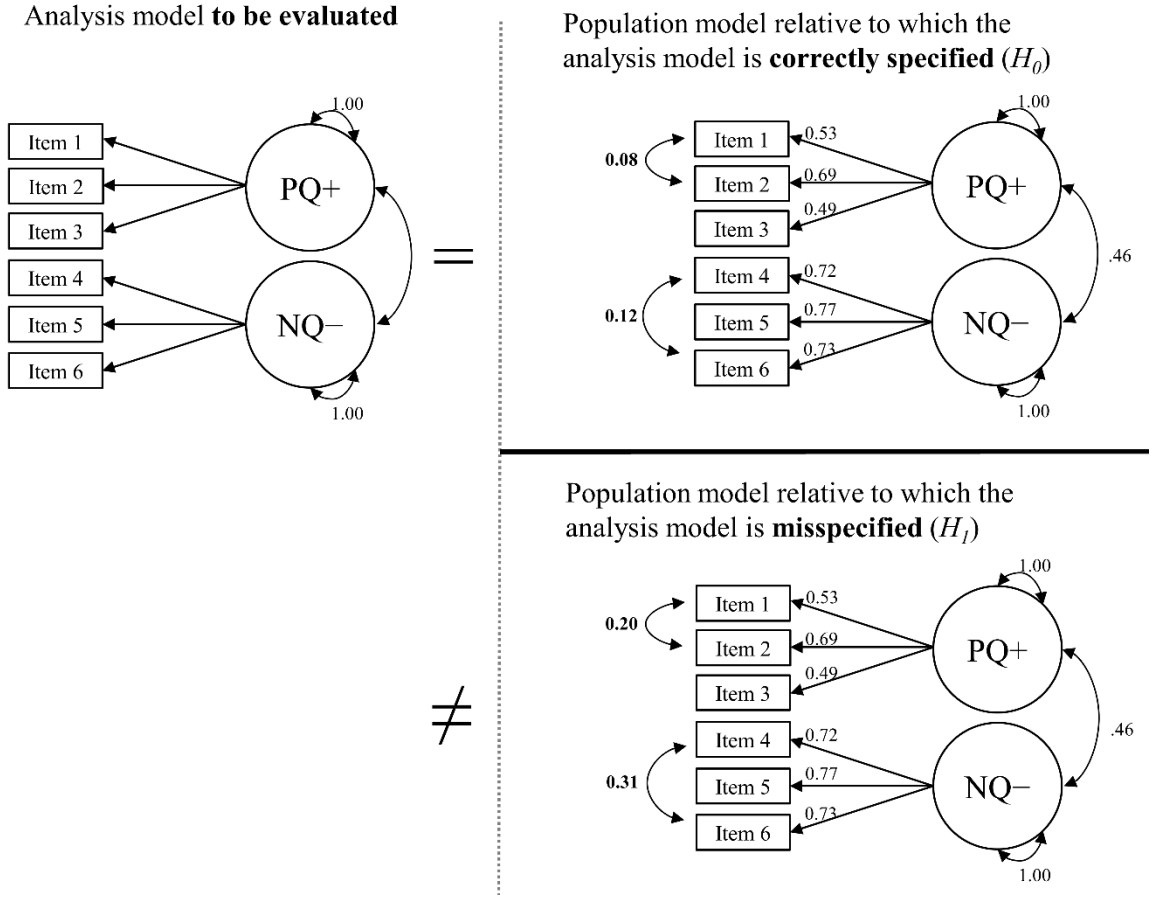
Figure S5: *Empirical Two-Factor Social Desirability-Gamma Short Scale Model*



Fit indices	χ^2	*df*	CFI	RMSEA	SRMR
Empirical values	32.06***	8	.947	.080	.048

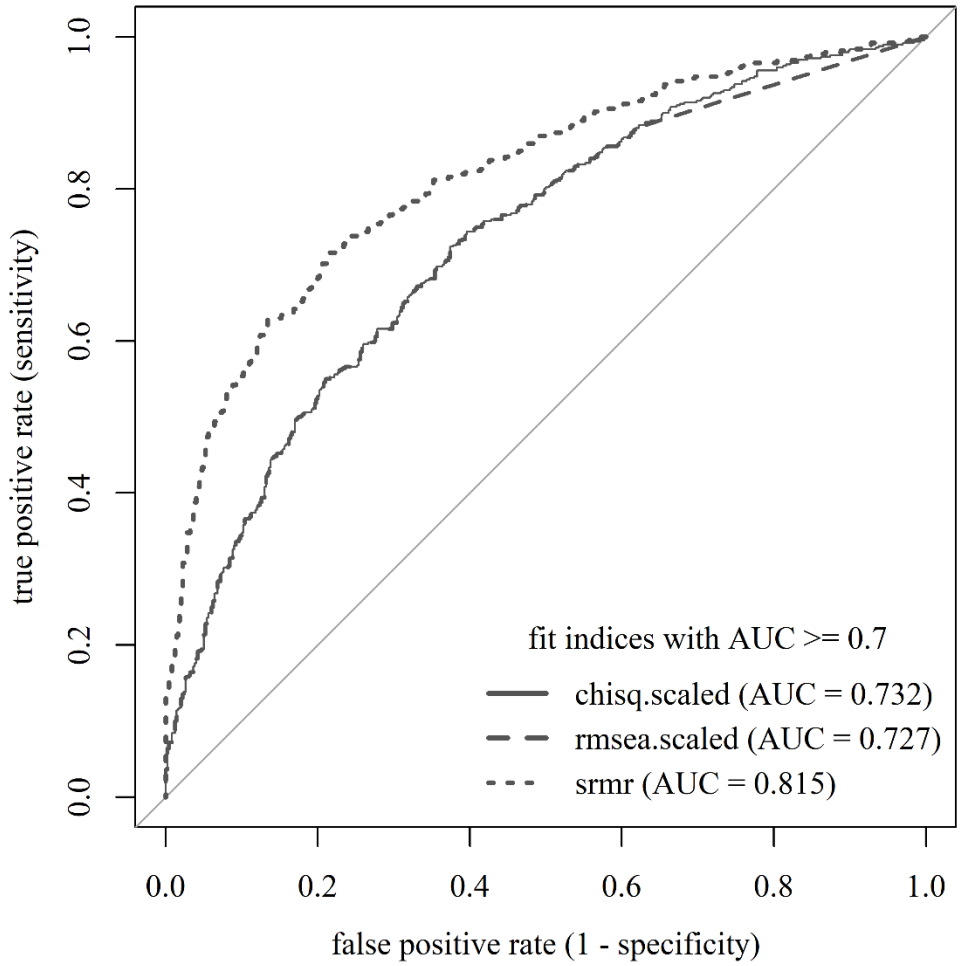
Note. Unstandardized coefficients. PQ+ = exaggerating positive qualities; NQ− = minimizing negative qualities. We recoded NQ− so that higher values imply more socially desirable responses. We omitted the residual variances and the mean structure for clarity. $N = 474$. *** $p < .001$.

Figure S6: *Proposed Analysis and Population Models of the Social Desirability-Gamma Short Scale for Testing Approximate Fit*



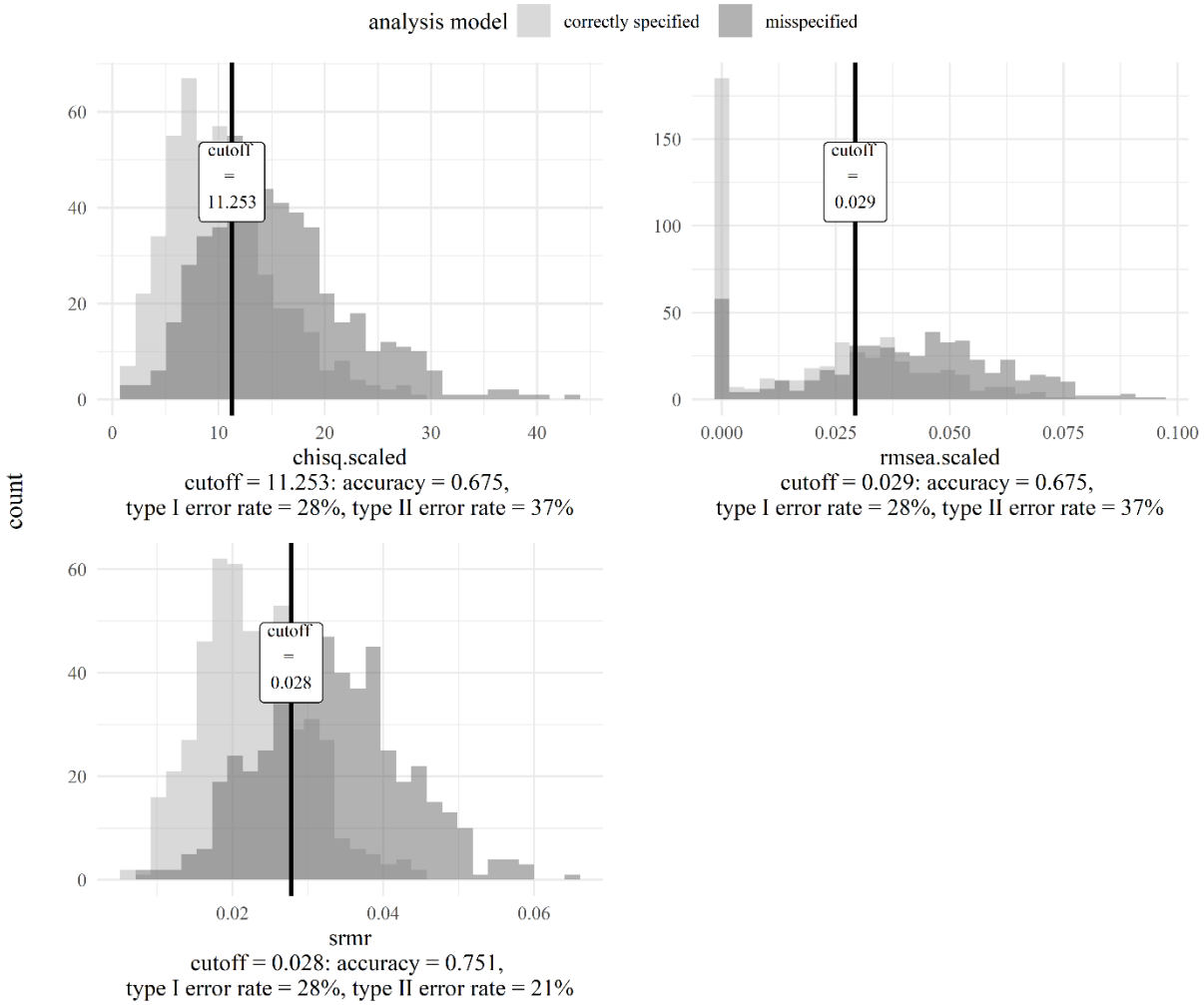
Note. We simulated data from both population models and fit the same analysis model to the data. Because the analysis model was structurally identical to the H_0 population model, it was correctly specified when fit to data generated from that population model. Because the analysis model differed from the H_1 population model, it was misspecified when fit to data generated from that population model. Unstandardized coefficients. PQ+ = exaggerating positive qualities; NQ− = minimizing negative qualities. We recoded NQ− so that higher values imply more socially desirable responses. We omitted the residual variances and the mean structure for clarity.

Figure S7: *ROC Curves for Fit Indices with AUC ≥ .70 Testing the Approximate Fit of the Social Desirability-Gamma Short Scale Model*



Note. Chisq.scaled is a χ^2 test statistic asymptotically equivalent to the robust Yuan-Bentler test statistic (Yuan & Bentler, 2000a) to account for non-normality. Rmsea.scaled is the RMSEA version calculated with this test statistic.

