Behavior Research Methods (2025) 57:135
https://doi.org/10.3758/s13428-025-02638-x

ORIGINAL MANUSCRIPT

The simulation-cum-ROC approach: A new approach to generate tailored cutoffs for fit indices through simulation and ROC analysis

Katharina Groskurth[1,2] · Nivedita Bhaktha[2,3] · Clemens M. Lechner[2]

Accepted: 5 February 2025

Abstract

To evaluate model fit in structural equation modeling, researchers commonly compare fit indices against fixed cutoff values (e.g., CFI ≥ .950). However, methodologists have cautioned against overgeneralizing cutoffs, highlighting that cutoffs permit valid judgments of model fit only in empirical settings similar to the simulation scenarios from which these cutoffs originate. This is because fit indices are not only sensitive to misspecification but are also susceptible to various model, estimation, and data characteristics. As a solution, methodologists have proposed four principal approaches to obtain so-called tailored cutoffs, which are generated specifically for a given setting. Here, we review these approaches. We find that none of these approaches provides guidelines on which fit index (out of all fit indices of interest) is best suited for evaluating whether the model fits the data in the setting of interest. Therefore, we propose a novel approach combining a Monte Carlo simulation with receiver operating characteristic (ROC) analysis. This so-called simulation-cum-ROC approach generates tailored cutoffs and additionally identifies the most reliable fit indices in the setting of interest. We provide R code and a Shiny app for an easy implementation of the approach. No prior knowledge of Monte Carlo simulations or ROC analysis is needed to generate tailored cutoffs with the simulation-cum-ROC approach.

Keywords Fit indices · Cutoff · Confirmatory factor analysis · Structural equation modeling · ROC

Introduction

To test the goodness of confirmatory factor analysis (CFA) models—and structural equation models more generally—researchers routinely rely on model fit indices (Jackson et al., 2009; Kline, 2016). Next to the chi-square test of exact model fit (χ^2; e.g., Bollen, 1989)[1], 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). Those fit indices quantify model–data (mis-)fit in a continuous way. However, because fit indices are hard to interpret in isolation, researchers usually rely on cutoffs (or "thresholds") for fit indices that enable them to make binary decisions about whether a model fits the data or not.

Methodologists commonly derive such cutoffs for fit indices from Monte Carlo simulation studies (for an overview and detailed description, see Boomsma, 2013). Such simulation studies examine how fit indices behave across controlled scenarios. Methodologists specify true data-generating (or population) models and determine misspecification of various forms (e.g., in loadings or number of factors) in the analysis model, the model to be tested. By repeatedly generating (i.e., simulating) random data from each population model and fitting the analysis model to each data, they obtain fit index distributions. A cutoff should then represent the fit index value that only rejects the analysis model if it is misspecified.

On the basis of such a simulation study, Hu and Bentler (1999) derived a set of cutoffs that have since become the de

[1] If not otherwise noted, we refer to the normal-distribution χ^2 test statistic here, also referred to as T_{ML} in other articles (e.g., Yuan, 2005).

✉ Katharina Groskurth
katharina.groskurth@gmx.de

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

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

[3] Indian Institute of Technology Kanpur, Kanpur, Uttar Pradesh, India

Published online: 01 April 2025

facto standard in the field. Their simulation study covered a limited set of scenarios assumed to represent typical empirical settings. More specifically, the scenarios always encompassed three-factor models with 15 items. Hu and Bentler specified those models to have varying item and factor distributions, drew samples of various sizes, and misspecified either loadings or factor covariances. Based on their investigation of those scenarios, Hu and Bentler proposed 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.

In empirical applications, researchers compare their obtained fit index values against these cutoffs to evaluate whether their model fits the data (i.e., is assumed to be correctly specified) or not (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 (oftentimes different from the initial simulation scenarios) has been common practice in research involving latent-variable models for decades (e.g., Jackson et al., 2009).

However, such fixed cutoffs for fit indices are more problematic than many researchers may realize. This is because 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 loading magnitudes, the type of estimator, the sample size, and interactions thereof, especially when the model is misspecified (e.g., Groskurth et al., 2024; 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 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 misspecifications, extraneous characteristics, and the interaction between them (Groskurth et al., 2024; Lai & Green, 2016; Moshagen & Auerswald, 2018).

The susceptibility of fit indices to such characteristics other than model misspecification leads to two key challenges in model evaluation. First, *the performance ability of fit indices to detect model misspecification can vary greatly across empirical settings.* Some fit indices react more strongly to misspecification than others in certain settings (and vice versa, e.g., Moshagen & Auerswald, 2018). This differential performance threatens the ability of fit indices to discriminate between correctly specified and misspecified models (e.g., Reußner, 2019). No fit index universally outperforms others (for an overview, see Groskurth et al., 2024; Niemand & Mai, 2018). Second, *cutoffs for fit indices pertain only to specific scenarios* (i.e., combinations of model, estimation, and data characteristics). Simulation studies can only cover a limited number of combinations of such characteristics. In empirical settings that diverge markedly from the simulation scenarios generating the cutoffs, these cutoffs may no longer allow for valid judgments of model fit (e.g., Hu & Bentler, 1999; McNeish & Wolf, 2023a).

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 scenarios. 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, 2023a). 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 erroneous conclusions regarding model fit (Marsh et al., 2004; for examples, see McNeish & Wolf, 2023a). Such erroneous results threaten the integrity of scientific findings.

A solution that has long been proposed is to use tailored 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, 2023a, b). Toward the ultimate aim of helping researchers transition to more valid model evaluation practices via tailored cutoffs, the first goal of this article was to review and summarize existing approaches to generating tailored cutoffs. Such a systematic overview is missing from the current literature. As this review will reveal, existing approaches to generating tailored cutoffs have unique strengths and, while generally superior to fixed cutoffs, share some limitations. Chief among these limitations is that none of the existing approaches allows an evaluation of the differential performance of fit indices. They provide no guidelines on which fit index (out of all fit indices of interest) reacts most strongly to misspecification and, thus, best discriminates between correctly specified and misspecified models in a given setting.

Therefore, the second goal of our article was to introduce a novel approach that builds on—and extends—prior approaches (e.g., McNeish & Wolf, 2023a, b; Millsap, 2013; Pornprasertmanit, 2014). It combines a Monte Carlo simulation, an often-used procedure in psychometrics, with a receiver operating characteristic (ROC) analysis. Our so-called simulation-cum-ROC approach answers two questions: (1) Which fit indices, if any, perform well (or even best) 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, whereas the

best-performing fit index is considered the most decisive for model evaluation. We illustrate this approach with empirical examples and provide complete R code as well as a Shiny app that facilitates its application.

The logic behind generating cutoffs for fit indices

In recent years, methodologists have advocated moving away from using fixed cutoffs and proposed several approaches to generate cutoffs tailored to the empirical setting of interest (e.g., McNeish & Wolf, 2023a; Millsap, 2013; Pornprasertmanit, 2014). Before introducing any of these approaches to tailored cutoffs, we must highlight two important distinctions foundational to generating cutoffs for fit indices, whether fixed or tailored. The first distinction is between the analysis model (i.e., the latent-variable model one seeks to test) and the population model (i.e., the true model that generated the data). The second distinction is between empirical settings and hypothetical scenarios. In an empirical setting (i.e., fitting the analysis model to empirical data to test its fit), one never knows whether the analysis model is correctly specified or misspecified because the true data-generating mechanism (i.e., the population model) is always unknown. By contrast, in a hypothetical scenario (which can be used for simulating data), one knows whether the analysis model is correctly specified or misspecified because one can define both the analysis model and the population model that generates the data. These distinctions between analysis and population models, as well as between empirical settings and hypothetical scenarios, are crucial for all approaches generating cutoffs for fit indices.

It is also pertinent to all approaches to define different hypotheses about how the empirical data might have been generated. Researchers usually follow the Neyman–Pearson approach to hypothesis testing (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 a population model identical (or nearly identical) to the analysis model has generated the data; the analysis model captures all relevant features of the population model. It is correctly specified. H_1 states that an alternative population model different from the analysis model has generated the data; the analysis model is underspecified (i.e., misspecified) compared to the population model to an intolerable degree and fails to capture its relevant features. It is misspecified.

Cutoffs for fit indices, in essence, are needed to discriminate between H_0 and H_1 in empirical settings where the population model is unknown. However, one cannot generate cutoffs in an empirical setting where the population model is unknown; one needs to generate cutoffs in a hypothetical scenario where the population model is known.

The general procedure to derive either fixed or tailored cutoffs is as follows: Fit index distributions for correctly specified (H_0) and misspecified (H_1) analysis models are derived. The goal is to choose a cutoff (e.g., corresponding to a certain percentile from the fit index distributions) that accurately classifies correctly specified models as correctly specified and misspecified models as misspecified. The chosen cutoff should minimize the misclassification of correctly specified models as misspecified (type I error rate) and of misspecified models as correctly specified (type II error rate).

Fixed cutoffs are generated to broadly cover a generic set of hypothetical scenarios that are assumed to occur regularly in empirical settings (e.g., three-factor models with 15 items in the case of Hu and Bentler, 1999). Once created, researchers use this single set of cutoffs across diverse empirical settings. In contrast, tailored cutoff approaches define the hypothetical scenario closely to the empirical setting of interest, such as using the same sample size of the empirical data and the same analysis model of interest. Each time researchers consider a new empirical setting, they must derive a new set of tailored cutoffs.

Once (either fixed or tailored) cutoffs are derived from hypothetical scenarios with known population models, one then uses these cutoffs to test which hypothesis—H_0 or H_1—is more plausible for their analysis model fit to empirical data generated from an unknown population model. If empirical fit index values pass their cutoffs, one accepts the analysis model. Accepting the analysis model means that H_0 seems more plausible than H_1, given the empirical data. If empirical fit index values fail their cutoffs, one rejects the analysis model. Rejecting the analysis model means that H_1 seems more plausible than H_0, given the empirical data. Whether H_0 or H_1 is indeed true will be left unanswered as the population model generating the empirical data always remains unknown (Neyman & Pearson, 1928).[2]

[2] Another way to test whether empirical evidence favors H_0 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 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; MacCallum et al., 1996; 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

Whereas fixed cutoffs are usually derived once in a single simulation study, covering a range of scenarios, various approaches have been specified to derive cutoffs tailored to the specific empirical setting at hand. Currently, there are four principal approaches to generating tailored cutoffs (Table 1)[3] that fall on a continuum from parametric to non-parametric:

(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; Groskurth et al., 2024). The regressions predict cutoffs from various model, estimation, and data characteristics, allowing the researcher to account for characteristics that influence fit indices.
(3) The dynamic simulation approach generates cutoffs based on fit index distributions from an analysis model fit to multiple samples from known population models (McNeish & Wolf, 2023a, b; Millsap, 2007, 2013; Mai et al., 2021; Niemand & Mai, 2018; Pornprasertmanit, 2014).
(4) The bootstrap approach generates cutoffs based on fit index distributions by fitting the analysis model to resampled empirical data transformed as if the analysis model does (or does not) fit it (Bollen & Stine, 1992; Kim & Millsap, 2014).

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, which seems to be partly based on MacCallum et al., 1996; see also Jak et al., 2021; Jobst et al., 2023). The core idea of the χ^2 distribution-based approach to tailored cutoffs is to infer the distributions of correctly specified and misspecified models from the known central and non-central χ^2 distributions. The central and non-central χ^2 distributions can then be used to determine cutoffs. This works both for the χ^2 test statistic itself and for fit indices that incorporate it (such as RMSEA).

