AMERICAN PSYCHOLOGICAL ASSOCIATION

American Psychologist

© 2023 American Psychological Association
ISSN: 0003-066X

2023, Vol. 78, No. 9, 1061–1075
https://doi.org/10.1037/amp0001213

Dynamic Fit Index Cutoffs for Categorical Factor Analysis With Likert-Type, Ordinal, or Binary Responses

Daniel McNeish
Department of Psychology, Arizona State University

Scale validation is vital to psychological research because it ensures that scores from measurement scales represent the intended construct. Fit indices are commonly used to provide quantitative evidence that a proposed factor structure is plausible. However, there is a mismatch between guidelines for evaluating fit of the factor models and the data that most researchers have. Namely, fit guidelines are based on the simulations that assume item responses are collected on a continuous scale whereas most researchers collect discrete responses such as with a Likert-type scale. In this article, we show that common guidelines derived from assuming continuous responses (e.g., root-mean-square error of approximation < 0.06, comparative fit index > 0.95) do not generalize to factor models applied to discrete responses. Specifically, discrete responses provide less information than continuous responses, so less information about misfit is passed to fit indices. Traditional guidelines, therefore, end up being too lenient and lose their ability to identify that a model may have a poor fit. We provide one possible solution by extending the recently developed dynamic fit index framework to accommodate discrete responses common in psychology. We conduct a simulation study to provide evidence that the proposed method consistently distinguishes between well-fitting and poorly fitting models. Results showed that our proposed cutoffs maintained at least 90% sensitivity to misspecification across studied conditions, whereas traditional cutoffs were highly inconsistent and frequently exhibited sensitivity below 50%. The proposed method is included in the `dynamic` R package and as a web-based Shiny application to make it easily accessible to psychologists.

Public Significance Statement
Conclusions in many subareas of psychology are reliant on conclusions drawn from statistical models, which rely on assumptions. This article discusses how the assumptions made by statistical guidelines in the key area of scale validation differ from the data that most psychologists collect, at best making it difficult to evaluate the evidence supporting psychological scales and, at worst, increasing the probability that conclusions in many studies may be inaccurate. This directly feeds into replication issues observed in psychology because scale scores may not be accurately measuring psychological traits, attitudes, or abilities, which increases noise in the data and reduces the validity of scores. This article proposes alternative guidelines for one common type of statistical model by making assumptions more closely match the characteristics of the data to which scale validation models are applied.

Keywords: psychometrics, scale validation, CFA, categorical data analysis, SEM

Supplemental materials: https://doi.org/10.1037/amp0001213.supp

Editor's Note. Daniel McNeish received the 2023 APA Award for Distinguished Scientific Early Career Contributions to Psychology. In association with the award, McNeish was invited to submit a manuscript to *American Psychologist*, which was peer reviewed. The article is published as part of the journal's annual Awards Issue.

Daniel McNeish https://orcid.org/0000-0003-1643-9408

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through R305D220003 to Daniel McNeish, Arizona State University. The opinions expressed are those of the authors and do not represent the views of the institute or the U.S. Department of Education.

The author was indebted to Melissa Wolf, who built much of the software architecture that allowed these statistical methods to be generalized and implemented at a large scale. She has since moved on from academia and did not have time to devote to ideas in this article specifically, but she was instrumental in the foundational work that allowed ideas presented in this article to develop. Additional online materials can be viewed on the Open Science Framework page associated with this project at https://osf.io/2t4zh/.

Daniel McNeish played a lead role in conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, validation, visualization, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to Daniel McNeish, Department of Psychology, Arizona State University, P.O. Box 871104, Tempe, AZ 85287, United States. Email: dmcneish@asu.edu

Measurement is a central part of research in psychology because many focal variables are constructs that cannot be physically measured (Nunnally, 1967). For instance, there is broad consensus that the construct "depression" exists, but there are no (currently available) physical instruments that can reliably and efficiently scan a person's brain to gauge the magnitude of their depression. Instead, items representing different aspects of depression are administered and—if the items reasonably capture relevant features of depression—a person's responses can be converted to numeric values to create scores to represent and compare individual differences on the physically unmeasurable "depression" construct (Borsboom, 2005). The same is true of other psychological constructs like "motivation," "intelligence," or "anxiety."

Measurement is essential to drawing accurate, reliable, and replicable conclusions in areas of psychology that heavily rely on scales because conclusions are only as trustworthy as scores that precede them (Flake & Fried, 2020). Poor measurement is an underappreciated source of replication issues (e.g., Loken & Gelman, 2017), particularly because replication studies often take the measurement at face value and replicate the protocol exactly without verifying that the initial measures were valid (Flake et al., 2022; Shaw et al., 2020). The downstream effects of poor measurement can therefore be significant because scaled scores are the foundational building block of any statistical analysis in many subareas of psychology (McNeish, 2022), and no amount of sophisticated statistical modeling can work around deficient measurement (Flake, 2021).