Figure S8*: Cutoffs for Fit Indices with AUC ≥ .70 Testing the Approximate Fit of the Social Desirability-Gamma Short Scale Model*



Note. Chisq.scaled is a χ^2 test statistic asymptotically equivalent to the robust Yuan-Bentler test statistic (Yuan & Bentler, 2000a) to account for non-normality. Rmsea.scaled is the RMSEA version calculated with this test statistic. The distribution colored in lighter gray originates from correctly specified models. The distribution colored in darker gray originates from misspecified models. Overlapping (parts of) distributions have an even darker gray color than the distribution from misspecified models. The vertical dash corresponds to the cutoff for each fit index (at the highest sum of sensitivity and specificity – 1).

Measurement Invariance Violation Indices (MIVIs): Effect sizes for (partial) non-invariance of items and item sets

Katharina Groskurth[1,2], Matthias Bluemke[1], and Clemens M. Lechner[1]

[1] GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany

[2] Graduate School of Economic and Social Sciences, University of Mannheim, Germany

Abstract

In many applications, measurement invariance does not hold. When a model with a certain level of invariance is rejected, the amount of non-invariance bias may either be consequential or practically irrelevant. So far, few attempts have been made to quantify the extent of bias due to the lack of measurement invariance. We derive new effect size measures from first principles called Measurement Invariance Violation Indices (MIVIs) for items and item sets. MIVIs assume that one can compare the basic measurement model across groups (i.e., configural invariance holds) but cannot compare some factor loadings, intercepts, and/or unique variances. Assuming partial invariance for a set of items, MIVIs quantify non-invariant factor loading, intercept, or uniqueness differences in relation to the pooled latent standard deviation (either per item or as an average for item sets). Thus, parameter differences can be interpreted in standard deviation units (of the pooled latent variable). One can further inspect the compensatory cancelation and non-compensatory aggregation effects of non-invariance bias when maintaining the directional information (signed MIVIs). MIVIs support the group-fair item selection, help to evaluate the questionnaire quality, and allow for assessing the amount of non-invariance bias when comparing simple, observed statistics (e.g., mean scores) across groups.

Introduction

Statistical comparisons are ubiquitous in the social sciences. Researchers may, for instance, compare numeracy scores of an exam across two school classes or the learning motivation across two entire school districts. Such comparisons are only valid if the test captures the numeracy scores, alternatively learning motivation, without bias across the two groups. The latent construct (e.g., numeracy or learning motivation) must be placed on a comparable scale. Otherwise, researchers compare apples with oranges (Chen, 2008). Examining the validity of numerical group comparisons is called measurement invariance testing in linear latent variable modeling (and differential item functioning in item response theory frameworks; Osterlind & Everson, 2009).

Crucially, some statistical comparisons (i.e., including variances or means) are only valid if specific levels of measurement invariance hold. If metric invariance does not hold, any differences (or lack of differences) in latent (co)variances might be spurious (Steenkamp & Baumgartner, 1998). If scalar invariance does not hold, any differences (or lack of differences) in latent or observed means might be spurious (Steenkamp & Baumgartner, 1998; Steinmetz, 2013; cf. DeShon, 2004; Wu et al., 2007). The uniqueness invariance level answers the question of whether any differences (or lack of differences) in observed (co)variances might be spurious (Steenkamp & Baumgartner, 1998; Millsap & Olivera-Aguilar, 2012).

Despite the necessity of establishing measurement invariance, applied researchers often compare group-level statistics (such as mean scores) without testing for measurement invariance (e.g., Boer et al., 2018). When researchers do report measurement invariance tests, they frequently find that invariance does not hold; especially uniqueness or scalar invariance is often hard to reach (e.g., Davidov et al., 2012, 2014; Dong & Dumas, 2020). A practical question thus arises: How consequential are any violations of invariance? In other words: How much non-invariance bias is present when comparing relevant group-level statistics (e.g., mean scores)?

This work proposes an intuitive way to empirically quantify non-invariance bias in factor loadings, intercepts, and/or unique variances. Whereas researchers should refrain from comparing statistics (such as mean scores) across groups when non-invariance bias is large (e.g., intercepts differ strongly due to the different use of response styles), small non-invariance bias may not be detrimental. Several effect size measures currently compete for acceptance among psychometricians and more widespread recognition by applied researchers (for a

review, see Gunn et al., 2020). Our proposed effect size measures, termed Measurement Invariance Violation Indices (MIVIs), quantify the differences in measurement model parameters (i.e., factor loadings, intercepts, or unique standard deviations) in units of latent standard deviations pooled across groups. MIVIs provide an intuitive metric for bias in items and item sets. The indices are readily implemented in standard structural equation modeling (SEM) software for confirmatory factor analysis (CFA). We provide examples for the few code lines that need to be added in Mplus or R to estimate MIVIs.

Testing for Measurement Invariance

Basics of Measurement Invariance Testing

Before defining MIVIs as effect size measures, we outline the basics of measurement invariance testing (Davidov et al., 2014; Millsap, 2011; Millsap & Olivera-Aguilar, 2012; Steenkamp & Baumgartner, 1998). Assuming a common-factor model holds, one can test measurement invariance across multiple groups. In the following, we consider a multi-group application in a unidimensional measurement scenario for individual *i*: Let $\mathbf{x}$ be a $p \times 1$ vector of observed measures, such as a set of approximately continuous rating scores across *p* items obtained from a social survey or psychological test. Let w_i be the individual score on the common factor forming the latent dimension underlying the correlated *p* measures to achieve model identification (if necessary, by additional constraints). We measure w_i by a set of items $\mathbf{x}_{ik}$ for an individual who belongs to a specific group, such as language, nationality, gender, school class, or age group. We can then express the congeneric factor model for the *k*th subgroup ($k = 1, ..., K$), assuming *p* items and one common factor, as

$$\mathbf{x}_{ik} = \boldsymbol{\tau}_k + \boldsymbol{\lambda}_k w_{ik} + \mathbf{u}_{ik}. \quad (1)$$

Here, $\boldsymbol{\tau}_k$ is a $p \times 1$ vector of regression intercept parameters, $\boldsymbol{\lambda}_k$ is a $p \times 1$ vector of factor loading parameters, and $\mathbf{u}_{ik}$ is a $p \times 1$ vector of unique scores. In line with classical test theory, unique scores are assumed to be statistically independent (i.e., uncorrelated) and do not correlate with common scores of the latent variable. The variance-covariance matrix, $\boldsymbol{\Sigma}$, of the factor model in the *k*th subgroup is

$$\boldsymbol{\Sigma}_k = \boldsymbol{\lambda}_k \Phi_k \boldsymbol{\lambda}_k' + \boldsymbol{\Theta}_k. \quad (2)$$

Here, Φ_k is the variance of the latent scores $\mathbf{w}_k$ and $\boldsymbol{\Theta}_k$ is the variance matrix of the unique scores $\mathbf{u}_k$. Then, the observed item means $E(\mathbf{x}_k)$ are related to the latent factor mean κ_k in the *k*th subgroup as

$$E(\mathbf{x}_k) = \boldsymbol{\tau}_k + \boldsymbol{\lambda}_k \kappa_k. \quad (3)$$

Because the expected means of uniqueness terms are zero, they drop from Equation 3 (Millsap, 2011; Millsap & Olivera-Aguilar, 2012; Steenkamp & Baumgartner, 1998).[1]

Measurement invariance holds if a multi-group CFA model fits after constraining cross-group parameters to equality: For a model to be invariant, the same item-to-factor structure (i.e., configural invariance), factor loadings (i.e., metric invariance; $\boldsymbol{\lambda}_1 = \ldots = \boldsymbol{\lambda}_K$) and intercepts (i.e., scalar invariance, $\boldsymbol{\tau}_1 = \ldots = \boldsymbol{\tau}_K$) must be equivalent across the *K* groups. For strict invariance to hold, even the unique item variances must be equal (i.e., uniqueness invariance, $\boldsymbol{\Theta}_1 = \ldots = \boldsymbol{\Theta}_K$). Readers can revisit the logic of invariance testing in Additional File 1 of the Supplementary Materials. Additional File 1 not only contains a detailed description of the traditional, sequential approach to measurement invariance testing (e.g., Meredith, 1993; Millsap, 2011; Steenkamp & Baumgartner, 1998) but also the description of a less well-known approach to measurement invariance testing by Raykov et al. (2013). As an alternative to the traditional bottom-up approach, Raykov et al. (2013) suggested a top-down procedure for measurement invariance testing.

The Problem of Establishing Measurement Invariance

Measurement invariance is indispensable for specific statistical comparisons to be valid but often hard to reach in empirical settings (Davidov et al., 2012, 2014). Lack of invariance hampers researchers' ability to perform group comparisons. These comparisons, all too often, motivated the research in the first place (e.g., rankings of countries, comparability of validity coefficients across gender, or mean-level comparisons across age groups).

However, the more restrictive the invariance level, the harder it is to reach. In Dong and Dumas's (2020) meta-analysis of personality constructs (across culture, gender, and age groups), only 8 out of 67 group comparisons (11.9%) reached uniqueness invariance, and only 22.4% reached scalar invariance. Most group comparisons only reached metric invariance (35.8%). Similarly, Zercher and colleagues (2015) conducted a measurement invariance test of items measuring Schwartz's human value of Universalism across 90 groups. They found that metric invariance held for all groups, but (partial) scalar invariance held for only 37 out of 90 groups (41.1%). In Dong and Dumas's (2020) meta-analysis of personality constructs, about a

[1] DeShon (2004) and Wu et al. (2007) argued that the means of uniqueness terms are rarely zero in reality. Oftentimes factor models are not correctly specified and unmodeled latent variables lead to means of uniqueness terms that are different from zero.

quarter reached merely configural invariance (23.9%). Some group comparisons did not even attain configural invariance (6.0%). Chen (2008), who conducted a meta-analysis of cross-cultural studies, supported the finding and found that 9 out of 130 group comparisons lacked configural invariance (6.9%).

If uniqueness, scalar, or metric invariance is unattainable, there is always the option to test for partial invariance (Byrne et al., 1989). Equality constraints on the factor loadings, intercepts, or uniquenesses are relaxed for some items, whereas the rest remains constrained (i.e., invariant). For partial invariance to hold, a pure CFA measurement model requires at least two equivalent items across groups (in terms of factor loadings, intercepts, or uniquenesses) with at least one group-specific (i.e., freely estimated) item (Byrne et al., 1989; Steenkamp & Baumgartner, 1998; Steinmetz, 2013). Partial invariance enables researchers to compare latent group statistics; this does not require full invariance. Differently, comparing observed group statistics requires full invariance (e.g., Steinmetz, 2013).

Underutilization of Effect Sizes in Invariance Testing

Typically, invariance testing stops after reaching either full or partial invariance in a binary accept-or-reject logic (see Additional File 1 of the Supplementary Materials for an overview). The binary accept-or-reject logic of measurement invariance testing discards essential information: When a certain invariance level does not (fully) hold, it remains unknown how large and serious the resulting bias in cross-group comparisons of observed statistics (e.g., group differences in mean scores) would be. Effect sizes can complement traditional model comparisons by enabling the researcher to quantify the size of non-invariant model parameter differences. In the development process of item sets, one might consider discarding items with large bias due to non-invariance but keeping those with only negligible bias. In fixed item sets, the size of non-invariance bias helps researchers evaluate (next to other information: McNeish & Wolf, 2020; Widaman & Revelle, 2022) whether cross-group comparisons of observed, simplified aggregates of the item set (such as unit-weighted mean scores) can be valid ways of approximating the cross-group differences in the latent variable representing the levels of the construct. If effect sizes indicate that non-invariance bias (a) adds up at the level of simplified aggregates or (b) is large for specific items or the whole item set, one might better refrain from comparing observed statistics across groups. If effect sizes indicate that non-invariance bias (a) cancels out at the level of simplified aggregates or (b) is so tiny that it is practically inconsequential, one might consider comparing observed statistics across groups.