The approach rests on the assumption that the χ^2 test statistic follows a central χ^2 distribution if the analysis model is correctly specified—but a non-central χ^2 distribution if the analysis model is misspecified. A non-centrality parameter determines how much the non-central χ^2 distribution deviates from the central χ^2 distribution. Crucially, this non-centrality parameter depends on the misspecification of the analysis model and the sample size (for a detailed description, see Bollen, 1989; Chun & Shapiro, 2009; Moshagen & Erdfelder, 2016).

To derive tailored cutoffs, users define an effect size difference (i.e., some degree of intolerable misspecification based on the non-centrality parameter) between the central and non-central χ^2 distribution. The expected value of the central χ^2 distribution equals the degrees of freedom of the analysis model of interest. It is the distribution for the χ^2 test statistic given that the analysis model is correctly specified. The expected value of the non-central χ^2 distribution equals the degrees of freedom of the analysis model of interest plus the effect size (i.e., the intolerable degree of misspecification defined by the non-centrality parameter). It is the distribution for the χ^2 test statistic given that the analysis model is misspecified. Those two distributions allow users to derive 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, users can only calculate tailored cutoffs from Moshagen and Erdfelder's (2016) χ^2 distribution-based approach under the assumption that items follow a multivariate normal distribution, in which case the χ^2 distribution 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. 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).

[3] One could also add the table-based approach for generating tailored cutoffs to this list (e.g., Groskurth et al., 2024). Reminiscent of looking up critical values for *z*-scores, users read out scenario-specific cutoffs from large tables originating from a prior simulation study. 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 dismissed 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, which 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 (Moshagen & Bader, 2024) Tutorial: Jobst et al. (2023)
Regression: Generating cutoffs based on meta-regressions	Nye & Drasgow (2011)	✓	✗	✗	Sample size and response distribution for RMSEA and SRMR	Regression equation: included in the article
	Groskurth et al. (2024)	✓	✗	✗	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 equation: included in the article R code: included in the article
Dynamic simulation: Generating cutoffs based on fit index distributions generated via a Monte Carlo simulation	Niemand & Mai (2018), Mai et al. (2021)	✓	✗	✗	All model, estimation, and data characteristics for available fit indices	R package: FCO (Niemand & Mai, 2025)
	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 (2023a, b)	✓	✓	✗	All model, estimation, and data characteristics for available fit indices	Shiny app: https://dynamicfit.app/__landing__/ R package: dynamic (Wolf & McNeish, 2022) Mplus code: included in the article
	Pornprasertmanit (2014)	✓	✓	✗	All model, estimation, and data characteristics for available fit indices	
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 article
	Yuan & Hayashi (2003), Yuan et al. (2004, 2007)	✓	✓	✗	All model, estimation, and data characteristics for available fit indices	R package: lavaan (Rosseel, 2012)
Dynamic simulation + ROC analysis	Present article (simulation-cum-ROC)	✓	✓	✓	All model, estimation, and data characteristics for available fit indices	Shiny app: https://kg11.shinyapps.io/tailoredcutoffs/ R code: included in Additional File 1 of the Supplementary Online Material

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 are based on a prior simulation study instead of being specifically derived for the setting of interest, we did not highlight them in this table

Regression-based approach

Another option to generate tailored cutoffs is via the parametric regression-based approach. The basic idea is to predict tailored cutoffs for a given empirical setting using a regression formula (e.g., Groskurth et al., 2024; Nye & Drasgow, 2011). This enables users to account for at least some of the characteristics of their empirical setting when choosing appropriate cutoffs.

The regression-based approach underlies a single, although typically very extensive, simulation study in the background. This simulation ideally covers many different model, estimation, and data characteristics (e.g., one- versus two-factor models, different numbers of items, and different distributions of the items). Cutoffs at certain type I or type II error rates are derived for each of the different scenarios. Cutoffs are considered dependent variables and regressed on the model, estimation, and data characteristics considered in the simulation. The regression formula thus comprises predictors with associated regression weights that contain information about how the various model, estimation, and data characteristics covered in the simulation (e.g., number of items, type of estimator, and distribution of responses) influence cutoffs at certain type I or type II error rates. To derive tailored cutoffs, users simply plug their model, estimation, and data characteristics of their setting of interest into the formula.

Hence, users predict tailored cutoffs by using a regression formula from a large, ideally extensive simulation study. Although the simulation study underlying the regression formula should be extensive, it does not necessarily cover a scenario similar to the empirical setting of interest. However, the formula allows for extrapolation; thus, it allows for the prediction of cutoffs for settings not covered in the initial simulation scenarios. Although extrapolation is only advisable for empirical settings that do not diverge strongly from the scenarios in the initial simulation, it helps to tailor cutoffs to a wider variety of settings than initially covered by the simulation scenarios.

Nye and Drasgow (2011) and Groskurth et al. (2024) followed the regression-based approach. Nye and Drasgow (2011) provided regression formulae for RMSEA and SRMR. Besides the cutoffs, they considered the response distribution, the sample size, and the type I error rate in the formulae. Their models had two factors, 15 items, and they estimated them with diagonally weighted least squares. Groskurth et al. (2024) considered more fit indices and a much wider range of characteristics: They provided regression formulae for χ^2, χ^2/degrees of freedom, CFI, RMSEA, and SRMR. Estimators, number of items, response distributions, response options, loading magnitudes, sample size, and number of factors served as predictors in the formulae.

Similar to the χ^2 distribution-based approach, the regression-based approach has the advantage of speed. Users merely have to plug the characteristics of their empirical setting into the formula, commonly solved by a statistical tool such as R. However, the regression formula is only as inclusive as the simulation study from which it was derived—although extrapolation is possible for settings different from the initial simulation scenarios. Further, users can only obtain cutoffs for those fit indices that are considered in the simulation study from which the regression formula hails. Akin to the χ^2 distribution-based approach, the regression-based approach limits the extent to which users can tailor cutoffs to their specific empirical setting and the range of fit indices (see Table 1).

Dynamic simulation approach

A third approach that allows for a much greater extent of tailoring cutoffs is what we call the "dynamic" simulation approach (following McNeish & Wolf, 2023a). Like the fixed cutoff approach (but also like the regression-based approach for tailored cutoffs), the dynamic approach uses Monte Carlo simulations to generate cutoffs. Crucially, however, the simulations are performed *for the specific empirical setting at hand* on a case-by-case basis—instead of relying on generic simulation results (McNeish & Wolf, 2023a, b; Millsap, 2007, 2013; Mai et al., 2021; Niemand & Mai, 2018; Pornprasertmanit, 2014; for nested models, see Pornprasertmanit et al., 2013).

Simulation scenarios are well known (dating back to the initial Hu & Bentler, 1999, article); we describe them in detail here to enable users to apply the dynamic simulation approach: Users need to define a population model, simulate data (i.e., draw multiple samples) from that population model, and fit an analysis model to the simulated data. 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. After fitting the analysis model to the data, users record the fit index values of each analysis model. A cutoff can then be set based on a specific percentile, commonly the 95th or 90th, of the resulting fit index distribution (equivalently, on the 5th or 10th for fit indices where higher values indicate better fit). At this percentile, the cutoff categorizes 95% or 90% of correctly specified models as correctly specified and 5% or 10% of correctly specified models as misspecified (i.e., the type I error rate).

Users may repeat the procedure with the same analysis model but a population model with more parameters than the analysis model. As such, one fixes non-zero parameters in the population model to zero in the analysis model (e.g., Hu & Bentler, 1998). The analysis model is, thus, underspecified (i.e., misspecified) relative to the population model; it

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).

Crucially, to arrive at cutoffs *tailored* to the setting of interest, the analysis and population models are not just any models but are chosen to match the given empirical setting. Each time users assess a new empirical setting (i.e., different model, estimation, and data characteristics), they must derive a new set of cutoffs via Monte Carlo simulations. This makes the approach dynamic and distinguishes it from approaches that rely on generic simulation studies (i.e., most prominently, the fixed cutoff approach). Thus, the dynamic simulation approach eliminates the problem that empirical settings may deviate from scenarios underlying the cutoffs by specifying the simulation scenario just like the empirical setting of interest.

The dynamic simulation approach is computationally intensive, more intensive than the χ^2 distribution-based and regression-based approaches, because a simulation study has to be run anew for every setting of interest. For the same reason, it has the advantage of being very flexible. It generates tailored cutoffs for *all* fit indices available in a given statistical program to the specific model, estimation, and data characteristics of the analysis setting at hand (see Table 1). Combined with computers' continuously increasing statistical power, this is one of the reasons why this approach has recently gained traction (McNeish & Wolf, 2023a, b).

Bootstrap approach

Cutoffs tailored to the given analysis setting at hand can not only be generated via (dynamic) Monte Carlo simulations, which are essentially parametric bootstrap approaches (simulating, i.e., resampling, data based on model parameters), but also via non-parametric bootstrap approaches (i.e., resampling observed data). A fourth approach uses such non-parametric bootstrapping to generate tailored cutoffs from empirical data transformed as if the analysis model does (or does not) fit it (Bollen & Stine, 1992; Kim & Millsap, 2014; Yuan & Hayashi, 2003; Yuan et al., 2004, 2007).

In the following, we illustrate Bollen and Stine's (1992) and Kim and Millsap's (2014) bootstrap approach in more detail. The 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, users obtain data that behaves as if the analysis model had generated it. The algorithm resamples the transformed data (with replacement), fits the analysis model to each resampled data, and records the values of fit indices for each. The bootstrap approach 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 distributions. 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 very flexible, similar to the dynamic simulation approach (see Table 1). Through repeated resampling, users can generate cutoffs for all available fit indices tailored to all choice characteristics. This comes at the expense of greater computational intensity than the χ^2 distribution-based and regression-based approaches.

Limitations of the existing approaches

All four approaches of generating tailored cutoffs have their merits and constitute a clear advancement over fixed cutoffs, allowing for more valid cutoffs that control type I and/or type II errors. 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 approaches, both parametric). Others 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 dynamic simulation approach and the non-parametric bootstrap approach).

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) reacts most strongly to misspecification. Knowing which fit index is, thus, best able to discern correctly specified from misspecified models in the setting of interest would 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 an approach that builds on previous approaches and extends them by (1) identifying well-performing (and best-performing) fit indices in a specific setting of interest while (2) generating tailored cutoffs that balance 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 various settings that researchers may encounter.

A novel approach to tailored cutoffs: The simulation-cum-ROC approach

Our so-called simulation-cum-ROC approach augments the dynamic simulation approach (e.g., McNeish & Wolf,

2023a; Millsap, 2013; Pornprasertmanit, 2014) that is currently gaining traction among applied researchers and builds on a long tradition of generating cutoffs through Monte Carlo simulations (dating back to the initial Hu & Bentler, 1999, article). The unique contribution of our approach is combining the dynamic simulation approach with receiver operating characteristic (ROC) analysis. The ROC analysis 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 article (i.e., χ^2, CFI, RMSEA, SRMR). Further, the dynamic simulation approach, in combination with ROC analysis, 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, the simulation-cum-ROC approach works as follows. First, we use a Monte Carlo simulation to generate data from two population models, each representing different assumptions about the true data-generating mechanism. One population model is structurally identical to the analysis model one seeks to test, 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. Those simulations are conducted for the empirical setting of interest and, thus, resemble what is done in other dynamic simulation approaches (e.g., McNeish & Wolf, 2023a; Millsap, 2013; Pornprasertmanit, 2014). Second, as a new feature, we analyze the fit index distributions with ROC analysis in addition to what is done in dynamic simulation approaches. 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 not for all fit indices of interest but only for those that appear well-performing in the given scenario—these cutoffs balance type I and type II error rates. We visualized the three steps to generate tailored cutoffs for well-performing fit indices in Fig. 1.