The main tenets of psychological measurement (also referred to as psychometrics) are *reliability* and *validity* (American Educational Research Association et al., 2014). Reliability quantifies (lack of) random error in scores and serves as a measure of precision or consistency (e.g., Bandalos, 2018; Furr, 2021). Validity is the degree to which evidence supports the interpretation of scores, which essentially means that the scores are an accurate measure of the intended construct, for a particular population (e.g., Borsboom et al., 2004; Slaney, 2017). Psychometric evidence is vital to psychology research because it establishes a connection between a score based on a combination of item responses and the manifestation of an unobserved construct, which permits interpretation of scale scores as approximations of physically unobservable constructs like "depression." Without this evidence, substituting scale scores for psychological constructs and interpreting them as synonymous with the construct in statistical analyses can be problematic.

Though there are many types of validity (Messick, 1989), *internal structure* is the most commonly reported in psychology, likely because it is among the easiest to quantify (Slaney, 2017). Internal structure evidence supports that the number of constructs underlying the item responses is consistent with how many constructs were intended and assesses whether the items are consistent with the intended construct. This provides evidence (but does not necessarily confirm) that a set of items reasonably measure a single construct and that it makes sense to combine items in some way to create an interpretable score for that construct.

Evidence of internal structure is typically provided by evaluating model fit from a factor analysis. We discuss different approaches shortly, but the approach that remains most popular (e.g., Zyphur et al., 2023) is comparing fit indices to traditional cutoffs established by Hu and Bentler (1999), such as root-mean-square error of approximation (RMSEA) $\leq$ 0.06, standardized root-mean-square residual (SRMR) $\leq$ 0.08, and comparative fit index (CFI) $\geq$ 0.95. The sizeable influence that these cutoffs have had on the scale validation literature (and on psychological research more broadly) is demonstrated by Hu and Bentler's article being cited over 100,000 times. These cutoffs are foundational to many subareas of psychology because these cutoffs are used to argue for validity of measurement scales and make claims that scale scores are suitably representing the construct of interest. Even if researchers are not performing psychometric work themselves, they are likely relying on or citing validation work done by others that directly cites or—at the very least—is guided by these cutoffs (Jackson et al., 2009).

Despite the wide adoption of these traditional cutoffs and their canonical status, the methodological literature has identified situations in which their performance deteriorates (e.g., Greiff & Heene, 2017; Heene et al., 2011; Marsh et al., 2004). The cutoffs were designed to maximize the probability of rejecting a misspecified model while minimizing the probability of rejecting an acceptable model. However, the traditional cutoffs can lose these optimal properties in some contexts in which they are commonly applied, including one-factor models (McNeish & Wolf, 2023a) or models for Likert-type responses (Nye & Drasgow, 2011).

In this article, we focus on issues of applying traditional cutoffs to binary, ordinal, or Likert-type item responses. The main problem is that the traditional cutoffs were derived assuming responses were collected on a continuous scale, whereas most psychological studies collect responses on a discrete scale (Flake et al., 2017; Jackson et al., 2009). Given that no dedicated cutoffs have been determined for factor models with discrete data, most psychologists continue to rely on the traditional cutoffs (Xia & Yang, 2019) even though the traditional cutoffs have been shown to perform suboptimally (or downright poorly) when applied to discrete data (e.g., Beauducel & Herzberg, 2006; Monroe & Cai, 2015; Xia & Yang, 2018). The potential consequence is that many psychology studies rely on inaccurate guidelines when evaluating the quality of their scores and measurement scales. This has implications for trustworthiness of conclusions (Borsboom et al., 2004; Crutzen & Peters, 2017), understanding mechanisms of thought and behavior (Yarkoni, 2022), and replication (Flake, 2021; Flake et al., 2022).

To address this issue, the goal of this article is to introduce a method to derive fit index cutoffs that are specifically suited for discrete data that psychologists typically collect. To outline the remainder of the article, we start by discussing some basic details about fit indices and the history of traditional cutoffs. We then review the methodological literature discussing issues with traditional cutoffs and how they do not always maintain optimal properties in all contexts, especially discrete item responses. The recently proposed dynamic fit index (DFI) framework for deriving more appropriate cutoff values is then reviewed. Limitations of the current implementation of DFI are discussed, including the fact that—much like Hu and Bentler (1999)—the method currently assumes continuous item responses. We propose an extension to DFI that would allow discrete responses to be accommodated such that dedicated cutoffs for discrete data could be derived. We provide a simulation study to demonstrate the effectiveness of our proposed discrete extension of DFI, both in absolute terms and relative to traditional cutoffs. We conclude with a discussion of take-home points, limitations, and extensions.