Millsap and Olivera-Aguilar (2012) provided such effect size measures for factor loadings, intercepts, and unique variances, which we call difference measures here.[1] To define a difference measure for factor loadings, Millsap and Olivera-Aguilar (2012, see also Pornprasertmanit, 2022) assumed a unidimensional CFA model that shows partial metric invariance across two groups ($K = 2$). Item j ($j = 1, \ldots, p$) has non-invariant loadings across groups. That is, loading λ_{1j} of item j in group 1 differs from loading λ_{2j} of item j in group 2, which are both, thus, freely estimated across groups. Millsap and Olivera-Aguilar are not explicit about the intercepts; however, to quantify only the loading difference, intercept τ_{1j} of item j in group 1 must be equal to intercept τ_{2j} of item j in group 2. Otherwise, the loading difference measure does not only quantify the non-invariant loading difference but also the intercept difference. At any common latent score w, Millsap and Olivera-Aguilar expressed the expected difference in observed item scores x_j between the two groups as

$$\begin{aligned} Loading\ Difference_j &= |E(x_{2j} - x_{1j}|w)| \\ &= |(\tau_{2j} - \tau_{1j}) + (\lambda_{2j} - \lambda_{1j})w| \,. \end{aligned} \quad (4)$$

Equation 4 represents the expected absolute difference in raw metrics between two people's item scores (both from different groups) that arise due to loading invariance despite identical values on the latent variables. Millsap and Olivera-Aguilar proposed to quantify the impact of loading differences (and/or intercept differences to be correct) on expected differences in observed item scores in three steps:

(1) Define a meaningful difference on the observed scale, which might be $x_d = x_{2j} - x_{1j}$. Millsap and Olivera-Aguilar did not propose guidelines for defining a meaningful difference in observed item scores but left it open to the applied researcher.

(2) Calculate the range of common latent scores $\mathbf{w}$ for which $\left|E(x_{2j} - x_{1j}|\mathbf{w})\right| \leq x_d$.

(3) Evaluate how many individuals are in the region $\left|E(x_{2j} - x_{1j}|\mathbf{w})\right| > x_d$. Means and variances of $\mathbf{w}$ estimated in each group (i.e., the latent variance Φ_k and the latent mean κ_k) should help this evaluation.

If the model has no loading differences (i.e., if metric invariance holds), the loading difference measure (Equation 4) is not useful anymore because the expected observed score difference

[1] Other approaches for quantifying measurement non-invariance, which we do not review here, do exist (e.g., Nye & Drasgow, 2011; Oberski, 2014). Yet, other effect size approaches are not as easy to apply as the effect size measures of Millsap and Olivera-Aguilar (2012).

will always equal the raw intercept difference: $E(x_{2j} - x_{1j}|w) = x_{2j} - x_{1j}$. Thus, Millsap and Olivera-Aguilar developed a specific intercept difference measure.

To define their intercept difference measure, Millsap and Olivera-Aguilar (2012) assumed a unidimensional CFA model showing partial scalar invariance across two groups (K = 2). Item j (j = 1, …, p) has non-invariant intercepts across groups. That is, intercept τ_{1j} of item j in group 1 differs from intercept τ_{2j} of item j in group 2, which are both, thus, freely estimated across groups. Millsap and Olivera-Aguilar's intercept difference measure represents the proportion of the cross-group difference in item means attributable to the difference in intercepts:

$$Intercept\ Difference_j = \frac{|\tau_{2j} - \tau_{1j}|}{|E(x_{2j}) - E(x_{1j})|}. \quad (5)$$

Here, $E(x_{1j})$ and $E(x_{2j})$ are approximated by the observed means of item j for groups 1 and 2 (alternatively by Equation 3). Thus, the numerator denotes the intercept difference of item j; the denominator reflects the mean difference of item j. If there is no intercept difference between groups (i.e., the numerator equals zero), the intercept difference measure will be zero (unless the item mean difference is zero, in which case the ratio is undefined).

Further, Millsap and Olivera-Aguilar (2012) developed a specific uniqueness difference measure to quantify non-invariance bias due to differing unique variances. For the uniqueness difference measure, Millsap and Olivera-Aguilar assumed a unidimensional CFA model showing partial uniqueness invariance across two groups (K = 2). Item j (j = 1, …, p) has non-invariant unique variances across groups. That is, unique variance θ_{1j} of item j in group 1 differs from unique variance θ_{2j} of item j in group 2, which are both, thus, freely estimated across groups. Millsap and Olivera-Aguilar's uniqueness difference measure represents the proportion of the cross-group difference in item variances attributable to the difference in unique variances:

$$Uniqueness\ Difference_j = \frac{|\theta_{2j} - \theta_{1j}|}{|\sigma^2_{2j} - \sigma^2_{1j}|}. \quad (6)$$

The item variance σ^2_{kj} can be approximated by the observed item variance of item j in group k (alternatively by $\sigma^2_{kj} = \lambda'_j \Phi_k \lambda_j + \theta_{kj}$ including loading λ_j of item j and latent variance Φ_k in group k). The numerator reflects the difference in unique variances of item j; the denominator reflects the difference in total variances of item j. If there is no difference in unique variances between groups (i.e., the numerator equals zero), the uniqueness difference measure will be zero (unless the item variance difference is zero, in which case the ratio is undefined).

Millsap and Olivera-Aguilar's (2012) difference measures help quantify non-invariance. They are easy to implement in standard software such as Mplus (Muthén & Muthén, 1998-2017) or R (using the lavaan package; Rosseel, 2012). Still, they have three crucial problems: First, they do not come with a unified rationale spanning different model parameters (i.e., factor loadings, intercepts, and unique variances). Although one should not compare effect sizes for different model parameters, an intuitive logic that applies to all parameters will surely aid the understanding and, in turn, the applicability of a set of effect size measures. Second, Millsap and Olivera-Aguilar only vaguely defined what constitutes a meaningful raw-metric difference in observed item scores, given a specific score on the latent variable, which hampers at least the application of the loading difference measure. Third, the denominators of Millsap and Olivera-Aguilar's intercept and uniqueness difference measures (i.e., the difference in item means or item variances) change their units with each item. Accordingly, researchers can only evaluate the bias item-by-item. Effect size measures that apply to single items and item sets alike would be favorable.

Objective

We develop new effect size measures to quantify the degree of non-invariance, which we term Measurement Invariance Violation Indices (MIVIs). MIVIs are derived for non-invariance in three different model parameters (i.e., factor loadings, intercepts, and unique variances) and for each model parameter in four different versions (item/item set × absolute/signed). Thus, MIVIs build on the difference measures proposed by Millsap and Olivera-Aguilar (2012) but improve on them in three regards: First, MIVIs quantify non-invariance bias with the same rationale (i.e., denominator) for all model parameter differences (in the numerator, i.e., factor loadings, intercepts, and unique variances). This makes MIVIs easy to grasp. Second, MIVIs have identical points of reference (i.e., denominators) comparable across multiple items. Consequently, they are informative about systematic—compensatory or non-compensatory—non-invariance bias. As a third asset, MIVIs consider sampling error by providing bootstrap confidence intervals around the point estimate (Cumming & Finch, 2001). MIVIs are also readily implemented in standard statistical programs such as Mplus or R.

Depending on the specific research stage (e.g., scale development vs. application of an extant and well-defined item set for measuring a construct comparably across groups), MIVIs have various advantages for quantifying the amount of non-invariance bias. In scale

development settings, MIVIs can support the group-fair item selection (i.e., the decision on keeping non-invariant items or dropping them from the item set). In settings with fixed item sets, MIVIs can help evaluate the quality of a questionnaire in a new context (e.g., an international large-scale assessment, say, with new countries onboarding a survey program). MIVIs are also helpful in evaluating the amount of bias in comparing observed statistics (e.g., mean scores) across groups.[1] We highlighted all advantages of MIVIs in Table 1.

Table 1: *Advantages of MIVIs*

	Absolute	Signed
Item level	**Scale development:** + MIVIs support group-fair item selection (decision on keeping or dropping non-invariant items)	
Item set level	**Settings with fixed item set:** + MIVIs evaluate the quality of a questionnaire	**Settings with fixed item set:** + MIVIs evaluate the amount of bias in comparing observed statistics across groups

[1] Notably, calculating mean scores requires additional assumptions such as tau-equivalent or parallel measurements. However, methodologists still debate about the necessity to which these assumptions must hold before using mean scores (McNeish & Wolf, 2020; Widaman & Revelle, 2022).

Measurement Invariance Violation Indices (MIVIs)

Preconsiderations and Prerequisites

Like Millsap and Olivera-Aguilar's (2012) difference measures, MIVIs are estimated based on a tenable (i.e., well-fitting) partial invariance model with at least one freely estimated parameter of interest (i.e., factor loadings, intercepts, or unique variances). We assume models to be unidimensional. When models are multidimensional (assuming no cross-loadings), MIVIs should be separately estimated for each common factor (or latent variable). Before defining MIVIs, we discuss the key idea of the new effect size measures.

The key idea of MIVIs is to use the pooled (across groups) standard deviation of the latent variable as a denominator and the (unstandardized) parameter differences of interest as a numerator. The pooled standard deviation $SD_{\text{LV, pooled}}$ of the latent variable (Cohen, 1988) is defined as

$$SD_{LV,\ pooled} = \sqrt{\frac{(n_2-1)\Phi_2 + (n_1-1)\Phi_1}{n_1+n_2-2}}. \quad (7)$$

One may ask why we chose the pooled standard deviation of the latent variable as a common denominator for all MIVIs, instead of the (observed) pooled standard deviation of an item or the item set (e.g., Pornprasertmanit, 2022). The pooled standard deviation of the latent variable as a common denominator is favorable in at least three regards: (1) It is the same across items in an item set, (2) it is independent of the number of items in an item set, and (3) it solely consists of true score variance.

First, a problem with an item's pooled standard deviation is that it differs across items of an item set. Consequently, not only differences in parameters of interest (i.e., factor loadings, intercepts, or unique variances) but also differences in the pooled standard deviation of the item would impact the effect size. The effect size would not be comparable across items of an item set; information about systematic—compensatory or non-compensatory—non-invariance bias would remain hidden. By using the pooled standard deviation of the latent variable as a common denominator, we make MIVIs comparable across items of an item set.

Second, a problem with an item set's pooled standard deviation is that it changes dramatically with the number of items in it. (The pooled standard deviation of the item set is estimated by the sum of item variances plus two times all item covariances, e.g., Nye & Drasgow, 2011). Thus, the effect size would drastically depend on the number of items in an

item set. Differently, the pooled standard deviation of the latent variable is relatively independent of the number of items in an item set.

Third, another problem with both the pooled standard deviation of the item and the item set is that they compound true score variation with fluctuating error variation. Differently, the pooled standard deviation of the latent variable *is* the true score variation.

However, taking the pooled standard deviation of the latent variable as a common denominator for an effect size also has its limitations: The effect size depends on the identification method. Typically, researchers use the (pooled) standard deviation per item to standardize model parameters (i.e., factor loadings, intercepts, or unique variances; Muthén, 1998-2004; Pornprasertmanit, 2022). Parameter estimates standardized with the (pooled) standard deviation per item are independent of the identification method (e.g., Putnick & Bornstein, 2016). That is, estimates are the same, independent of whether the model is identified by fixing the latent variance and latent mean or by fixing an anchor item's loading or intercept to a certain value. Because MIVIs include raw parameters in the numerator, standardized by the (pooled) standard deviation of the latent variable (not the item), MIVIs change with the identification method. To obtain common ground for MIVIs, we propose identifying the final model by fixing the reference group's latent variance to 1 (and its latent mean to 0). As MIVIs are based on partial invariance models, latent variances can be freely estimated in all groups other than the reference group.