Fundamentals of ROC analysis

Before outlining the details of our simulation-cum-ROC 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 popularity in machine learning (for an overview, see Majnik & Bosnić, 2013).

Fit indices are essentially continuous classifiers that typically indicate better fit for correctly specified models and poorer fit for misspecified models. The cutoffs for these fit indices serve as decision thresholds. These cutoffs should be selected to maximize the share of analysis models that are classified as either 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

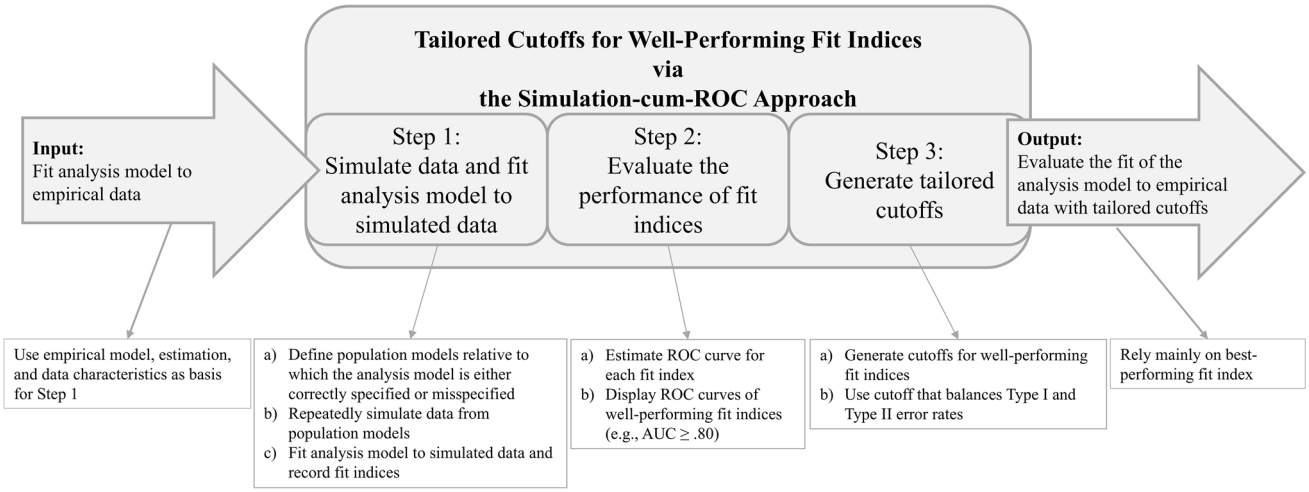


Fig. 1 The simulation-cum-ROC approach

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

$$Sensitivity\,(or\ true-positive\ rate) = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives + Number\ of\ False\ Negatives} \tag{1}$$

$$Specificity\,(or\ true-negative\ rate) = \frac{Number\ of\ True\ Negatives}{Number\ of\ True\ Negatives + Number\ of\ False\ Positives} \tag{2}$$

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, i.e., the 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 (i.e., the total number of converged models in simulation runs):

$$Accuracy = \frac{Number\ of\ True\ Positives\ + Number\ of\ True\ Negatives}{Number\ of\ True\ Positives + Number\ of\ True\ Negatives + Number\ of\ False\ Positives + Number\ of\ False\ Negatives} \tag{3}$$

An 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 done here, following Thiele & Hirschfeld, 2021). The graph visualizing the ROC curve has 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

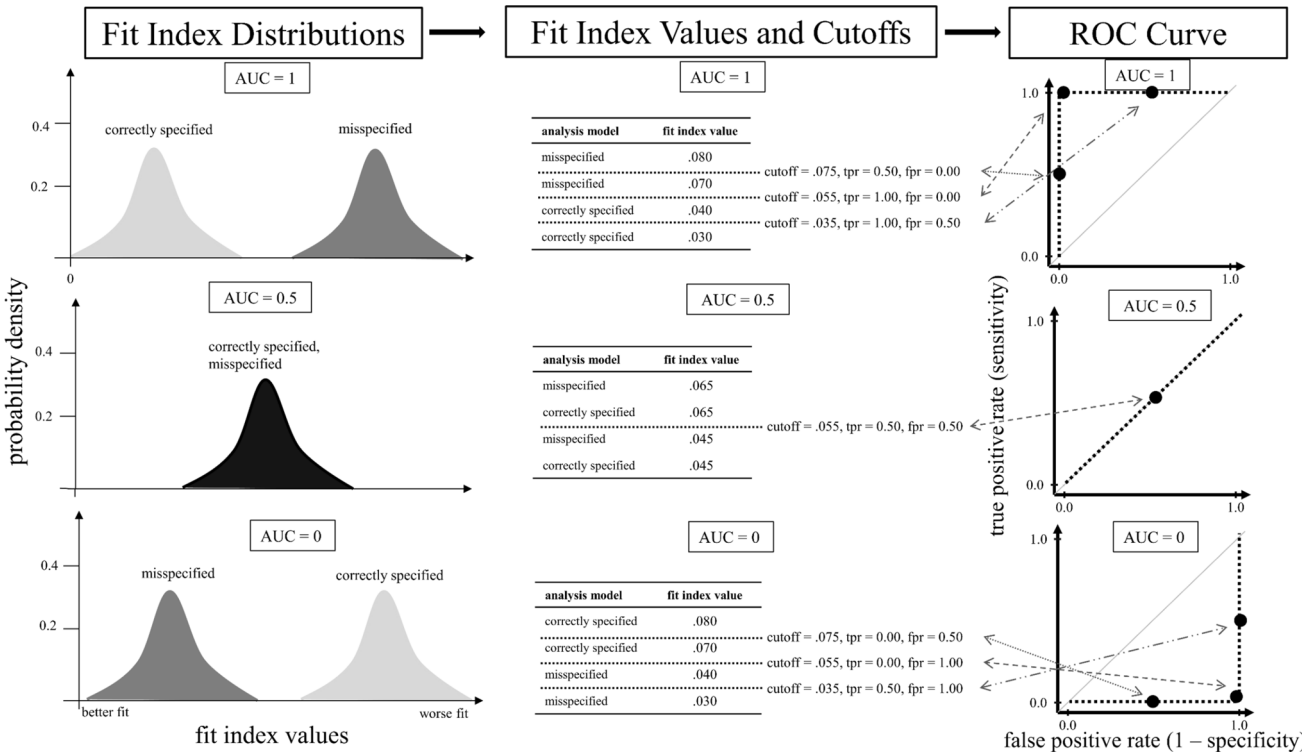


Fig. 2 Relation of fit index distributions, cutoffs, and the ROC curve for different AUCs. *Note.* The figure shows fit index distributions and a sample of their fit index values. It further includes true- and false-positive rates of cutoffs estimated from the sample of fit index values. The ROC curve 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 AUCs. tpr = true-positive rate; fpr = false-positive rate

a fit index, true- and false-positive rates of cutoffs, ROC curve, and AUC values in Fig. 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-positive rate of 0. Thus, 100% of the time, the fit index will correctly discriminate 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-positive 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 (see Fig. 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—the ROC curve is an ascending diagonal. Fit index distributions from correctly specified and misspecified models completely overlap; no distinction is possible (see Fig. 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 Fig. 2).

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 simulation-cum-ROC approach to (1) evaluate the performance of fit indices and (2) generate tailored cutoffs. We walk the reader through each step of the procedure shown in Fig. 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, forms the basis of our empirical setting. We fit our analysis model—a one-factor CFA model—to the empirical data using robust maximum likelihood (MLR)[4] estimation.

We aim to test two hypotheses. H_0 states that a population model identical (or nearly identical) to the analysis model (i.e., a one-factor model) has generated the empirical data; the analysis model captures all relevant features of the population model. If empirical evidence favors H_0, we want to accept this analysis model. H_1 states that an alternative population model different from the analysis model has generated the empirical data; the analysis model is misspecified compared to the population model to an intolerable degree and fails to capture its relevant features. If empirical evidence favors H_1, we want to reject the analysis model. Thus, we define two diverging states of the world that describe how the empirical data may have been generated (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 obtain empirical fit index values fitting the analysis model to empirical data and compare those against cutoffs tailored to the specific characteristics of our empirical setting. We generate 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, loading magnitudes), the analytical strategy (e.g., MLR estimator), and the data characteristics (e.g., $N = 500$, multivariate distribution). The simulation-cum-ROC approach shares this basic idea with other dynamic simulation approaches (e.g., McNeish & Wolf, 2023a, b).

More specifically, following the Neyman–Pearson approach, we operationalize the two competing hypotheses, H_0 versus H_1, about the population model that may have generated the empirical data in the setting of interest through a Monte Carlo simulation scenario. Thereby, researchers need to define the H_0 and H_1 population models (e.g., Millsap, 2007, 2013; cf. McNeish & Wolf, 2023a, b). Whereas the H_0 population model oftentimes simply equals the analysis model, the H_1 population model is harder to define; researchers need to specify a certain degree of intolerable

[4] MLR (Muthén & Muthén, 1998-2017; Yuan & Bentler, 2000a) is a variant of the most commonly used ML estimator (Bollen, 1989; Jackson et al., 2009). MLR assumes continuous data but corrects the χ^2 test statistic and standard errors of ML-estimated parameters for non-normality with the help of scaling factors. It yields relatively unbiased model parameters for ordered categorical data with a sufficient number of response options (i.e., five or more; Beauducel & Herzberg, 2006; Rhemtulla et al., 2012). Thus, MLR is very well-known and appropriate in a wide variety of empirical settings, which is why we chose it throughout this article.

misspecification of the analysis model compared to the H_1 population model.

As H_1 population models are hard to define, they are usually predefined in dynamic simulation approaches (e.g., McNeish & Wolf, 2023a, b) and, thus, hidden from the researcher. For example, McNeish and Wolf (2023a, b) always use the same H_1 population model in terms of model structure and additional parameters for all analysis models of the same type (e.g., one-factor CFA models).

We decided not to predefine the H_1 population model but leave the definition of the H_1 population model to the researcher (e.g., Millsap, 2007, 2013). To aid researchers in defining H_1 population models, we provide guidance on defining the form and quantifying the degree of misspecification in the Discussion section. This guidance should make the definition of H_1 population models more comparable and, thus, objective.

In our view, having the researcher explicitly specify the H_1 population model is favorable to relying on implicit ones. It makes assumptions about the H_1 population model transparent; researchers need to think about and justify their definition of intolerable misspecification. It is very important that researchers transparently outline their choices and provide a strong rationale for their hypotheses and models. Providing a strong rationale aligns 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).

After defining the population models, we simulate data from the H_0 population model structurally identical to the analysis model of interest (i.e., a one-factor CFA model). We also simulate data from the H_1 population model that diverges substantially from the analysis model. For example, an H_1 population model could have two factors, whereas the analysis model of interest has one factor. Notably, model are usually nested but they do not necessarily need 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 flexible regarding model definition.[5,6]

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 H_0) and misspecified models (under H_1).