Origin of Traditional Fit Index Cutoffs

In classic regression-based models, model fit is defined as variance explained by predictors. However, model fit in factor analysis is not about explained variance. Instead, if a proposed factor model is a plausible explanation for the process underlying the data, the covariance matrix implied by the model should look similar to the covariance matrix of the observed data (Tomarken & Waller, 2003). Therefore, factor analytic fit metrics are based on quantifying the discrepancy between the model-implied and observed covariance matrices.

This comparison can be *local* or *global*. Local fit metrics assess how closely each individual element of the covariance matrix was reproduced (Browne et al., 2002; McDonald & Ho, 2002), which can be useful for identifying specific locations where misspecification may be present. Local fit is much less commonly reported (e.g., a review by Ropovik, 2015 reports only 3% of studies to use such metrics; a review by M. F. Zhang et al., 2021 reports 17%). Global fit summarizes the total discrepancy throughout the entire covariance matrix with a single value. This article focuses on metrics to assess global fit, but local fit is less informative and local fit is a recommended component of any comprehensive fit assessment (Appelbaum et al., 2018; West et al., 2023).

An early approach to evaluate global fit placed the discrepancy within the null hypothesis testing framework (Jöreskog, 1969). The null hypothesis states that the model-implied covariance matrix equals the observed covariance matrix. This test is classified as an *exact fit* test because the focus is *equality* of the observed and model-implied matrices. This test is commonly referred to as the χ^2 test because the test statistic that quantifies the discrepancy between matrices asymptotically follows a χ^2 distribution. The χ^2 test is praised for providing a clear hypothesis with an unambiguous interpretation (Barrett, 2007). However, others have criticized the χ^2 test because its null hypothesis may not necessarily be of interest. As Browne and Cudeck (1993) noted,

> In applications of the analysis of covariance structures in the social sciences it is implausible that any model that we use is anything more than an approximation to reality. Since a null hypothesis that a model fits exactly in some population is known a priori to be false, it seems pointless even to try and test whether it is true. (p. 137)

Rejecting the null hypothesis of exact fit may be triggered by small discrepancies or modest departures from normality when sample sizes are large (Hu et al., 1992; Tanaka, 1987) and may not necessarily speak to the model's practical utility (Ropovik, 2015). Consequently, the χ^2 test is commonly reported in empirical studies (Jackson et al., 2009), but it rarely is the primary piece of evidence and it is almost universally supported with other fit metrics (West et al., 2023).

As an alternative, *approximate* fit indices have been proposed such as SRMR, RMSEA, and CFI. Whereas the χ^2 test focuses on the *presence* of any discrepancy between the model-implied and observed covariance matrices, approximate fit indices focus on the *magnitude* of the discrepancy to allow researchers to evaluate the extent of misfit. As put by Millsap (2007),

> Given that a proposed model does not provide an exact fit, an approximate fit index will summarize the degree of misfit. The tacit rationale for such indices is that the degree of misfit is relevant information when deciding whether the model is scientifically useful. (p. 876)

However, approximate fit indices do not have innate values that indicate misfit has reached an unreasonable level. That is, a frequent question emerging from the use of approximate fit indices is, "How much misfit is too much?" (Tanaka, 1987). For instance, does an RMSEA of 0.03 indicate a good fit? How about 0.05? Or 0.08? At which values of RMSEA should we be concerned?

Which Fit Index Values Indicate "Good" Fit?

Initially, values indicating a "good" fit were based on heuristics or personal experience. For instance, when discussing the reasonable values of RMSEA, Browne and Cudeck (1993) suggested,

> Practical experience has made us feel that a value of the RMSEA of about 0.05 or less would indicate a close fit of the model. … We are also of the opinion that a value of about 0.08 or less for the RMSEA would indicate a reasonable error of approximation and would not want to employ a model with an RMSEA greater than 0.1. (p. 144)

Similarly, when discussing reasonable values of CFI, Bentler and Bonett (1980) wrote