Effect Size Measures for Items

Absolute Effect Size Measures

To define MIVIs for items, we start on the assumption of a tenable partial invariance model with at least one freely estimated factor loading, intercept, or unique standard deviation across groups.[1] We define MIVIs for an (unstandardized) loading λ_{kj}, intercept τ_{kj}, or unique standard deviation $\sqrt{\theta_{kj}}$ of the non-invariant item *j* in group *k* as

$$MIVI\text{-}L_{Item\ j|absolute} = \frac{|\lambda_{2j} - \lambda_{1j}|}{SD_{LV,\ pooled}}, \tag{8a}$$

[1] We include differences in unique standard deviations instead of differences in unique variances in the numerator to be consistent with the entity of the denominator (which is also a standard deviation). If one prefers to retain the metric of unique variances (as it is in the measurement model), one might use differences in unique variances in the numerator as an alternative version of $MIVI - U$.

$$MIVI\text{-}I_{Item\ j|absolute} = \frac{|\tau_{2j} - \tau_{1j}|}{SD_{LV,\ pooled}}, \quad (8b)$$

$$MIVI\text{-}U_{Item\ j|absolute} = \frac{\left|\sqrt{\theta_{2j}} - \sqrt{\theta_{1j}}\right|}{SD_{LV,\ pooled}}. \quad (8c)$$

These MIVIs pertain to the item level, are naturally bounded at zero on one end, and reflect an absolute loading, intercept, or uniqueness difference relative to the pooled standard deviation of the latent variable.

To adequately interpret and understand MIVI, one needs to revisit the logic of Equation 3: In a common factor model, common scores on the latent variable and observed item scores relate to each other through linear regression of the observed scores on the latent scores (e.g., Wu et al., 2007, for a visualization). Whereas loadings represent the steepness (or slope/weight) of regression lines, intercepts represent their origin. Unique scores describe the differences between the scores of the latent variable on the regression line and the observed item scores. For a factor model to be invariant across groups, the regression lines that relate the observed units to the latent property should be identical across groups. Further, the variance of the unique scores should be identical across groups (i.e., observed scores have an equal precision across groups; see DeShon, 2004). MIVIs quantify the differences in those regression lines across groups: $MIVI\text{-}L_{Item\ j|absolute}$ quantifies the differences in regression weights, $MIVI\text{-}I_{Item\ j|absolute}$ the differences in regression origins, and $MIVI\text{-}U_{Item\ j|absolute}$ the differences in the standard deviations of unique scores. Ideally for group comparisons, all differences would be zero.

The different meanings of the three parameters (i.e., factor loadings, intercepts, and unique variances represent different features of a regression) have two important implications. First, an intercept and a unique variance should only be fixed across groups if the corresponding loading can be fixed without inducing misfit. If a loading is non-invariant across groups, the item's intercept (as well as its unique variance) should not undergo invariance tests. They, too, should be estimated freely across groups. Thus, metric non-invariance of an item implies scalar and uniqueness non-invariance of that item. If the units (i.e., loadings) differ across groups, unit shifts (i.e., intercepts) and unique variances are barely comparable (theoretically and, often, empirically; e.g., Millsap & Olivera-Aguilar, 2012). Then, one stops the quantification of item non-invariance at the level of loadings (i.e., $MIVI\text{-}L$)—although it might be helpful to obtain a rough estimate of intercept or uniqueness non-invariance too (i.e., $MIVI\text{-}I$ or $MIVI\text{-}U$) in some cases. Because the expected mean of unique scores is zero (see Equation 3),

scalar invariance is not a prerequisite for testing uniqueness invariance (Steinmetz, 2013; cf. DeShon, 2004; Wu et al., 2007).

Second, one cannot compare the three types of MIVI (i.e., $MIVI-L$, $MIVI-I$, $MIVI-U$) to each other and conclude, for instance, that $MIVI-L_{Item\ j|absolute}$ is larger than $MIVI-I_{Item\ j|absolute}$. Loadings, intercepts, and uniquenesses are entirely different parameters, and so are the associated effect sizes. However (and more importantly), the different types of MIVI can be compared across items of an item set. Thus, one can conclude, for instance, that $MIVI-L_{Item\ j|absolute}$ is larger than $MIVI-L_{Item\ j+1|absolute}$.

Signed Effect Size Measures

Violations of measurement invariance in factor loadings, intercepts, or uniquenesses across groups manifest as interaction effects of the respective parameter and group membership (e.g., Bauer, 2017), quantified through MIVIs. Such an interaction term can favor one or the other group (i.e., group 1 or group 2 has a larger non-invariant parameter). So far, $MIVI_{Item\ j|absolute}$ always has a positive sign; it cannot disentangle settings where group 1 has a larger parameter than group 2 from those where group 2 has a larger parameter than group 1. Without loss of generality, we can define a signed variant, $MIVI_{Item\ j|signed}$, that traces the mathematical signs of the non-invariant loading, intercept, or unique standard deviation differences back to the groups,

$$MIVI-L_{Item\ j|signed} = \frac{\lambda_{2j} - \lambda_{1j}}{SD_{LV,\ pooled}}, \quad (9a)$$

$$MIVI-I_{Item\ j|signed} = \frac{\tau_{2j} - \tau_{1j}}{SD_{LV,\ pooled}}, \quad (9b)$$

$$MIVI-U_{Item\ j|signed} = \frac{\sqrt{\theta_{2j}} - \sqrt{\theta_{1j}}}{SD_{LV,\ pooled}}. \quad (9c)$$

Different from $MIVI_{Item\ j|absolute}$ estimates, $MIVI_{Item\ j|signed}$ estimates are unbounded.

Item-level MIVIs allow quantifying the size of parameter differences for a single item. Knowing the size of loading, intercept, or uniqueness differences for a single item is highly informative in the scale development process when researchers face the question of retaining or discarding items in an item set. However, in applications with fixed item sets, one usually wants to know whether non-invariance substantially impacts group comparisons, especially when the goal is to compare observed summary statistics. Next, we present MIVI variants that consider multiple items in tandem and can answer this question.

Effect Size Measures for Item Sets

Absolute Effect Size Measures

Let us extend the logic of MIVIs for single items to the aggregate level, that is, calculating average MIVIs by considering *all* items of an item set simultaneously. Evaluating the bias by average MIVIs across all items of an item set is useful when the goal is to compare summative indices such as observed mean scores or when the goal is to use MIVIs as a rough, overall indicator of non-invariance bias at the level of item sets. As with item-level MIVIs, we define absolute and signed versions.

Suppose we have a partial metric invariance model if we want to quantify loading differences, a partial scalar invariance model if we want to quantify intercept differences, and a partial uniqueness invariance model if we want to quantify differences in unique standard deviations. For absolute MIVIs at the level of item sets, numerators summarize all biases of single items; once again, they do so in terms of absolute parameter differences (i.e., invariant parameters do not contribute to aggregate MIVIs). Again, denominators contain the pooled standard deviation of the latent variable. Dividing the fraction by the number of items serves as the average amount of bias introduced per item:

$$MIVI\text{-}L_{Item\ set|absolute} = \frac{|\lambda_{21} - \lambda_{11}| + |\lambda_{22} - \lambda_{12}| + \ldots + |\lambda_{2p} - \lambda_{1p}|}{SD_{LV,\ pooled}}/p = \frac{\sum_{j=1}^{p} |\lambda_{2j} - \lambda_{1j}|}{SD_{LV,\ pooled}}/p, \quad (10a)$$

$$MIVI\text{-}I_{Item\ set|absolute} = \frac{|\tau_{21} - \tau_{11}| + |\tau_{22} - \tau_{12}| + \ldots + |\tau_{2p} - \tau_{1p}|}{SD_{LV,\ pooled}}/p = \frac{\sum_{j=1}^{p} |\tau_{2j} - \tau_{1j}|}{SD_{LV,\ pooled}}/p, \quad (10b)$$

$$MIVI\text{-}U_{Item\ set|absolute} = \frac{|\sqrt{\theta_{21}} - \sqrt{\theta_{11}}| + |\sqrt{\theta_{22}} - \sqrt{\theta_{12}}| + \ldots + |\sqrt{\theta_{2p}} - \sqrt{\theta_{1p}}|}{SD_{LV,\ pooled}}/p = \frac{\sum_{j=1}^{p} |\sqrt{\theta_{2j}} - \sqrt{\theta_{1j}}|}{SD_{LV,\ pooled}}/p. \quad (10c)$$

$MIVI_{Item\ set|absolute}$ indicates the average between-group difference in factor loadings, intercepts, or unique standard deviations for a set of items in units of the pooled standard deviation of the latent variable. $MIVI_{Item\ set|absolute}$ quantifies the average amount of bias that arises from all cross-group differences in factor loadings, intercepts, or unique standard deviations in the latent representation of the item set, and it does so in the same metric as all MIVI versions.

Given the absolute values of constituents in the numerator, $MIVI_{Item\ set|absolute}$ is restricted to values ≥ 0. $MIVI_{Item\ set|absolute}$ gives an overall, absolute impression and quantifies *how biased* an item set is on average. Crucially, by dividing the fraction by the number of items, $MIVI_{Item\ set|absolute}$ takes the length of an item set into account. Thereby, it

considers that, for instance, two non-invariant items are more problematic in an item set of four rather than eight items.

From an applied perspective, bias due to non-invariant parameters may cancel out (at least in part). Thus, despite biased items in an item set, the overall between-group differences in the total variance or total mean score of the item set may remain unbiased (for a formula-based definition, see Nye & Drasgow, 2011). This is reminiscent of Chen (2007; see also Robitzsch & Lüdtke, 2020, for the context of item response theory), who distinguished between two patterns of non-invariance: uniform (i.e., one group has higher parameter values, such as intercepts, on all non-invariant items) and mixed (i.e., one group has higher parameter values on some non-invariant items but lower parameter values on other non-invariant items).

Crucially, any such non-compensatory aggregation or compensatory cancelation effects may be evaluated for intercepts (Equation 3) or unique variances (Equation 2). Factor loadings work as amplifiers; they result in multiplicative effects (Equations 2 and 3). They multiply with the latent variances/means when producing observed variances/means of the items (or item sets). Thus, factor loadings ultimately connect to the variances and means of the latent variable, which can differ across groups in partial invariance models. Compensation across groups is, thus, not immediately conceivable for loading differences.

Signed Effect Size Measures

A variant of $MIVI_{Item\ set|absolute}$ with each summand keeping its direction (i.e., sign), $MIVI_{Item\ set|signed}$, allows for complete (or incomplete) compensation of parameter differences of intercepts and unique standard deviations,

$$MIVI\text{-}L_{Item\ set|signed} = \textit{not applicable}, \tag{11a}$$

$$MIVI\text{-}I_{Item\ set|signed} = \frac{(\tau_{21} - \tau_{11}) + (\tau_{22} - \tau_{12}) + \ldots + (\tau_{2p} - \tau_{1p})}{SD_{LV,\ pooled}} / p = \frac{\sum_{j=1}^{p} (\tau_{2j} - \tau_{1j})}{SD_{LV,\ pooled}} / p, \tag{11b}$$

$$MIVI\text{-}U_{Item\ set|signed} = \frac{\sqrt{\theta_{21}} - \sqrt{\theta_{11}} + \sqrt{\theta_{22}} - \sqrt{\theta_{12}} + \ldots + \sqrt{\theta_{2p}} - \sqrt{\theta_{1p}}}{SD_{LV,\ pooled}} / p = \frac{\sum_{j=1}^{p} \sqrt{\theta_{2j}} - \sqrt{\theta_{1j}}}{SD_{LV,\ pooled}} / p. \tag{11c}$$

Given signed differences as constituents in the numerator, estimates of $MIVI_{Item\ set|signed}$ are unbounded.

Although MIVIs might show that non-invariance bias is negligible or cancels out completely, one should always be aware that the invariance assumption is violated to a certain degree in the item set. MIVIs are no legitimation to ignore non-invariance bias but rather a tool to assess its size and impact, depending on the research focus.

How Much of a Non-Invariant Parameter Difference is Too Large?