Step 2: Evaluate the performance of fit indices

After simulating data and obtaining fit index distributions, we evaluate and rank the performance of fit indices on the simulated data via the ROC curve and the AUC in particular, which is a unique feature of the simulation-cum-ROC approach. Both the ROC curve and the AUC reflect the balance of a fit index between the true-positive rate, or sensitivity, and the false positive rate, or 1 – specificity, at different potential cutoffs. The closer the fit index's AUC is to 1, the higher its sensitivity and specificity across different potential cutoffs. Thus, the AUC quantifies how well a fit index discriminates between correctly specified and misspecified models—as such, we can rank fit indices according to their ability to detect misspecification in the specific scenario. Hence, the AUC provides guidance regarding which fit index (or indices) is best to judge the model's fit. The idea of the simulation-cum-ROC approach is to consider only well-performing fit indices in the evaluation of model fit, with the best-performing fit index being the most decisive one.

In the following, we define those fit indices as well-performing 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 will correctly discriminate between correctly specified and misspecified models at different potential cutoffs (e.g., D'Agostino et al., 2013). Notably, the 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 thresholds—especially as type I and type II error rates of the corresponding cutoffs can exceed conventional levels of 5% or 10% at such an AUC threshold.[7]

[5] Readers should not confuse H_1 population models, to which we refer here, with saturated H_1 models (e.g., Savalei & Rosseel, 2022). Whereas a saturated model perfectly reproduces the structure of the empirical data, the H_1 population model used in the simulation-cum-ROC approach is just any reasonable population model compared to which the analysis model of interest is considered misspecified.

[6] 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 decision between H_0 and H_1 for the analysis model tested with empirical data.

[7] The AUC might be unexpectedly low even if the models under H_0 and H_1 actually differ substantially. This can be the case if the empirical, and accordingly the simulated data which is orientated upon the empirical one, has heavy tails. If models are fit to data with heavy tails, parameter estimates can be inefficient, which might lead to a bad discrimination ability of fit indices between the analysis model under H_0 and H_1. Robust methods that downweigh cases in the tail area, applied to the empirical and simulated data, lead to more efficient parameter estimates. Thus, those methods can help to improve the separation of fit index values under H_0 and H_1 (see Yuan et al., 2004, on identifying the optimal robust procedure for the data of interest).

Although we focus on well-performing fit indices with an AUC above a certain threshold (e.g., .80) to evaluate a model's fit, inspecting the distributions of low-performing fit indices can also be informative. Strongly overlapping fit index distributions (i.e., AUC around .50) imply that a fit index cannot discern correctly specified from misspecified models. If few fit indices have strongly overlapping distributions, those particular fit indices might not be able to detect misfit in the scenario of interest. However, if several fit indices have strongly overlapping distributions, the misspecification of the analysis model relative to the H_1 population model might not be strong enough to be detected in the scenario of interest. Similarly, the analysis model might be flexible enough to account for data from both H_0 and H_1 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). Thus, even strongly overlapping distributions (i.e., AUC around .50) may provide important insights.

Generally, it is important to bear in mind that 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). This is why fit indices perform differently well in different scenarios. Thus, the shape and overlap of distributions for each fit index help understand models (and accordingly misfit) further—as fit indices characterize models differently (see Browne et al., 2002; Lai & Green, 2016; Moshagen & Auerswald, 2018).

Step 3: Generate tailored cutoffs

After identifying well-performing fit indices (e.g., AUC $\geq$ .80) and screening out the others, we can identify optimal cutoffs. To arrive at optimal cutoffs with the simulation-cum-ROC approach, we employ ROC analysis to select an optimal cutoff at the highest sum of sensitivity and specificity and, thus, the highest accuracy for each fit index. We interpret the type I error rate (i.e., 1 – specificity) and type II error rate (i.e., 1 – sensitivity) as equally problematic.[8] At cutoffs with balanced type I and type II error rates, fit indices can best classify correctly specified models as correctly specified and misspecified models as misspecified.

We visualize the fit index distributions from correctly specified and misspecified models in a graph and provide cutoffs along with their accuracy, type I error rate, and type II error rate. 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).[9] An essential advantage of the simulation-cum-ROC approach is that it returns the error rates associated with applying cutoffs. It draws researchers' attention to how well cutoffs discriminate between correctly specified and misspecified models in the context of interest (quantified through type I and type II error rates).

Optimal cutoffs are not only identified in the simulation-cum-ROC approach but also in dynamic simulation approaches, though the strategies of the two approaches are different. The idea of the simulation-cum-ROC approach is to rank the fit indices by their performance. Optimal cutoffs are derived for all fit indices that meet an AUC threshold (commonly .80) and are, thus, considered well-performing. Common dynamic simulation approaches do not incorporate the feature to rank fit indices; optimal cutoffs are provided for all fit indices that meet conventional requirements of type I or type II error rates (commonly 5%/95% or 10%/90%, e.g., McNeish & Wolf, 2023a, b).

Thus, optimal cutoffs obtained via the simulation-cum-ROC approach do not need to meet certain requirements of type I or type II error rates and are typically derived at balanced error rates. However, those cutoffs might exceed conventional type I and type II error rates (i.e., 5%/95% and 10%/90%). Researchers then have the freedom to decide if they are willing to accept those so-obtained type I and type II error rates—and, accordingly, they have the freedom to use those so-obtained cutoffs.

Suppose researchers deem the type I and type II error rates to be too large. In that case, they need to impose stronger misspecification by redefining the H_1 population model and, thus, adjust their assumptions about the level of misfit they want to reject. The derivation of tailored cutoffs needs to be redone. The initial and revised assumptions must be explicitly outlined and justified.

Thus, different from previous dynamic simulation approaches (e.g., McNeish & Wolf, 2023a, b), the simulation-cum-ROC approach allows for more researcher degrees

[8] The simulation-cum-ROC approach also allows to consider one error as more problematic than the other by changing the *metric* for cutpointr (Thiele & Hirschfeld, 2021) in the R code to either *sens_constrain* (if the type II error is considered more problematic) or *spec_constrain* (if the type I error is considered more problematic).

[9] 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 AUCs of different fit indices are only marginally different from each other.

of freedom but also forces the researcher to think about their choices regarding the hypotheses and models carefully, make them transparent, and provide a strong rationale for them.

Output: Evaluate the fit of the analysis model to empirical data with tailored cutoffs

Having generated tailored cutoffs for well-performing fit indices with the simulation-cum-ROC approach, we can evaluate how well our analysis model (i.e., the 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, the analysis model has a good fit. Given the empirical data, H_0 seems more plausible than H_1. We can accept the analysis model. If all empirical values of fit indices fail the proposed tailored cutoffs, the analysis model has a poor fit. Given the empirical data, H_1 seems more plausible than H_0. 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 uniquely provided by the simulation-cum-ROC approach. If there is a best-performing fit index and its empirical value suggests that the analysis model fits (i.e., it passes its tailored cutoff), H_0 seems more plausible than H_1. We accept the analysis model. If the best-performing fit index suggests that the analysis model does not fit, H_1 seems more plausible than H_0. We reject the analysis model. Thus, in those less-straightforward settings, we prioritize the best-performing fit index and its corresponding cutoff for our decision on model fit.

Rejecting the analysis model implies that the model is misspecified to the extent it was misspecified compared to the H_1 population model—or even to a larger extent. Hence, rejecting the analysis model informs us about the severity of misspecification relative to the H_1 population model. It does not inform us about the specific alternative model that has generated the data—this remains unknown as in all empirical settings.

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 must repeat the above procedure (Steps 1 to 3 of the simulation-cum-ROC approach) once we state a new H_0 and H_1.

Application of the simulation-cum-ROC approach

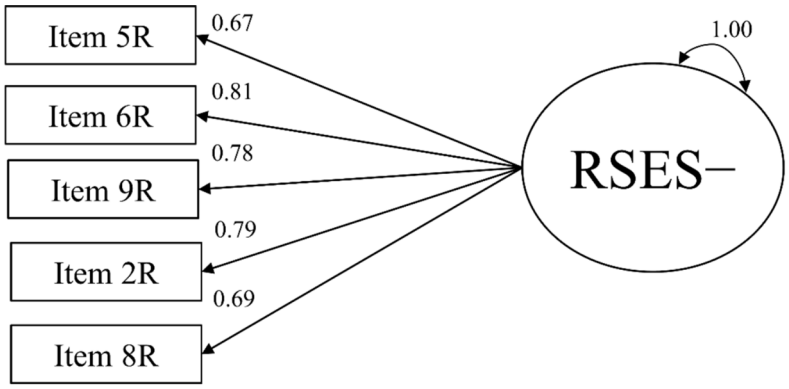
In the following, we provide two examples that illustrate the simulation-cum-ROC approach. The aim of the first example is to illustrate the three steps to generate and apply tailored cutoffs. We chose a simple example without complications for the purpose of this illustration. All fit indices performed equally well in this example, which is not always guaranteed in real-life empirical applications.

The 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, the fit indices of interest differed in their performance; 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 4.4.1; R Core Team, 2024). We used the R package lavaan to fit the models (version 0.6.19; Rosseel, 2012), simsem to simulate the data (version 0.5.16; Pornprasertmanit et al., 2021), pROC to plot the ROC curves (version 1.18.5; Robin et al., 2011), and cutpointr to obtain cutoffs for fit indices (version 1.1.2; Thiele & Hirschfeld, 2021). We documented all other packages used in the R code. Additional File 1 of the Supplementary Online Material includes the computational code.

We also programmed a Shiny app available under https://kg11.shinyapps.io/tailoredcutoffs/. Specifically, one needs to plug in their analysis model, population models, marginal skewness and excess kurtosis of the response distribution (used to obtain multivariate non-normal data with Vale and Maurelli's method, 1983[10]), estimator, sample size, number of simulation runs, fit indices one is interested in, and the AUC threshold. The Shiny app internally runs through Steps 1 to 3 of the simulation-cum ROC approach. It allows convenient downloading of the ROC curves from Step 2 and of the fit index distributions and tailored cutoffs from Step 3. Users need not execute any statistical program locally; the Shiny app does all the computational work to arrive at tailored cutoffs within the simulation-cum-ROC approach.

[10] Olvera Astivia and Zumbo (2015) scrutinized Vale and Maurelli's method and found that estimates of skewness and kurtosis were downward-biased, especially in small samples. We still relied on Vale and Maurelli's method to obtain multivariate non-normal data as it is the standard method in the simsem package (Pornprasertmanit et al., 2021) used here. However, researchers can adapt the code provided in this article estimating multivariate non-normal data using other functions (e.g., the functions of the R package covsim, Grønneberg et al., 2022, can be called in simsem's function to simulate data, *sim*, but covsim's functions require to specify a population variance-covariance matrix instead of a population model).