We have not yet outlined which parameter differences in units of the pooled latent standard deviation can be considered either negligible or substantial, although some preliminary guidance for applied settings may be helpful. Our goal is not to prescribe a single fixed rule. Many components, such as the substantive research question, type of analysis, empirical context, and, especially, the parameter of interest (i.e., factor loading, intercept, or uniqueness), may influence what is considered a critical value (e.g., Steinmetz, 2013).

Although Cohen's (1988, 1992) effect size guideline has recently come under scrutiny for not being applicable in all research contexts (e.g., Schäfer & Schwarz, 2019), researchers are well-acquainted with his cutoffs and their meanings. For his *d*, a difference measure between independent means relative to the pooled standard deviation, he suggested that $|d| = 0.2$ may indicate a substantial but small, $|d| = 0.5$ a medium, and $|d| = 0.8$ a large effect. These values may serve as initial thresholds (especially for $MIVI - I$ quantifying intercept differences) that, when surpassed, alert researchers to become cautious about biased group comparisons in their substantive analyses. We reiterate that this guideline is extremely rough; an appropriate guideline must be tailored to substantive considerations, different parameters, and the specific research context of interest.

Applications

Example Setup

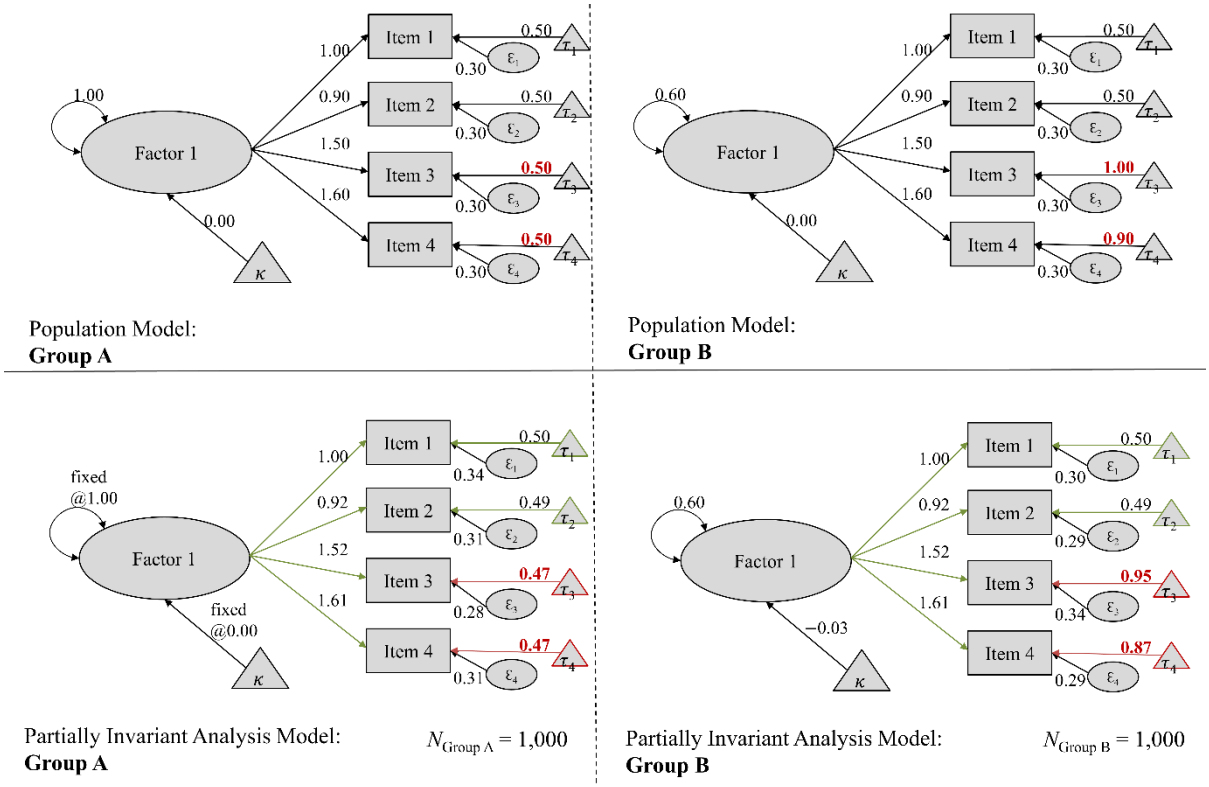
We simulated data to exemplify the applicability of the MIVI versions. The advantage of simulations is that we can specify beforehand all model, data, and estimation characteristics, such as the magnitude of the model parameters and, accordingly, the magnitude of bias due to non-invariance. We conducted all analyses with Mplus (version 8.4; Muthén & Muthén, 1998-2017). We estimated bootstrap confidence intervals around the point estimate of all MIVIs. We supplied the Mplus code for the Simulated Examples 1 and 2 in the Supplementary Materials (see Additional Files 2 and 3). Additional File 4 includes R code for a MIVI application similar to Example 1 (i.e., using the same population model for data generation). Some intercepts were non-invariant in both examples; all factor loadings were invariant. Thus, full metric and partial scalar invariance held. In empirical reality, this is not always the case. As we also derived MIVIs to quantify loading non-invariance, MIVIs can also be applied when only partial metric invariance holds. Further, MIVIs can also be applied when partial uniqueness invariance holds,

as we also derived MIVIs for uniqueness invariance. The logic of applying MIVIs in different empirical settings is analogous to the exemplary one.

Simulated Example 1

To demonstrate the utility of the MIVI versions, we simulated data for a hypothetical item set that conforms to a partial scalar invariance model across two groups. We simulated data for four continuous indicator variables underlying a single latent variable. We set the number of observations to 1,000 per group. The intercepts of Items 1 and 2 in Group A equaled those in Group B. The intercepts of Items 3 and 4 differed across groups. We fit the partial scalar invariance model with the maximum likelihood (ML) estimator. Figure 1 shows the predetermined coefficients from the population model. Further, it shows the empirical coefficients resulting from the partial scalar analysis model. Apart from sampling variation, the empirical coefficients mirrored the predetermined ones. We identified the partial scalar analysis model by fixing the latent variance in Group A to 1 and the latent mean in Group A to 0. The latent variance and latent mean in Group B are freely estimated. Table 2 includes the resulting estimates of the MIVI versions (item- or item-set-level × absolute or signed) and accompanying bootstrap confidence intervals.

***Figure 1**: Population and Analysis Models (Example 1)*



Note. Non-invariant parameters are colored in red, invariant parameters in green.

Table 2: *MIVIs for Item and Item Set Levels (Example 1)*

		Absolute	Signed
Item level	Item 3	$MIVI\text{-}I_{Item\ 3\|absolute}$ (Eq. 8b) $= \frac{\|0.95 - 0.47\|}{\sqrt{(0.60 + 1.00) / 2}}$ = 0.54 $CI_{95\%}$[0.45; 0.63]	$MIVI\text{-}I_{Item\ 3\|signed}$ (Eq. 9b) $= \frac{0.95 - 0.47}{\sqrt{(0.60 + 1.00) / 2}}$ = 0.54 $CI_{95\%}$[0.45; 0.63]
	Item 4	$MIVI\text{-}I_{Item\ 4\|absolute}$ (Eq. 8b) $= \frac{\|0.87 - 0.47\|}{\sqrt{(0.60 + 1.00) / 2}}$ = 0.46 $CI_{95\%}$[0.37; 0.54]	$MIVI\text{-}I_{Item\ 4\|signed}$ (Eq. 9b) $= \frac{0.87 - 0.47}{\sqrt{(0.60 + 1.00) / 2}}$ = 0.46 $CI_{95\%}$[0.37; 0.54]
Item set level		$MIVI\text{-}I_{Item\ set\|absolute}$ (Eq. 10b) $= \frac{\|0.95 - 0.47\| + \|0.87 - 0.47\|}{\sqrt{(0.60 + 1.00) / 2}} / 4$ = 0.25 $CI_{95\%}$[0.21; 0.29]	$MIVI\text{-}I_{Item\ set\|signed}$ (Eq. 11b) $= \frac{0.95 - 0.47 + 0.87 - 0.47}{\sqrt{(0.60 + 1.00) / 2}} / 4$ = 0.25 $CI_{95\%}$[0.21; 0.29]

We first inspected the (relative) size of the bias due to non-invariant intercepts by looking at the absolute MIVIs on the item level. As intended, bias was more substantial in Item 3 than in Item 4. Whereas Item 3 had a non-invariant intercept difference of 0.54 pooled standard deviations of the latent variable, Item 4 had a non-invariant intercept difference of 0.46 pooled standard deviations. Both corresponded to a substantial medium effect, following Cohen's guidelines. The signed MIVIs on the item level were equal to the absolute ones, indicating that intercept bias was larger in Group B than in Group A for both items.

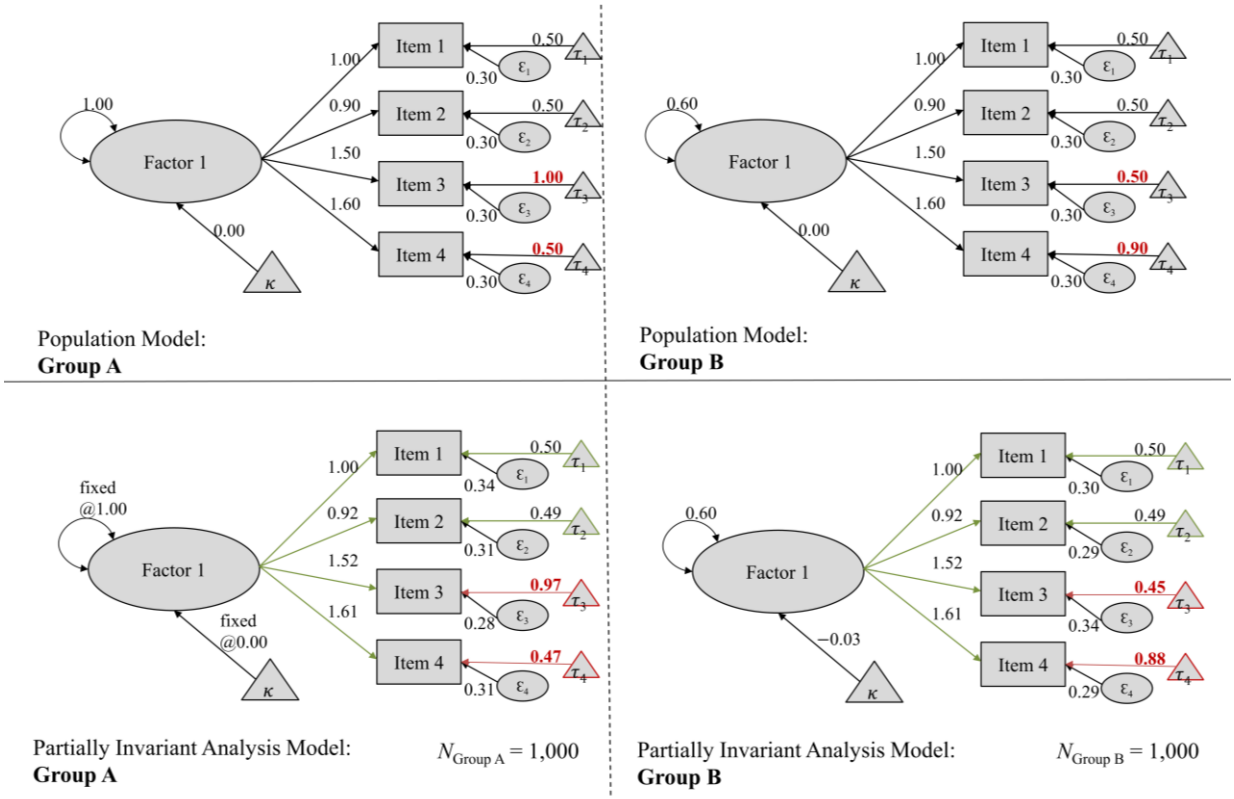
As a next step, we took a closer look at the size and shape of the total non-invariance bias of the complete item set. The absolute bias due to all non-invariant intercept differences adding up in the item set was non-negligible according to $MIVI\text{-}I_{Item\ set|absolute}$: The average intercept difference was as large as 0.25 pooled standard deviations per item of the item set. The signed version of $MIVI\text{-}I_{Item\ set}$ equaled the absolute one and, thus, showed that intercept differences did not cancel out but accumulated at the level of the item set.

Considering all MIVI values, we would not consider the bias due to non-invariant intercepts negligible. Especially the absolute and signed versions of $MIVI\text{-}I_{Item\ set}$ clearly demonstrated that non-invariance bias added up and substantially contaminated the item set in total (and, accordingly, simplified statistics of the item set such as observed mean scores).

Simulated Example 2

The following example illustrates the applicability of MIVIs in scenarios with compensatory non-invariance bias. We simulated data in the same way as in Example 1. The crucial difference was the opposite direction of intercept differences across Items 3 and 4. Figure 2 shows the predetermined coefficients from the population model and the estimated coefficients from the partial scalar analysis model. Table 3 includes the resulting estimates of the MIVI versions and accompanying bootstrap confidence intervals.

Figure 2: *Population and Analysis Models (Example 2)*



Note. Non-invariant parameters are colored in red, invariant parameters in green.

Table 3: *MIVIs for Item and Item Set Levels (Example 2)*

		Absolute	Signed
Item level	Item 3	$MIVI\text{-}I_{Item\ 3\|absolute}$ (Eq. 8b) $= \frac{\|0.45 - 0.97\|}{\sqrt{(0.60 + 1.00)\ /\ 2}}$ $= 0.58$ $CI_{95\%}[0.49; 0.67]$	$MIVI\text{-}I_{Item\ 3\|signed}$ (Eq. 9b) $= \frac{0.45 - 0.97}{\sqrt{(0.60 + 1.00)\ /\ 2}}$ $= -0.58$ $CI_{95\%}[-0.67; -0.49]$
	Item 4	$MIVI\text{-}I_{Item\ 4\|absolute}$ (Eq. 8b) $= \frac{\|0.88 - 0.47\|}{\sqrt{(0.60 + 1.00)\ /\ 2}}$ $= 0.46$ $CI_{95\%}[0.37; 0.54]$	$MIVI\text{-}I_{Item\ 4\|signed}$ (Eq. 9b) $= \frac{0.88 - 0.47}{\sqrt{(0.60 + 1.00)\ /\ 2}}$ $= 0.46$ $CI_{95\%}[0.37; 0.54]$
Item set level		$MIVI\text{-}I_{Item\ set\|absolute}$ (Eq. 10b) $= \frac{\|0.45 - 0.97\| + \|0.88 - 0.47\|}{\sqrt{(0.60 + 1.00)\ /\ 2}} / 4$ $= 0.26$ $CI_{95\%}[0.24; 0.28]$	$MIVI\text{-}I_{Item\ set\|signed}$ (Eq. 11b) $= \frac{0.45 - 0.97 + 0.88 - 0.47}{\sqrt{(0.60 + 1.00)\ /\ 2}} / 4$ $= -0.03$ $CI_{95\%}[-0.07; 0.01]$

Again, the absolute non-invariance bias of each item was relatively substantial, with the intercept difference of Item 3 ($MIVI\text{–}I_{Item\ 3|absolute}$ = 0.58) being larger than that of Item 4 ($MIVI\text{–}I_{Item\ 4|absolute}$ = 0.46). However, $MIVI_{Item\ 3|signed}$ had a negative sign for the intercept non-invariance bias ($MIVI\text{–}I_{Item\ 3|signed} = -0.58$), whereas $MIVI_{Item\ 4|signed}$ had a positive sign ($MIVI\text{–}I_{Item\ 4|absolute}$ = 0.46). The non-invariance bias of Items 3 and 4 operated differently: Item 3 had a larger intercept in Group B, Item 4 had a larger intercept in Group A.

To further investigate the mutual compensation of intercept differences, we considered the average non-invariance bias of the whole item set. The absolute biasing effect was non-negligible for intercepts ($MIVI\text{–}I_{Item\ set|absolute}$) and resulted in an average difference of 0.26 pooled standard deviations per item. Given this non-negligible effect size at first inspection, one might refrain from treating the item set as invariant. However, calculating $MIVI_{Item\ set|signed}$ provides further information on possible compensatory differences (e.g., non-invariant intercepts canceling each other out), which are relevant for valid cross-group comparisons of total mean scores (in the case of non-invariant intercepts). $MIVI_{Item\ set|signed}$ showed that the two intercept differences accounted for an average non-invariance bias per item of |0.03| latent standard deviations ($MIVI\text{–}I_{Item\ set|signed} = -0.03$). The confidence interval was compatible even with the notion of the total absence of any biasing effect ($CI_{95\%}$ = [−0.07; 0.01]). The non-negligible amount of (absolute) bias present in all items was negligible when considering the direction of item-specific bias.

The relatively small values of the signed MIVI at the item set level suggested that non-invariant intercept differences compensated for each other. Some researchers may conclude that the item set is practically invariant in this example to justify the comparison of observed group statistics, especially if one favors treating the items as a fixed item set without the possibility to further select items before running substantive analyses. In different words, applied researchers might conclude from the analysis of all MIVIs (and especially $MIVI_{Item\ set|signed}$) that the non-invariant intercepts did not substantially impact the difference in observed mean scores of the item set.

Discussion and Conclusion

The Usefulness of Measurement Invariance Violation Indices

Research questions involving numerical between-group comparisons of constructs are ubiquitous in psychological research and beyond. However, for group comparisons to be valid, researchers must first establish the measurement invariance of the constructs in question. Unfortunately, as any experienced analyst can testify, measurement non-invariance is a frequently encountered problem. The question then becomes: How wrong are assumptions of invariance? Stated differently: How much non-invariance bias is present?

This paper provided a novel set of tools to answer this essential question. Specifically, we proposed an effect size approach called MIVI that quantifies the non-invariance of factor loadings, intercepts, and uniquenesses in a unified and principled way. MIVIs are handy effect size measures that overcome the limitations of existing approaches. Specifically, MIVIs provide a common metric for different parameters (i.e., factor loadings, intercepts, and uniquenesses) that is comparable across all items of an item set. At the item level, MIVIs relate each item's cross-group difference in factor loadings, intercepts, or uniquenesses to the pooled standard deviation of the latent variable. At the item set level, MIVIs reflect the average cross-group difference in factor loadings, intercepts, or uniquenesses in units of the pooled latent standard deviation across all items. Deriving absolute and directional MIVI versions, we can inspect the cancelation of non-invariance bias. Confidence intervals display the accuracy of the MIVI estimates. MIVI estimates and confidence intervals can easily be computed by adding a few code lines for standard SEM software such as Mplus or R.

MIVIs are especially helpful for scale development as well as in settings with fixed item sets: In scale development, MIVIs can support the decision on keeping non-invariant items or dropping them from the item set—depending on the size of the specific parameter non-invariance MIVIs identify on the level of single items and item sets. By quantifying the non-invariance bias in a fixed item set, MIVIs can help evaluate the quality of a questionnaire, for instance, in a new context. Further, MIVIs allow evaluation of the amount of bias in cross-group comparisons of simple proxies such as observed mean scores.

Preconditions of Measurement Invariance Violation Indices

Without loss of generality, we have sketched MIVIs as useful for quantifying non-invariance bias when comparing two groups. Researchers can legitimately apply MIVI variants to scenarios with more than two groups: They may pick a reference group and compare all non-invariant parameters against this reference group, yielding multiple reference-group comparisons.

MIVIs also rest on other assumptions or preconditions that must be fulfilled. The analysis model should be unidimensional (without cross-loadings) to employ MIVIs correctly. If analysis models are multidimensional (assuming no cross-loadings), separate MIVIs should be calculated for each dimension. Alternatively, one might estimate how proximate the multidimensional model is to a unidimensional one, indicating whether the multidimensional model can be treated as essentially unidimensional (Raykov & Bluemke, 2020).

Further, the multi-group measurement model, which is the basis for all MIVI versions, must be correctly identified. Multi-group measurement models are commonly identified by constraining the latent variance or a factor loading to a fixed value (commonly to 1). Additionally, to identify the mean structure, either the latent mean or an intercept must be constrained to a fixed value (commonly to 0). The crucial assumption for measurement invariance tests to be valid is that those fixed parameters are equal (or invariant) across groups (Putnick & Bornstein, 2016). If the fixed parameters are not invariant, measurement invariance tests might be misleading. Whereas one can only assume that latent variances and latent means are equal across groups, several approaches guide identifying so-called reference indicators or anchor items (e.g., Kopf et al., 2015, Schulze & Pohl, 2021; Thompson et al., 2021). Raykov et al.'s (2013) approach to measurement invariance, which we outlined in Additional File 1 of the Supplementary Information, does not require identifying a reference item (as it constrains multiple items at once). It is specifically suitable for detecting non-invariant intercepts (Thompson et al., 2021).

To properly quantify bias due to non-invariant parameters (i.e., factor loadings, intercepts, or uniquenesses) and, thus, to estimate MIVIs, one needs a common, congeneric factor model (i.e., configural invariance) as the starting point. Otherwise, one cannot compare parameters at all. Furthermore, partial metric invariance must hold. Otherwise, one cannot compare and pool latent variances for the computation of the MIVI denominators.

For MIVIs to be interpretable for intercepts and unique variances, the non-invariant items should also have equal loadings. If the units (i.e., loadings) differ across groups, unit

shifts (i.e., intercepts) or variances of unique scores are barely comparable (theoretically and, often, empirically; e.g., Millsap & Olivera-Aguilar, 2012). Then, MIVIs should only be estimated for loadings—although a rough estimate of intercept non-invariance (i.e., $MIVI\text{–}I$) or uniqueness non-invariance (i.e., $MIVI\text{–}U$) might be helpful in some cases.

Further, MIVIs rely on raw parameters (i.e., factor loadings, intercepts, and uniquenesses), partially standardized by the pooled standard deviation of the latent variable. Thus, the method of model identification influences the MIVI estimates. Either constraining the latent variance or a factor loading to a fixed value, commonly to 1, identifies the (co)variance structure of the model. Either the latent mean or an intercept constrained to a fixed value, commonly to 0, identifies the mean structure of the model. The way to identify the model influences the estimates of model parameters (such as cross-group estimates of latent variances, factor loadings, or intercepts) and, in turn, MIVI estimates. To have a common ground for MIVIs, we propose identifying the final partial invariance model by fixing the latent variance to 1 and the latent mean to 0. As MIVIs are estimated from partial invariance models, the latent variance (and the latent mean if at least partial scalar invariance holds) must only be fixed in a reference group. Still, it can be freely estimated in the other group(s).

Limitations and Future Directions

Although MIVIs might show that non-invariance is non-substantial or cancels out at the level of the item set, one should always be aware that non-invariance is present. One should not use MIVI as a legitimation to completely ignore non-invariance but rather as a tool to quantify non-invariance or assess its impact on simple statistics (e.g., observed mean scores). Theoretical considerations and empirical models, such as multiple-indicators multiple-causes models (or even more flexible variations of it, e.g., Bauer, 2017), can help explain the non-invariance and should still be reflected in addition to applying MIVIs.

Specific heuristics help evaluate whether an effect is small, medium, or large (e.g., Cohen, 1988, 1992). In this paper, we tentatively recommended Cohen's effect size guidelines as initial thresholds to evaluate MIVI values. However, Cohen's guidelines are no more than starting points. One needs to understand what constitutes small, medium, or large effects in the specific setting of interest. Unless enough empirical evidence is gathered to understand how non-invariant parameter differences distribute in empirical data (e.g., Gignac & Szodorai, 2016), simulations may aid in evaluating the impact of non-invariant parameter differences on summary statistics (e.g., Nye et al., 2019).