Fit indices	χ^2	*df*	CFI	RMSEA	SRMR
Empirical values	17.97**	5	.981	.074	.021

Fig. 3 Empirical one-factor Rosenberg Self-Esteem Scale model (negative feelings). *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. $N = 468$. ** $p < .01$

Example 1: The Rosenberg Self-Esteem Scale

We chose the Rosenberg Self-Esteem Scale as a first example for generating tailored cutoffs via the simulation-cum-ROC approach (Rosenberg, 1965). The Rosenberg Self-Esteem Scale measures global self-esteem with ten items (five referring to positive feelings and five to negative ones) rated on a four-point Likert scale. Initially thought to measure 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). In this example, we focused only on one of the two factors, the one for negative feelings, and evaluated its unidimensionality. 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 one-factor model to the empirical data using MLR. Figure 3 depicts the analysis model and the empirical fit index values. We evaluated whether empirical evidence favors H_0 or H_1 for the one-factor model using tailored cutoffs. We would accept the one-factor model if empirical evidence favored H_0, stating that a population model identical (or nearly identical) to the one-factor model had generated the data; the one-factor model captured all relevant features of the population model. We would reject the one-factor model if empirical evidence favored H_1, stating that a population model different from the one-factor model had generated the data; a one-factor model failed to capture relevant features of the population model to an intolerable degree.

Step 1: Simulate data and fit analysis model to simulated data After fitting the one-factor model to empirical data, we defined H_0 and H_1 for the Monte Carlo simulation. The one-factor model served as an analysis model in the simulation. The structure and parameter estimates of the one-factor model fit to empirical data served as the H_0 population model. Relative to the one-factor population model, the analysis model was correctly specified.

Next, we must define a theoretically justifiable H_1 population model. A good candidate for an H_1 population model could be a two-factor rather than a one-factor model, as the question of dimensionality is at the core of model testing (Brown, 2015). We set the factor correlation between two factors (which equals 1 for the one-factor model) to .70, inducing a misspecification of $r = .30$, a correlation considered medium (Cohen, 1992). Thus, we chose an H_1 population model identical to the H_0 population model (and, thus, the analysis model) in the parameter estimates (i.e., loadings and residual variances) but split into two factors correlating at .70 (with one factor containing the items explicitly referring to feelings). Relative to the two-factor population model, the analysis model was 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 one-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 data 500 times from each population model.

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 only considered fit indices

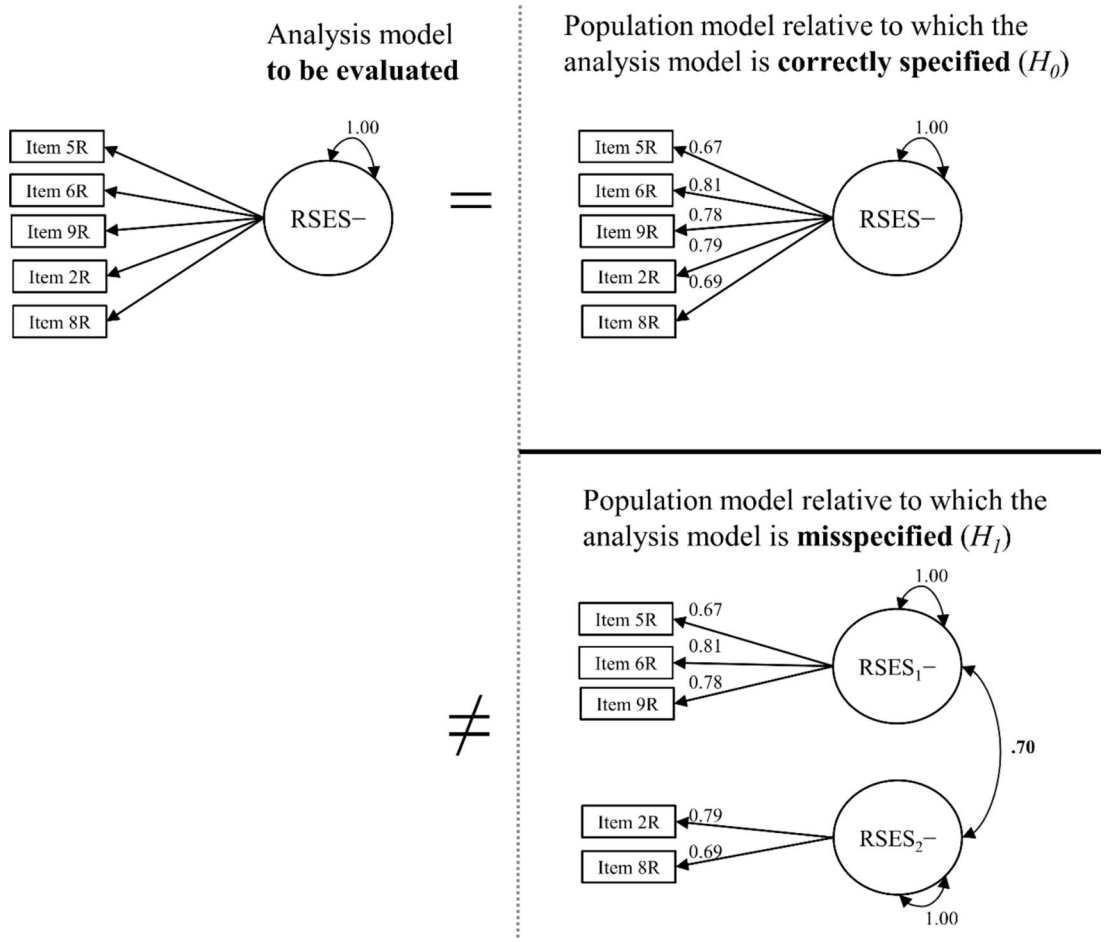


Fig. 4 Proposed analysis and population models of the Rosenberg Self-Esteem Scale (negative feelings). *Note.* Unstandardized coefficients. RSES = Rosenberg Self-Esteem Scale. We recoded the items so that higher values imply higher self-esteem. We omitted the residual variances and the mean structure for clarity

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). In Fig. 5, the different lines representing the different ROC curves completely overlap for all fit indices. All fit indices had an AUC equal to or higher than .80, namely an AUC of 1. Thus, the ROC curves and AUCs were the same for *all* fit indices, which implies that all fit indices discriminated equally well between correctly specified and misspecified models. This is certainly not always the case, as shown in the second example.

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).[11] The cutoffs were the following: $\chi^2(5) \leq 28.03$, CFI ≥

[11] As evident from Fig. 6, we 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 poor model fit (e.g., CFI = 0.00, RMSEA = 1.00). It evaluates sensitivity and specificity at each potential cutoff. 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 poor model fit). We let the algorithm take the mean between the optimal cutpoint (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. By this, we obtain a revised optimal cutpoint, which is the mean of the optimal cutpoint and the next fit index value.

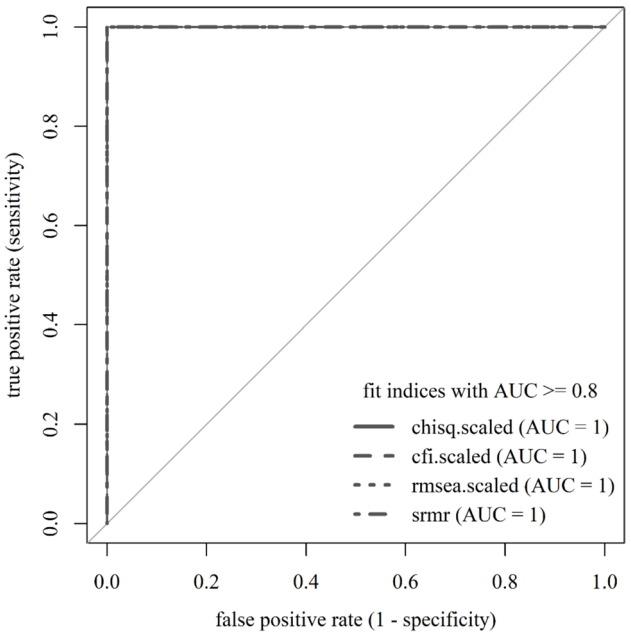


Fig. 5 ROC curves for fit indices with AUC ≥ .80 of the Rosenberg Self-Esteem Scale model (negative feelings). *Note.* Chisq.scaled is a χ^2 test statistic asymptotically equivalent to the robust Yuan–Bentler test statistic (Yuan & Bentler, 2000a) to account for non-normality. Cfi.scaled is the CFI version and rmsea.scaled is the RMSEA version calculated with this test statistic

.972, RMSEA ≤ .097, SRMR ≤ .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 scenario. Again, perfectly discriminating cutoffs do not reflect the usual case, as shown in the second example.

Output: Evaluate the fit of the analysis model to empirical data with tailored cutoffs Judged against the tailored cutoffs, we accepted the one-factor model for the negative feelings of the Rosenberg Self-Esteem Scale fit to empirical data. All of the empirical fit index values for the one-factor model ($\chi^2(5) = 17.97$, $p < .01$; CFI = .981; RMSEA = .074; SRMR = .021) passed the tailored cutoffs (i.e., $\chi^2(5) \leq 28.03$; CFI ≥ .972; RMSEA ≤ .097; SRMR ≤ .031). Given the empirical data, H_0 seemed more plausible, stating that the one-factor model generated the data.

Whereas according to fixed cutoffs of CFI around .950 and SRMR around .080 (but not RMSEA around .060; Hu & Bentler, 1999), the one-factor would also fit and be accepted, we were more confident that the tailored cutoffs correctly classified the one-factor model as correctly specified. Those fixed cutoffs were generated from three-factor models with 15 items in total (Hu & Bentler, 1999)—largely different from the empirical setting at hand. The tailored cutoffs applied here, in turn, were explicitly targeted at our one-factor model with five items (and all the other characteristics of the empirical setting at hand). Additionally, we knew that all fit indices performed equally well and are, thus, equally decisive for model evaluation. This question would be left unanswered with fixed cutoffs for fit indices as well as other approaches to tailored cutoffs.

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 in a favorable light (e.g., 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 Likert 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. We evaluated whether empirical evidence favors H_0 or H_1 for the two-factor model using tailored cutoffs. We would accept the two-factor model if empirical evidence favored $H_{0,}$ stating that a population model identical (or nearly identical) to the two-factor model had generated the data; the two-factor model captured all relevant features of the population model. We would reject the two-factor model if empirical evidence favored $H_{1,}$ stating that a population model different from the two-factor model had generated the data; a two-factor model failed to capture relevant features of the population model to an intolerable degree.

Step 1: Simulate data and fit analysis model to simulated data After fitting the two-factor model to empirical data, we defined H_0 and H_1 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 analysis model was correctly specified.