Conclusion

MIVIs are tools for quantifying measurement non-invariance in an easy-to-apply and continuous manner. They are available by adding only a few lines of code in standard statistical programs. MIVIs quantify non-invariance bias from partial metric, scalar, or uniqueness invariance models. They help overcome the practice of merely accepting or rejecting an invariance model in a binary accept-or-reject logic. Their advantages are manifold: MIVIs can support the group-fair item selection by quantifying the item non-invariance during the scale development process. In settings with fixed item sets, MIVIs can help evaluate the quality of a questionnaire in a new context or evaluate the amount of non-invariance bias in cross-group comparisons of simple index scores. Taken together, MIVIs advance researchers' focus and alert them to the practical consequences of measurement non-invariance.

References

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

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology*, *49*(5), 713–734. https://doi.org/10.1177/0022022117749042

Bollen, K. A. (1989). *Structural equations with latent variables.* Wiley.

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

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*(5), 1005–1018. https://doi.org/10.1037/a0013193

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*(4), 532–574. https://doi.org/10.1177/0013164401614002

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., & Meuleman, B. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology, 43*(4), 558–575. https://doi.org/10.1177/0022022112438397

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*(1), 55–75. https://doi.org/10.1146/annurev-soc-071913-043137

DeShon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science, 46*(1), 137–149.

Dong, Y., & Dumas, D. (2020). Are personality measures valid for different populations? A systematic review of measurement invariance across cultures, gender, and age. *Personality and Individual Differences, 160*, Article 109956. https://doi.org/10.1016/j.paid.2020.109956

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Gunn, H. J., Grimm, K. J., & Edwards, M. C. (2020). Evaluation of six effect sizes of measurement non-invariance for continuous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(4), 503–514. https://doi.org/10.1080/10705511.2019.1689507

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

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22–56. https://doi.org/10.1177/0013164414529792

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

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. https://doi.org/10.1007/BF02294825

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

Millsap, R. E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. Hoyle, D. Kaplan, G. A. Marcoulides, & S. West (Eds.), *Handbook of Structural Equation Modeling* (pp. 380–392). Guilford Press.

Muthén, B.O. (1998-2014). *Mplus technical appendices*. Muthén & Muthén.

Muthén, L.K., & Muthén, B.O. (1998-2017). *Mplus User's Guide* (8th ed.). Muthén & Muthén.

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology, 96*(5), 966–980. https://doi.org/10.1037/a0022955

Nye, C. D., Bradburn, J., Olenick, J., Bialko, C., & Drasgow, F. (2019). How big are my effects? Examining the magnitude of effect sizes in studies of measurement equivalence. *Organizational Research Methods, 22*(3), 678–709. https://doi.org/10.1177/1094428118761122

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis, 22*(1), 45–60. https://doi.org/10.1093/pan/mpt014

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Sage Publications.

Pornprasertmanit, S. (2022). *A note on effect size for measurement invariance.* https://cran.r-project.org/web/packages/semTools/vignettes/partialInvariance.pdf

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

R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Raykov, T., & Bluemke, M. (2020). Examining multidimensional measuring instruments for proximity to unidimensional structure using latent variable modeling. *Educational and Psychological Measurement, 81*(2), 319–339. https://doi.org/10.1177/0013164420940764

Raykov, T., Marcoulides, G. A., & Millsap, R. E. (2013). Factorial invariance in multiple populations: A multiple testing procedure. *Educational and Psychological Measurement, 73*(4), 713–727. https://doi.org/10.1177/0013164412451978

Robitzsch, A., & Lüdtke, O. (2020). A review of different scaling approaches under full invariance, partial invariance, and noninvariance for cross-sectional country comparisons in large-scale assessments. *Psychological Test and Assessment Modeling, 62*(2), 233–279.

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

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology, 10,* Article 813. https://doi.org/10.3389/fpsyg.2019.00813

Schulze, D., & Pohl, S. (2021). Finding clusters of measurement invariant items for continuous covariates. *Structural Equation Modeling, 28*(2), 219–228. https://doi.org/10.1080/10705511.2020.1771186

Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78–90. https://doi.org/10.1086/209528

Steinmetz, H. (2013). Analyzing observed composite differences across groups: Is partial measurement invariance enough? *Methodology, 9*(1), 1–12. https://doi.org/10.1027/1614-2241/a000049

Thompson, Y. T., Song, H., Shi, D., & Liu, Z. (2021). It matters: Reference indicator selection in measurement invariance tests. *Educational and Psychological Measurement, 81*(1), 5–38. https://doi.org/10.1177/0013164420926565

Widaman, K.F., & Revelle, W. (2022). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-022-01849-w

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research, and Evaluation, 12*, Article 3.

Zercher, F., Schmidt, P., Cieciuch, J., & Davidov, E. (2015). The comparability of the universalism value over time and across countries in the European Social Survey: Exact vs. approximate measurement invariance. *Frontiers in Psychology, 6*, Article 733. https://doi.org/10.3389/fpsyg.2015.00733

Additional File 1: How to Test for Measurement Invariance?

Researchers often test measurement invariance in a bottom-up manner, though a preferable top-down approach exists. In the standard bottom-up procedure of measurement invariance testing, researchers test three levels of invariance: configural, metric, and scalar invariance (e.g., Meredith, 1993; Millsap, 2011; Steenkamp & Baumgartner, 1998). The levels are nested such that metric invariance cannot exist without configural invariance, and scalar invariance cannot (meaningfully) exist without metric invariance. Thus, after testing (and possibly accepting) the invariance of the basic item-factor structure without any parameter constraints (configural invariance) across the $k = 1, 2, \ldots, K$ tested groups, one proceeds to test the invariance of factor loadings (metric invariance; $\boldsymbol{\lambda}_1 = \ldots = \boldsymbol{\lambda}_K$). Only after accepting metric invariance, one can proceed to test the invariance of intercepts (scalar invariance; $\boldsymbol{\tau}_1 = \ldots = \boldsymbol{\tau}_K$). Depending on the research question, researchers may finally test the invariance of uniqueness terms (residual invariance, $\boldsymbol{\Theta}_1 = \ldots = \boldsymbol{\Theta}_K$).

Metric invariance implies identical scaling and unbiased comparisons of latent (co)variances across groups. Scalar invariance implies equivalent locations of measurement indicators and unbiased comparisons of mean structures across groups. Uniqueness invariance implies that the observed error variances impact each group alike. The invariance levels obtained from sequential tests of metric, scalar, and uniqueness invariance are also known as weak, strong, and strict invariance tests. In short, measurement invariance holds if the K groups' parameters are identical for the $k = 1, 2, \ldots, K$ tested groups; hence k might be dropped as an index from the formulae without losing information.

Whether a specific level of measurement invariance is accepted may be decided upon inspection of the (significance of the) chi-square test statistic (Bollen, 1989) or the difference in chi-square test statistics when comparing a more restrictive model to a less restrictive model with the former being nested in the latter. Additionally, or alternatively, researchers rely on information criteria (such as the Akaike information criterion, AIC, and the Bayesian information criterion, BIC) or fit indices, such as the comparative fit index (CFI), the root mean square error of approximation (RMSEA), or the standardized root mean squared residual (SRMR). Fit indices are commonly compared against cutoffs, such as those suggested by Hu and Bentler (1999) for overall model fit and those suggested by Chen (2007) for relative model fit. If values exceed those critical thresholds, the stricter model is rejected, resulting in a binary logic for accepting or rejecting the invariance assumption.

As an alternative to the classic bottom-up testing procedure, Raykov et al. (2013) developed a top-down procedure for measurement invariance testing. The approach proceeds in seven major steps: (1) The researcher starts on a fully constrained invariance model, Model *M0*. (2) Irrespective of obtaining sufficient model fit for *M0*, each constrained item parameter is released individually (and fixed afterward), while all other parameters remain equal across groups. Each inspected model M_s, for s = 1, 2, …, S tested parameters, differs only by one degree of freedom from the initial model *M0*. (3) For each tested model, M_s, a *p*-value results that correspond to the chi-square difference test between Model M_s and Model *M0*. (4) Ranking the *p*-values in ascending order first, (5) one next addresses multiple testing and estimates so-called *l*-values that correspond to the order of *p*-values, in line with the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995). The first *l*-value (for the model with the lowest *p*-value) is obtained by dividing the *p*-value cutoff, commonly controlling the Type I error rate at α = .05, through a series of m (= number of models) ratios times m,

$$l_1 = \frac{\alpha}{m*(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{m})}. \quad (1)$$

The second *l*-value is obtained by multiplying l_1 with 2, followed by 3 until m. After estimating the *l*-values, (6) the largest *p*-value, which is below its corresponding *l*-value, *p**, is identified. (7) All coefficients with *p*-values below *p** are flagged as non-invariant. See Raykov et al. (2013) for more details on the procedure.

Additional File 2: Mplus Code for the Simulated Example 1

Data Simulation (Mplus Input File)

```
TITLE: SIMULATING DATA IN GROUP 1 AND GROUP 2 FOR THE SIMULATED EXAMPLE 1;

MONTECARLO:
NAMES = y1-y4; ! FOUR ITEMS
NOBSERVATIONS = 1000 1000; ! 1000 OBSERVATIONS PER GROUP
NGROUPS = 2 ! TWO GROUPS
NREPS = 1;
SEED = 53487; ! SET SEED FOR REPLICATION
SAVE = DAT_G12_EX1.DAT;

MODEL POPULATION: ! POPULATION MODEL FROM WHICH DATA IN GROUP 1 IS GENERATED
f1 BY y1*1 y2*0.9 y3*1.5 y4*1.6; ! LOADINGS RANGING FROM 0.9 TO 1.6
f1@1; ! LATENT VARIANCE FIXED TO 1
[f1@0]; ! LATENT MEAN FIXED TO 0
y1-y4*0.3; ! RESIDUALS ARE 0.3 IN THE POPULATION
[y1-y4*0.5]; ! INTERCEPTS ARE ALL 0.5 IN THE POPULATION IN GROUP 1

MODEL POPULATION-g2: ! POPULATION MODEL FROM WHICH DATA IN GROUP 2 IS GENERATED
f1 BY y1*1 y2*0.9 y3*1.5 y4*1.5; ! LOADINGS RANGING FROM 0.9 TO 1.6
f1@0.6; ! LATENT VARIANCE FIXED TO 0.6
[f1@0]; ! LATENT MEAN FIXED TO 0
y1-y4*0.3; ! RESIDUALS ARE 0.3 IN THE POPULATION
[y1-y2*0.5]; ! INTERCEPTS ARE 0.5 FOR ITEMS 1-2 IN THE POPULATION
[y3*1.0]; ! INTERCEPT IS 1.0 FOR ITEM 3 IN GROUP 2 IN THE POPULATION
[y4*0.9]; ! INTERCEPT IS 0.9 FOR ITEM 4 IN GROUP 2 IN THE POPULATION

MODEL: ! ANALYSIS MODEL FIT TO THE DATA OF GROUP 1
       ! (EQUALS RESPECTIVE POPULATION MODEL)
f1 BY y1*1 y2*.9 y3*1.5 y4*1.6;
f1@1;
[f1@0];
y1-y4*0.3;
[y1-y4*0.5];

MODEL g2: ! ANALYSIS MODEL FIT TO THE DATA OF GROUP 2
          ! (EQUALS RESPECTIVE POPULATION MODEL)
f1 BY y1*1 y2*.9 y3*1.5 y4*1.6;
f1@0.6;
[f1@0];
y1-y2*.3;
[y1-y2*0.5];
[y3*1.0];
[y4*0.9];

OUTPUT: TECH9;
```