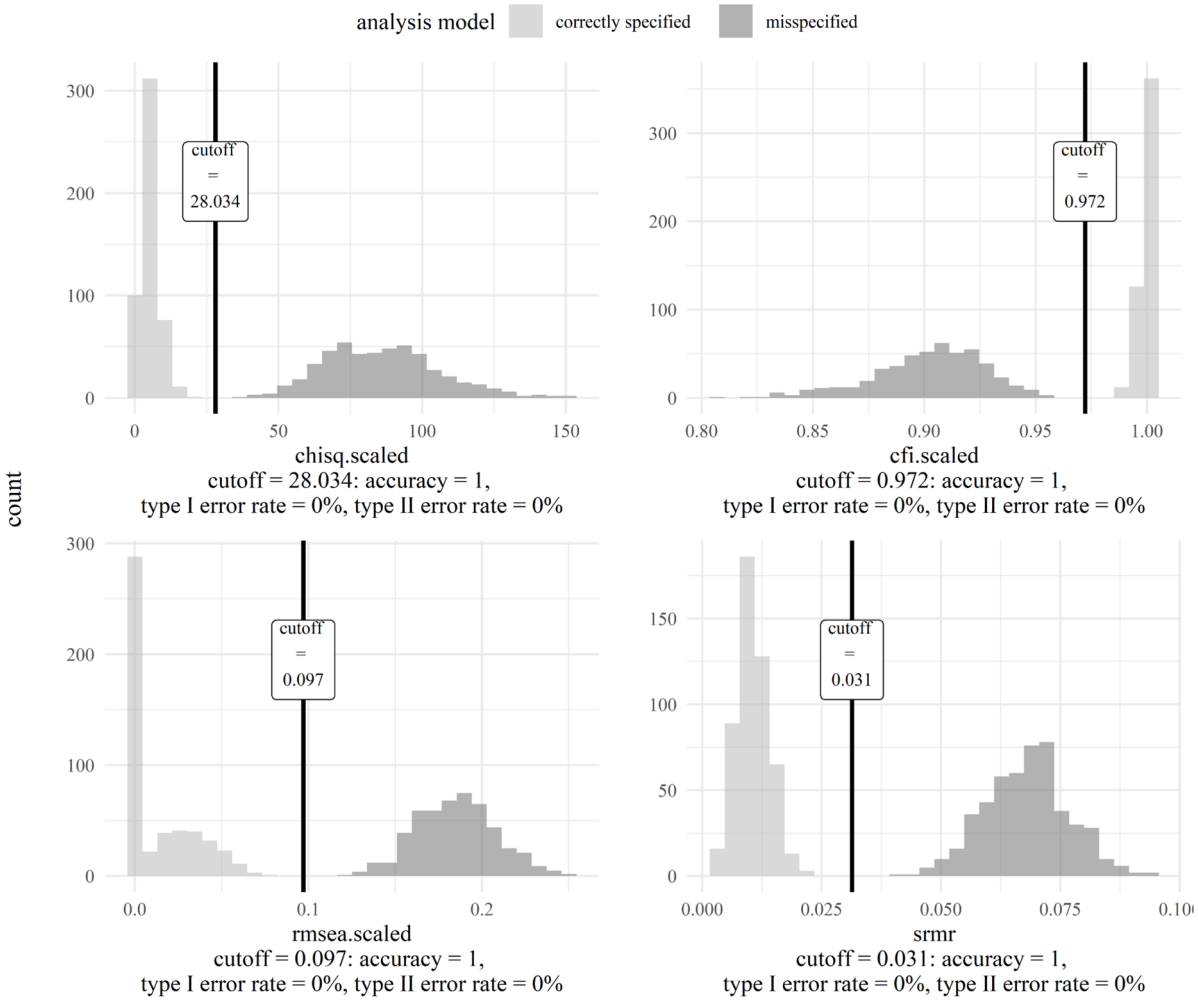
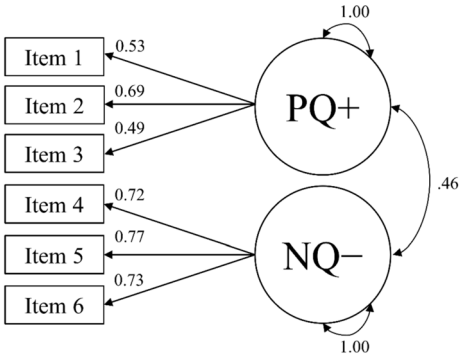


Fig. 6 Cutoffs for fit indices with AUC ≥ .80 of the Rosenberg Self-Esteem Scale model (negative feelings). *Note.* Chisq.scaled is a χ^2 test statistic asymptotically equivalent to the robust Yuan–Bentler test statistic (Yuan & Bentler, 2000a) to account for non-normality. Cfi.scaled is the CFI version 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 are colored in an even darker gray 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)

Next, we must define a theoretically justifiable H_1 population model. 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 fully account for the covariances among items 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, 2023a). 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 two items of the PQ+ factor (resulting in a residual covariance of 0.20), both asking for emotional control. We modeled another residual correlation between the first and third items of the NQ− factor (resulting in a residual covariance of 0.31), both referring to behavior in social interactions. Relative to this H_1 population model, the analysis model was underspecified (i.e., misspecified). Figure 8 shows the population and analysis models for examining H_0 and H_1.

We simulated data from the population models, fit the analysis model to each simulated data, and recorded the fit indices. Essential features for the simulation mimicked the empirical setting in terms of the sample size (i.e., $N = 474$),



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

Fig. 7 Empirical two-factor Social Desirability-Gamma Short Scale model. *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. $N = 474$. *** $p < .001$

estimator (i.e., MLR), and the multivariate response distribution. We simulated data 500 times from each population model.

Step 2: Evaluate the performance of fit indices Unlike the previous example, not all fit indices passed the AUC ≥ .80 benchmark, and the AUCs were generally lower (i.e., below 1). 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 scenario. 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 (= .94) and was, thus, the best-performing fit index in our scenario.

Step 3: Generate tailored cutoffs We generated cutoffs only for the three well-performing fit indices in the following. The cutoff for χ^2 was 11.68, RMSEA .031, and SRMR .025 (Fig. 10). In line with the AUC, the cutoff for SRMR had the highest accuracy (= .87) as well as the lowest type I error rate (= 14%) and type II error rate (= 12%). It better

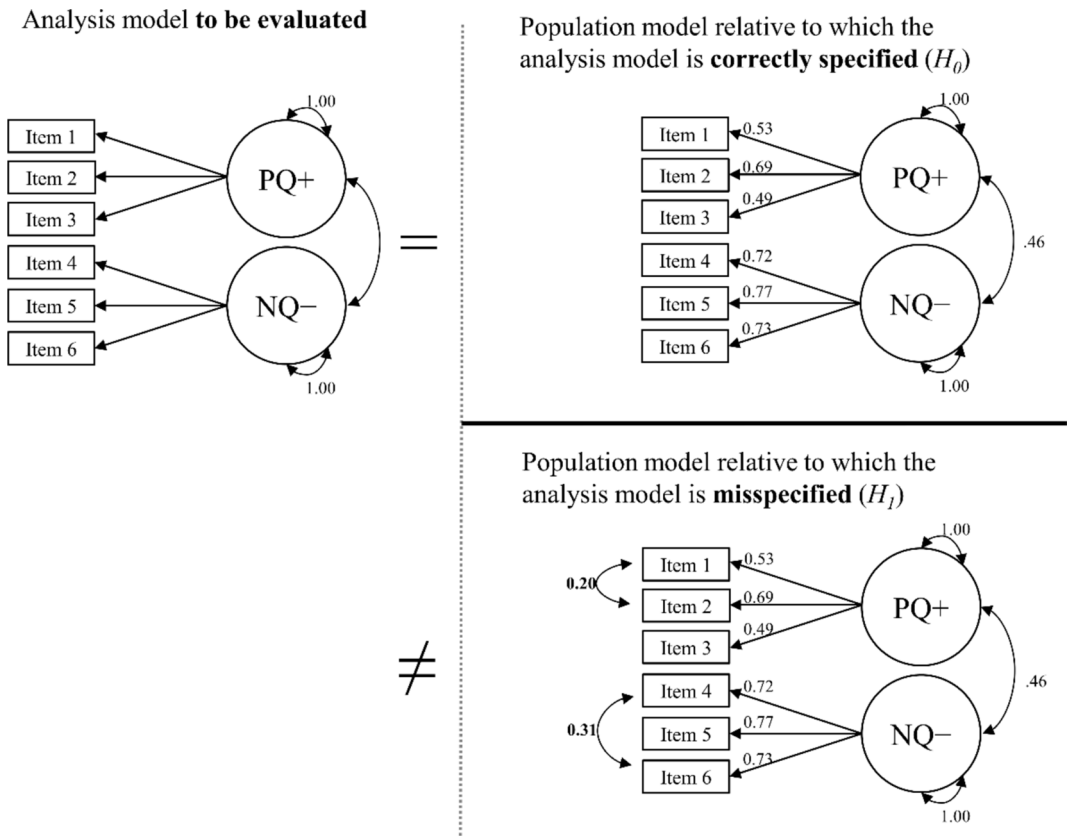


Fig. 8 Proposed analysis and population models of the Social Desirability-Gamma Short Scale. *Note.* Unstandardized coefficients. PQ+ = exaggerating positive qualities; NQ− = minimizing negative qualities. We recoded NQ− so that higher values imply more socially desirable responses. We omitted the residual variances and the mean structure for clarity

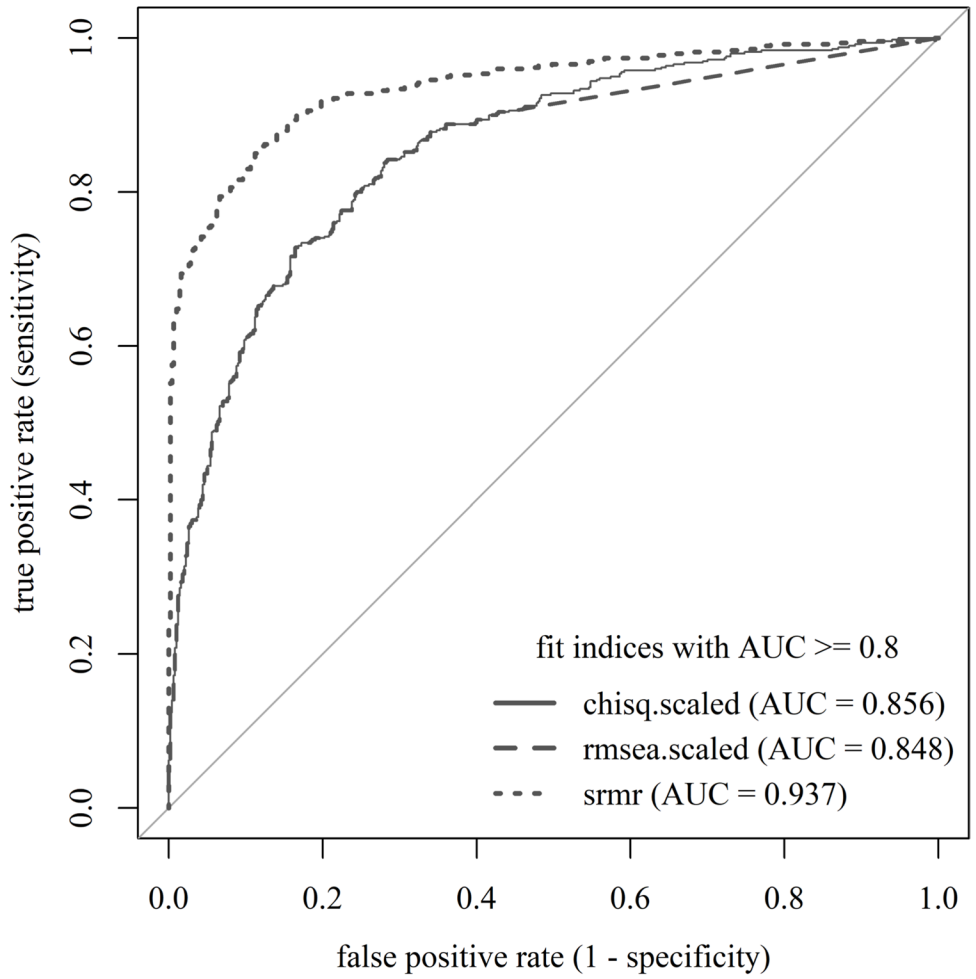


Fig. 9 ROC curves for fit indices with AUC ≥ .80 of the Social Desirability-Gamma Short Scale model. *Note.* Chisq.scaled is a χ^2 test statistic asymptotically equivalent to the robust Yuan–Bentler test statistic (Yuan & Bentler, 2000a) to account for non-normality. Rmsea.scaled is the RMSEA version calculated with this test statistic

categorized correctly specified models as correctly specified and misspecified models as misspecified than cutoffs for other fit indices. Thus, the SRMR, with its corresponding cutoff, had the best ability to demarcate between correctly specified and misspecified models in the scenario of interest. Further, it implies that the greatest difference between correctly specified and misspecified models in the specific scenario was due to average standardized residuals.

The reader may have noted that these cutoffs' type I and type II error rates are above conventional levels of 5% or 10%. 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 rates. Here, we deemed the error rates acceptable (especially the ones of SRMR) because we explicitly wanted to retain the definitions of population models as outlined and justified in this example. 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 might 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).

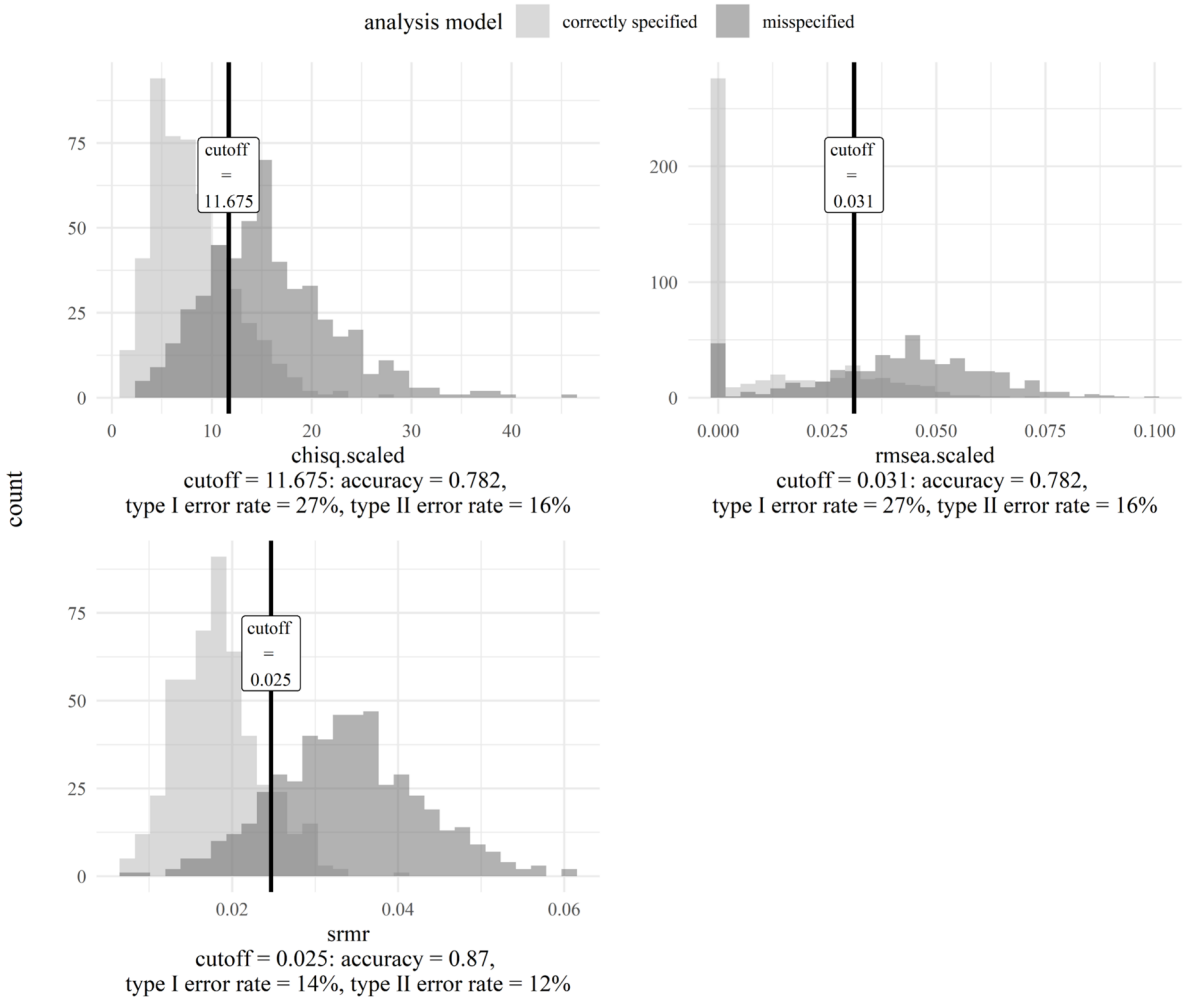


Fig. 10 Cutoffs for fit indices with AUC ≥ .80 of the Social Desirability-Gamma Short Scale model. *Note.* Chisq.scaled is a χ^2 test statistic asymptotically equivalent to the robust Yuan–Bentler test statistic (Yuan & Bentler, 2000a) to account for non-normality. Rmsea.scaled is the RMSEA version calculated 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 are colored in an even darker gray 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 rejected 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 all tailored cutoffs ($\chi^2(8) \leq 11.68$; CFI = should not be considered; RMSEA ≤ .031; SRMR ≤ .025). Thus, H_1 seemed more plausible than H_0, concluding that a model different from a two-factor one is likely to have generated the data.

Notably, fixed cutoffs of CFI around .950 and SRMR around .080 (but not RMSEA around .060; Hu & Bentler, 1999) were far off the tailored cutoffs and would wrongly lead to accepting the two-factor model. This underscores that fixed cutoffs would not have provided valid assessments of model fit in settings markedly different from the simulation scenarios they originated from (i.e., three-factor models with 15 items). Additionally, we knew that the SRMR was most decisive for decisions about model fit (if fit indices would disagree about model acceptance or rejection)—something that would remain unknown with fixed cutoffs for fit indices and other approaches to tailored cutoffs.

As we rejected the two-factor model, we must modify the model and test the modified model again. A modified model can be considered a new empirical setting, so testing that modified model requires a new set of tailored cutoffs. We demonstrated 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 Online Material. We made use of the different performance of fit indices in that example as their decisions on model fit disagreed.

Discussion

Fixed cutoffs for fit indices are far more problematic than many researchers realize (e.g., Groskurth et al., 2024; 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 misspecifications they should detect (for an overview, see Groskurth et al., 2024; Niemand & Mai, 2018; McNeish & Wolf, 2023a, b; 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, 2023a; Niemand & Mai, 2018; Nye & Drasgow, 2011).

The current article reviewed four principal approaches to generating tailored cutoffs. This is the first article to comprehensively review and synthesize the approaches to tailored cutoffs. 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 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 dynamic simulation approach to tailored cutoffs that has gained traction in recent literature (e.g., McNeish & Wolf, 2023a, b; 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 certain type I and type II error rates (i.e., balanced ones for the simulation-cum-ROC approach) for several fit indices across different settings. However, it conceptually advances previous approaches by (2) ranking the performance of fit indices (i.e., their ability to discriminate between correctly specified and misspecified models) for 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 on (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 intended 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 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).[12] 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 criticism. 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 conventional fit indices we used in the examples. Notably, the simulation-cum-ROC approach can be generalized to include any other fit index, 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,

[12] 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 both the χ^2 and the RMSEA in the examples because both are regularly used for model evaluation (Jackson et al., 2009).

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 for lower type I and type II error rates. Moreover, we selected a 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 might maximize sensitivity given a minimal specificity value to obtain optimal cutoffs (or vice versa).

We provided R code in Additional File 1 of the Supplementary Online Material and programmed a Shiny app available under https://kg11.shinyapps.io/tailoredcutoffs/ to ease the application of the simulation-cum-ROC approach. Executing the simulation-cum-ROC approach for our examples took two to three minutes on a standard computer using R (single-threaded).

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; Nye & Drasgow, 2011; McNeish & Wolf, 2023a).

Advanced definitions of population models

A challenge in applying the simulation-cum-ROC approach—one that it shares with similar dynamic simulation approaches (e.g., Pornprasertmanit, 2014)—concerns the definition of the H_0 and H_1 population models (cf. McNeish & Wolf, 2023a, b, who already predefined 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 an analysis model is negligibly underspecified (i.e., misspecified) to test for approximate fit, as suggested by Millsap (2007, 2013) and Pornprasertmanit (2014). We indeed believe that alternative definitions of population models can be fruitful, so 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 analysis models were always identical to the H_0 population models. We generated cutoffs based on an analysis model that exactly fits the data generated by (i.e., simulated from) that H_0 population model. Only sampling fluctuations influenced the resulting fit index distributions and, accordingly, the cutoffs (Cudeck & Henly, 1991; MacCallum, 2003; MacCallum & Tucker, 1991). Testing the assumption of exact fit has guided model evaluation for years; the entire distributional assumptions of the χ^2 test statistic rely on the test of exact fit (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 (more precisely, specific features, e.g., broad factors in models) that do not solely reproduce a specific population but are generalizable to different populations (a broad array of, e.g., demographic populations; Cudeck & Henly, 1991; Millsap, 2007; see also Wu & Browne, 2015). 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). In this context, model error 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 trivial 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 others (for an overview and more in-depth discussion, see MacCallum, 2003; see Wu & Browne, 2015, for the related concept of adventitious error that defines the differences between the sampled and theoretically hypothesized population). 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, 2023a; 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 such that one tests approximate (instead of exact) fit

in Additional File 2 of the Supplementary Online Material. As the example demonstrates, testing approximate fit via the simulation-cum-ROC approach works 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_0 / H_1 population model to test the fit of an analysis model of interest. As defined by 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_0 / H_1 population model), we take only one form of misspecification (e.g., omitted residual correlation of $r = .50$) out of all possible misspecifications in the space of conceivable models into account.

To ensure decisions about accepting or rejecting a model are generalizable to other forms of misspecification, we could, for example, repeatedly follow the fixed method and conduct so-called robustness checks. To conduct robustness checks, we define different forms of misspecification and derive new cutoffs for each of them. The degree of misspecification should roughly stay the same to compare the cutoffs' robustness across different forms of misspecification. These robustness checks investigate whether we will make the same decision about accepting or rejecting a model with different forms of misspecification. We included an example of a robustness check for the Social Desirability-Gamma Short Scale example in Additional File 2 of the Supplementary Online Material. In the Guidelines on Forms and Degrees of Misspecification section, we provide some guidance on defining forms and quantifying degrees of misspecification.

Pornprasertmanit (2014) proposed new methods that directly take a wide variety of misspecification forms into account (e.g., omitted residual correlations; omitted cross-loadings) when deriving a set of cutoffs (ideally at the same degree of misspecification). Like the robustness checks for the fixed method, the new methods apply to both H_0 population models (to test approximate fit) and H_1 population models. Recall that the H_0 population model implies trivial, acceptable misspecification (to test approximate fit), and the H_1 population model implies severe, unacceptable misspecification of the analysis model relative to that population model.

In the *random method*, one defines several H_0 / H_1 population models relative to a misspecified analysis model. The analysis model is trivially misspecified to H_0 population models and severely misspecified to H_1 population models. The algorithm randomly picks a new H_0 / H_1 population model from the several possible H_0 / H_1 population models (defined initially) each time it starts simulating data. This approach considers multiple H_0 / H_1 population models relative to a misspecified analysis model. The 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 a misspecified 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 principles and mechanisms of our simulation-cum-ROC approach), we encourage researchers to consider the random and maximal/minimal methods in future applications of the simulation-cum-ROC approach. 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.