Data Analysis (Mplus Input File)

```
TITLE:      ESTIMATING PROPOSED MEASUREMENT INVARIANCE VIOLATION
            INDICES USING DATA SIMULATED WITH A CFA MODEL
            NON-INVARIANT INTERCEPTS FOR ITEMS Y3 AND Y4
            EXAMPLE 1;

DATA:       FILE = DAT_G12_EX1.DAT; !("FULL" DATA SET);

VARIABLE:   NAMES = Y1-Y4 GROUP;
            GROUPING = GROUP(1 = G1, 2 = G2);

ANALYSIS:   ESTIMATOR = ML;
            BOOTSTRAP = 1000;

MODEL:      f1 BY Y1* (L1)
                 Y2-Y4 (L2-L4); ! LABELED FACTOR LOADINGS;
            f1@1; ! FIXED LATENT VARIANCE IN GROUP 1;
            [f1@0]; ! FIXED LATENT MEAN IN GROUP 1;
            [y1-y4] (TAU11-TAU14);


            MODEL G2:
            f1 (LV_G2); ! FREE LATENT VARIANCE IN GROUP 2;
            [f1] (LM_G2); ! FREE LATENT MEAN IN GROUP 2;
            [y1-y2] (TAU11-TAU12); ! FIXED INTERCEPTS ACROSS GROUPS 1&2;
            [y3*] (TAU23); ! FREE INTERCEPT ACROSS GROUPS 1&2;
            [y4*] (TAU24); ! FREE INTERCEPT ACROSS GROUPS 1&2;

MODEL CONSTRAINT:
            NEW(M3_A, M3_S, M4_A, M4_S, M_A, M_S);

            ! M3_A = MIVI - ITEM 3, ABSOLUTE
              M3_A = SQRT((TAU23-TAU13)^2)/SQRT((LV_G2+1)/2);
            ! M3_S = MIVI - ITEM 3, SIGNED
              M3_S = (TAU23-TAU13)/SQRT((LV_G2+1)/2);
            ! M4_A = MIVI - ITEM 4, ABSOLUTE
              M4_A = SQRT((TAU24-TAU14)^2)/SQRT((LV_G2+1)/2);
            ! M4_S = MIVI - ITEM 4, SIGNED
              M4_S = (TAU24-TAU14)/SQRT((LV_G2+1)/2);
            ! M_A = MIVI - ITEM SET, ABSOLUTE
              M_A = ((SQRT((TAU24-TAU14)^2)+SQRT((TAU23-TAU13)^2))/
                    SQRT((LV_G2+1)/2))/4;
            ! MSCALE_S = MIVI - ITEM SET, SIGNED
              M_S = ((TAU24-TAU14+TAU23-TAU13)/
                    SQRT((LV_G2+1)/2))/4;

OUTPUT:     CINTERVAL(BOOTSTRAP) STDYX;
```

Additional File 3: Mplus Code for the Simulated Example 2

Data Simulation (Mplus Input File)

```
TITLE: SIMULATING DATA IN GROUP 1 AND GROUP 2 FOR THE SIMULATED EXAMPLE 2;

MONTECARLO:
NAMES = y1-y4; ! FOUR ITEMS
NOBSERVATIONS = 1000 1000; ! 1000 OBSERVATIONS PER GROUP
NGROUPS = 2 ! TWO GROUPS
NREPS = 1;
SEED = 53487; ! SET SEED FOR REPLICATION
SAVE = DAT_G12_EX2.DAT;

MODEL POPULATION: ! POPULATION MODEL FROM WHICH DATA IN GROUP 1 IS GENERATED
f1 BY y1*1 y2*0.9 y3*1.5 y4*1.6; ! LOADINGS RANGING FROM 0.9 TO 1.6
f1@1; ! LATENT VARIANCE FIXED TO 1
[f1@0]; ! LATENT MEAN FIXED TO 0
y1-y4*0.3; ! RESIDUALS ARE 0.3 IN THE POPULATION
[y1-y2*0.5]; ! INTERCEPTS ARE 0.5 FOR ITEMS 1-2 IN THE POPULATION
[y3*1.0]; ! INTERCEPT IS 1.0 FOR ITEM 3 IN GROUP 1 IN THE POPULATION
[y4*0.5]; ! INTERCEPT IS 0.5 FOR ITEM 4 IN GROUP 1 IN THE POPULATION

MODEL POPULATION-g2: ! POPULATION MODEL FROM WHICH DATA IN GROUP 2 IS GENERATED
f1 BY y1*1 y2*0.9 y3*1.5 y4*1.5; ! LOADINGS RANGING FROM 0.9 TO 1.6
f1@0.6; ! LATENT VARIANCE FIXED TO 0.6
[f1@0]; ! LATENT MEAN FIXED TO 0
y1-y4*0.3; ! RESIDUALS ARE 0.3 IN THE POPULATION
[y1-y2*0.5]; ! INTERCEPTS ARE 0.5 FOR ITEMS 1-2 IN THE POPULATION
[y3*0.5]; ! INTERCEPT IS 0.5 FOR ITEM 3 IN GROUP 2 IN THE POPULATION
[y4*0.9]; ! INTERCEPT IS 0.9 FOR ITEM 4 IN GROUP 2 IN THE POPULATION

MODEL: ! ANALYSIS MODEL FIT TO THE DATA OF GROUP 1
       ! (EQUALS RESPECTIVE POPULATION MODEL)
f1 BY y1*1 y2*.9 y3*1.5 y4*1.6;
f1@1;
[f1@0];
y1-y4*.3;
[y1-y2*0.5];
[y3*1.0];
[y4*0.5];

MODEL g2: ! ANALYSIS MODEL FIT TO THE DATA OF GROUP 2
          ! (EQUALS RESPECTIVE POPULATION MODEL)
f1 BY y1*1 y2*.9 y3*1.5 y4*1.6;
f1@0.6;
[f1@0];
y1-y2*.3;
[y1-y2*0.5];
[y3*0.5];
[y4*0.9];

OUTPUT: TECH9;
```

Data Analysis (Mplus Input File)

```
TITLE:      ESTIMATING PROPOSED MEASUREMENT INVARIANCE VIOLATION
            INDICES USING DATA SIMULATED WITH A CFA MODEL
            NON-INVARIANT INTERCEPTS FOR ITEMS Y3 AND Y4
            EXAMPLE 2;

DATA:       FILE = DAT_G12_EX2.DAT; !("FULL" DATA SET);

VARIABLE:   NAMES = Y1-Y4 GROUP;
            GROUPING = GROUP(1 = G1, 2 = G2);

ANALYSIS:   ESTIMATOR = ML;
            BOOTSTRAP = 1000;

MODEL:      f1 BY Y1* (L1)
                 Y2-Y4 (L2-L4); ! LABELED FACTOR LOADINGS;
            f1@1; ! FIXED LATENT VARIANCE IN GROUP 1;
            [f1@0]; ! FIXED LATENT MEAN IN GROUP 1;
            [y1-y4] (TAU11-TAU14);


            MODEL G2:
            f1 (LV_G2); ! FREE LATENT VARIANCE IN GROUP 2;
            [f1] (LM_G2); ! FREE LATENT MEAN IN GROUP 2;
            [y1-y2] (TAU11-TAU12); ! FIXED INTERCEPTS ACROSS GROUPS 1&2;
            [y3*] (TAU23); ! FREE INTERCEPT ACROSS GROUPS 1&2;
            [y4*] (TAU24); ! FREE INTERCEPT ACROSS GROUPS 1&2;

MODEL CONSTRAINT:
            NEW(M3_A, M3_S, M4_A, M4_S, M_A, M_S);

            ! M3_A = MIVI - ITEM 3, ABSOLUTE
              M3_A = SQRT((TAU23-TAU13)^2)/SQRT((LV_G2+1)/2);
            ! M3_S = MIVI - ITEM 3, SIGNED
              M3_S = (TAU23-TAU13)/SQRT((LV_G2+1)/2);
            ! M4_A = MIVI - ITEM 4, ABSOLUTE
              M4_A = SQRT((TAU24-TAU14)^2)/SQRT((LV_G2+1)/2);
            ! M4_S = MIVI - ITEM 4, SIGNED
              M4_S = (TAU24-TAU14)/SQRT((LV_G2+1)/2);
            ! M_A = MIVI - ITEM SET, ABSOLUTE
              M_A = ((SQRT((TAU24-TAU14)^2)+SQRT((TAU23-TAU13)^2))/
                    SQRT((LV_G2+1)/2))/4;
            ! MSCALE_S = MIVI - ITEM SET, SIGNED
              M_S = ((TAU24-TAU14+TAU23-TAU13)/
                    SQRT((LV_G2+1)/2))/4;

OUTPUT:     CINTERVAL(BOOTSTRAP) STDYX;
```

Additional File 4: R Code for the Simulated Example 1

```
##################################
#EXAMPLE TO QUANTIFY NON-INVARIANCE VIA MIVIS
##################################

#R version 4.1.1
R.Version()$version.string

#clear working space
rm(list = ls())

#load relevant packages
if (!require(simsem)) { install.packages("simsem"); require(simsem) }
packageVersion("simsem") #version 0.5-16
if (!require(lavaan)) { install.packages("lavaan"); require(lavaan) }
packageVersion("lavaan") #version 0.6-11
if (!require(psych)) { install.packages("psych"); require(psych) }
packageVersion("psych") #version 2.1.9

#specify population model
populationModel <- "
f1 =~ 1*y1 + 0.9*y2 + 1.5*y3 + 1.6*y4
f1 ~~ c(1, 0.6)*f1
f1 ~ 0*1
y1 + y2 ~ 0.50*1
y3 ~ c(0.50, 1.0)*1
y4 ~ c(0.50, 0.9)*1
y1 ~~ 0.30*y1
y2 ~~ 0.30*y2
y3 ~~ 0.30*y3
y4 ~~ 0.30*y4
"

RNGkind(sample.kind = "Rounding")
set.seed(1234)

#sample data from population model
data <- simsem::generate(model=populationModel,n=c(1000,1000), group=group, seed = 1234)

#specify model including MIVIs for items and item sets

#cave: up to the newest lavaan version (0.6-11) it was not possible
#to identify the model via the latent variance in Group 1 while freely estimating
#and labeling the latent variance of Group 2.
#Until this bug persists, one must freely estimate the latent variance
#and plug it, in a second step, into the MIVI formulae.
#This also leads to a bootstrapped confidence interval,
#which is smaller than anticipated (as the standard error of the latent variance in Group 2
is not included).
cfa.MIVI <- "
f1 =~ c(a1,a1)*y1 + c(a2,a2)*y2 + c(a3,a3)*y3 + c(a4,a4)*y4
f1 ~~ c(1,NA)*f1
f1 ~ c(M1,M2)*1
y1 ~ c(b1,b1)*1
y2 ~ c(b2,b2)*1
y3 ~ c(b31,b32)*1
y4 ~ c(b41,b42)*1
y1 ~~ NA*y1
y2 ~~ NA*y2
y3 ~~ NA*y3
y4 ~~ NA*y4

M1 == 0
```

```
#MIVIs - ITEM
MIVI3_Intercept_Absolute := abs(b32-b31)/sqrt((1+0.648)/2)
MIVI3_Intercept_Signed := (b32-b31)/sqrt((1+0.648)/2)

MIVI4_Intercept_Absolute := abs(b42-b41)/sqrt((1+0.648)/2)
MIVI4_Intercept_Signed := (b42-b41)/sqrt((1+0.648)/2)

#MIVIs - ITEM SET
MIVI_Intercept_Absolute := ((abs(b32-b31)+abs(b42-b41))/sqrt((1+0.648)/2))/4
MIVI_Intercept_Signed := ((b32-b31+b42-b41)/sqrt((1+0.648)/2))/4
"

#calculate MIVIs for items and item sets
cfa.MIVI.fit <- lavaan::cfa(cfa.MIVI, data = data, group = "group", estimator = "ml",
missing = "fiml", std.lv=TRUE, se = "bootstrap", bootstrap = 1000)
summary(cfa.MIVI.fit, fit.measures = T)

#bootstrap confidence intervals for MIVIs
lavaan::parameterEstimates(cfa.MIVI.fit)
```