Guidelines on forms and degrees of misspecification

In the previous section, we have outlined how to incorporate several forms of misspecification (ideally at the same degree of misspecification) into derivations of tailored cutoffs. Both different forms of the same degree and different degrees of the same form of misspecification influence the fit index performance and, thus, tailored cutoffs, as shown by several studies (e.g., Groskurth et al., 2024; McNeish & Wolf, 2023a; Moshagen & Erdfelder, 2016). Thus, the form and degree of misspecification are both relevant for deriving tailored cutoffs, so we want to guide researchers in defining the form and quantifying the degree of misspecification.

Similar to several other authors (e.g., Curran et al., 1996; Hu & Benter, 1998, 1999; McNeish & Wolf, 2023a; Millsap, 2013; Yuan & Bentler, 1997), the analysis models in our examples were (either trivially or severely) misspecified relative to the population models, as they either propose a different model structure or omit specific parameters of a particular size. Thereby, we have already shown different forms of misspecification: A single factor of an analysis model can be misspecified by splitting it into two factors in the population model (e.g., Rosenberg Self-Esteem Scale

example), an analysis model of at least two factors can be misspecified by adding cross-loadings[13] to the population model (e.g., robustness check in Additional File 2 of the Supplementary Online Material), and an analysis model can be misspecified by adding residual covariances to the population model (e.g., Social Desirability-Gamma Short Scale example).

To make the incorporated misspecification (independent of its form) more comparable and thus objective across different scenarios, we can quantify the degree of misspecification in an effect size logic (see also Moshagen & Auerswald, 2018).[14] For instance, we can quantify the degree of misspecification in terms of the non-centrality parameter (see Jak et al., 2021, who developed a Shiny app for this). The non-centrality parameter can then be transformed into a comparable effect size metric such as χ^2/df or RMSEA, both considered on the population level (Moshagen & Erdfelder, 2016). This effect size helps to quantify and compare degrees of misspecification within or across scenarios. We evaluated the degrees of misspecification induced in our examples in Additional File 3 of the Supplementary Online Material.

However, what constitutes a reasonable population model and a trivial, medium, or severe misspecification of the analysis model relative to that population model depends on many characteristics, such as the research question, study design, and empirical data. Researchers need to justify their definitions of population models based on those characteristics. By requiring that the population models be made explicit, editors, reviewers, and readers of the article can judge the appropriateness of the assumptions about the population model—we believe that this transparency is a major advantage of our simulation-cum-ROC approach.

Checklist

Overall, the simulation-cum-ROC approach applies to a broad range of empirical settings in which cutoffs must be tailored to the needs of the setting at hand. This also comes with a certain level of subjectivity; the researcher needs to make several decisions, for instance, on the definition of H_0 and H_1 population models. To guide researchers through this process, we have defined a checklist for evaluating an analysis model with tailored cutoffs using the simulation-cum-ROC approach. This checklist is based on the decisions and pathways outlined throughout this article and can be found in Additional File 4 of the Supplementary Online Material.

Conclusion

Tailored cutoffs are ideally suited to the empirical setting at hand because they account for the many model, estimation, and data characteristics that can influence fit indices and render fixed cutoffs questionable. This article reviewed four principal approaches researchers can employ to generate tailored cutoffs. We then presented a novel approach, the simulation-cum-ROC approach, that extends previous tailored cutoff approaches, more specifically the dynamic simulation ones, by introducing ROC analysis. Introducing ROC analysis to model fit evaluation is a contribution that uniquely characterizes our approach. It allows for evaluating the performance of fit indices in a given scenario, thus enabling researchers to make informed choices regarding the fit indices on which to rely (or to which to assign the greatest weight). 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. The simulation-cum-ROC approach can derive tailored cutoffs for any fit index that a researcher may want to use, including yet-to-be-developed ones. Our procedure to obtain tailored cutoffs 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 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 valid judgments about model fit and ultimately increase the replicability of research findings. By reviewing possible extensions of our approach, we hope to encourage methodologists to advance further—and help disseminate—the current approaches to generating tailored cutoffs (including our simulation-cum-ROC approach).

Acknowledgements We want to thank Hansjörg Plieninger for the initial idea and example, Thorsten Meiser for suggestions on conceptual questions and comments on earlier versions of the article, and Thomas Knopf for comments on earlier versions of the code. This work was supported by the University of Mannheim's Graduate School of Economic and Social Sciences.

Authors' contributions Katharina Groskurth: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Visualization.

[13] Whereas misspecification increases with a larger magnitude of unmodeled cross-loadings, adding more unmodeled cross-loadings to the population model leads to "lower" misspecification in the analysis model (if factors are correlated). Other parameters can compensate for the added cross-loadings (Groskurth et al., 2024).

[14] Some authors (Cudeck & Browne, 1992; Yuan et al., 2007) have proposed to directly define misspecification in terms of an effect size logic, without defining the particular form (or location) of misspecification. They define (or approximate) the population variance-covariance matrix at a given distance from the analysis model structure.

Nivedita Bhaktha: Conceptualization, Methodology, Software, Validation, Writing – Review & Editing, Visualization.

Clemens M. Lechner: Conceptualization, Methodology, Validation, Resources, Writing – Original Draft, Writing – Review & Editing, Visualization, Supervision.

Funding Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials We did not preregister the design and analysis of this article. The data (Nießen et al., 2018, 2020) to reproduce the analysis and results of this article is publicly available on the GESIS SowiDataNet | datorium repository (https://doi.org/10.7802/2080 and https://doi.org/10.7802/1752). We also programmed a Shiny app, which is available at https://kg11.shinyapps.io/tailoredcutoffs/. All Additional Files from the Supplementary Online Material of this article are available on the Open Science Framework (https://osf.io/vk94q/).

Code availability The code to reproduce this article's simulation, analysis, and results is available in Additional File 1 of the Supplementary Online Material.

Declarations

Competing interests We have no competing interests to disclose.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Prior dissemination We have published the article as a preprint on PsyArXiv (https://doi.org/10.31234/osf.io/62j89) and advertised it on ResearchGate and Twitter. We have also published a previous version of this article on MADOC (https://madoc.bib.uni-mannheim.de/64707/), an online repository of the University of Mannheim, as part of Katharina Groskurth's doctoral thesis. Further, we narratively presented prior versions of this article at three online conferences in 2021: (1) Meeting of the Working Group on Structural Equation Modelling (SEM) on March 18–19, 2021, (2) International Meeting of the Psychometric Society (IMPS) on July 19–23, 2021, and (3) the 15th Conference of the Section 'Methods and Evaluation' in the German Psychological Society (FGME) on September 15–17, 2021. We confirm that this article has not been externally presented, submitted, or published elsewhere.

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

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., Miklošić, I., & Halberstadt, J. (2016). Measuring cross-cultural supernatural beliefs with self- and peer-reports. *PLoS ONE, 11*(10), 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/10.1177/0049124192021002004

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

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford Publications.

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, R. B., Sr., Pencina, M. J., Massaro, J. M., & Coady, S. (2013). Cardiovascular disease risk assessment: Insights from Framingham. *Global Heart, 8*(1), 11–23. https://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

Grønneberg, S., Foldnes, N., & Marcoulides, K. M. (2022). covsim: An R package for simulating non-normal data for structural equation models using copulas. *Journal of Statistical Software, 102*(3), 1–45. https://doi.org/10.18637/jss.v102.i03

Groskurth, K., Bluemke, M., & Lechner, C. M. (2024). Why we need to abandon fixed cutoffs for goodness-of-fit indices: An extensive simulation and possible solutions. *Behavior Research Methods, 56*(4), 3891–3914. https://doi.org/10.3758/s13428-023-02193-3

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science, 16*(4), 789–802. 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, J. A., Jr., & 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. (2023). A tutorial on assessing statistical power and determining sample size for structural equation models. *Psychological Methods, 28*(1), 207–221. https://doi.org/10.1037/met0000423

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

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

Kemper, C. J., Beierlein, C., Bensch, D., Kovaleva, A., & Rammstedt, B. (2014). Soziale Erwünschtheit-Gamma (KSE-G) [Social Desirability-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

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

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*, 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. (2023). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods, 28*(1), 61–88. https://doi.org/10.1037/met0000425

McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for one-factor models. *Behavior Research Methods, 55*(3), 1157–1174. 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., & Bader, M. (2024). semPower: Power Analyses for SEM. R package version 2.1.1. Retrieved March 9, 2025, from https://cran.r-project.org/package=semPower

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

Muthén, L. K., & Muthén, B.O. (1998-2017). *Mplus User's Guide* (8th ed.). Muthén & Muthén. Retrieved January 29, 2021, from https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf

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://doi.org/10.1098/rsta.1933.0009

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) [Data set]. GESIS – Leibniz Institute for the Social Sciences. https://doi.org/10.7802/1752

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) [Data set]. GESIS – Leibniz Institute for the Social Sciences. https://doi.org/10.7802/2080

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*, 2. https://doi.org/10.1186/s42409-018-0005-1

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

Niemand, T., & Mai, R. (2025). FCO: Flexible cutoffs for model fit evaluation in covariance-based structural models. R package version 0.8.0. Retrieved March 9, 2025, from https://cran.r-project.org/package=FCO

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/10.1177/0013164414548889

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*, 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

Pornprasertmanit, S. (2014). *The unified approach for model evaluation in structural equation modeling* [Unpublished doctoral dissertation]. University of Kansas. Retrieved August 31, 2021, from 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

Pornprasertmanit, S., Miller, P., Schoemann, A., & Jorgensen, T. D. (2021). *simsem: SIMulated structural equation modeling*. R package version 0.5.16. Retrieved November 12, 2024, from https://CRAN.R-project.org/package=simsem

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved November 12, 2024, from 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.

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

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*, 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. https://doi.org/10.18637/jss.v048.i02

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. https://doi.org/10.23668/psycharchives.12784

Schmalbach, B., Irmer, J. P., & Schultze, M. (2019). *ezCutoffs: Fit measure cutoffs in SEM*. R package version 1.0.1. Retrieved November 12, 2024, from 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

Wolf, M. G., & McNeish, D. (2022). dynamic: DFI cutoffs for latent variable models. R package version 1.1.0. Retrieved March 9, 2025, from https://cran.rproject.org/package=dynamic

Wu, H., & Browne, M. W. (2015). Quantifying adventitious error in a covariance structure as a random effect. *Psychometrika, 80*(3), 571–600. https://doi.org/10.1007/s11336-015-9451-3

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

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.

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. (2000). 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. (2000). 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., 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

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

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

Open practices statement We did not preregister the design and analysis of this article. The code to reproduce this article's simulation, analysis, and results is available in Additional File 1 of the Supplementary Online Material. The data (Nießen et al., 2018, 2020) to reproduce the analysis and results of this article is publicly available in the GESIS SowiDataNet | datorium repository (https://doi.org/10.7802/2080 and https://doi.org/10.7802/1752). We also programmed a Shiny app, which is available at https://kg11.shinyapps.io/tailoredcutoffs/. All Additional Files from the Supplementary Online Material of this article are available on the Open Science Framework (https://osf.io/vk94q/).