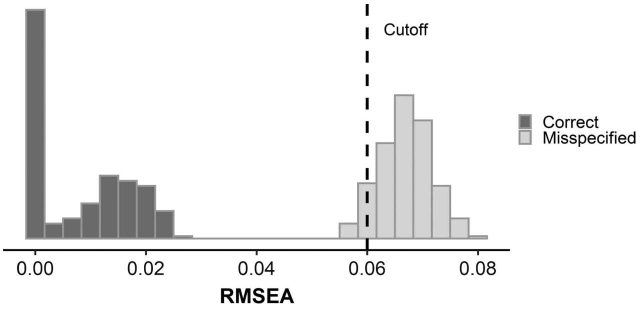
> Experience will be required to establish values of the indices that are associated with various degrees of meaningfulness of results. In our experience, models with over fit indices of less than .9 can usually be improved substantially. (p. 600)

Hu and Bentler (1999; hereinafter, referred to as HB) noted that the adequacy of heuristic benchmarks had not been formally evaluated up to 1999 (p. 4) and sought to empirically evaluate commonly suggested benchmarks. HB conducted a large simulation study where they fit models that they knew were either correct or misspecified to many simulated data sets. They then explored distributions of the resultant fit indices to determine which values of which indices could reliably differentiate between acceptable and misspecified models.

Figure 1 demonstrates HB's process for RMSEA. The dark grey distribution shows RMSEA values from 200 replications of a model known to be consistent with simulated data and the light grey distribution shows RMSEA values from 200 replications of a model known to be inconsistent with simulated data (a cross-loading is omitted). The cutoff is established by the RMSEA value that optimally distinguishes between the two distributions. In this case, RMSEA ≤ 0.06 rejects essentially all misspecified models in light grey while retaining all of the acceptable models in dark grey. Results from this simulation led to now-prominent cutoffs, which—to this day—remain the most common guidelines for evaluating fit (Zyphur et al., 2023).

Figure 1

Demonstration of How Traditional Cutoffs Were Derived in Hu and Bentler (1999)



Note. The dark gray distribution represents RMSEA values from 200 simulation replications, where the fitted model is known to be consistent with the data. The light gray distribution represents RMSEA values from 200 simulation replications where the fitted model is known to be inconsistent with the data because a nonnegligible cross-loading has been omitted from the model. The cutoffs are then derived by determining the value of the fit index that is sensitive enough to correctly classify most misspecified replications as poorly fitting without incorrectly classifying correct models as poorly fitting. RMSEA = root-mean-square error of approximation.

Criticism of Traditional Fit Index Cutoffs

HB did not intend for their cutoffs to become a panacea and were aware that these cutoffs may not generalize broadly (e.g., Hu & Bentler, 1998, pp. 446, 450; Hu & Bentler, 1999, p. 27). Nonetheless, empirical researchers were elated to have empirically based cutoffs to cite as evidence of good fit and disregarded HB's cautions about potentially limited generalizability. Traditional cutoffs soon became "golden rules" enforced by reviewers and editors. Correspondingly, methodological research began to examine issues with overgeneralizing traditional cutoffs shortly after they appeared in the literature (e.g., Marsh et al., 2004).

Primarily, traditional cutoffs were derived from simulation and therefore only generalize to the conditions studied by HB (Paxton et al., 2001), which includes models with three factors, 15 items, and standardized loadings between 0.70 and 0.80. Recent studies have noted that traditional cutoffs are not stable and replicating HB's study with different conditions for the number of items, number of factors, standardized loading magnitude, model type, amount of missing data, or misspecification type results in different cutoffs being derived (Browne et al., 2002; Chen et al., 2008; Davey, 2005; Fan & Sivo, 2007; Greiff & Heene, 2017; Hancock & Mueller, 2011; Heene et al., 2011; Miles & Shevlin, 2007; Savalei, 2012; Shi et al., 2019; Sivo et al., 2006; X. Zhang & Savalei, 2020). For instance, McNeish et al. (2018) showed that replicating HB's simulation with different factor loading strengths produces RMSEA cutoffs as low as 0.04 or as high as 0.20. Relating this to Figure 1, the central tendency and variability of the light and dark grey distributions shift depending on model characteristics, causing the cutoff with optimal properties to shift as a function of model and data characteristics.

The root of these criticisms is that HB's simulation was essentially a power analysis. Traditionally, a power analysis strives to determine the sample size at which a statistical test is sensitive to the presence of an effect with some probability (usually 80% or above). Analogously, the goal of HB's simulation was to determine fit index values that are sensitive to the presence of misspecification with some probability (90% or 95%). The essential idea is the same—to determine the value of a target quantity that optimizes sensitivity to the presence of a phenomena. In power analyses, the target quantity is sample size and the phenomenon is (usually) an effect of a predictor or group assignment. In HB, the target quantity is a fit index value and the phenomenon is model misspecification. Just as the sufficient sample size (the target quantity) with optimal sensitivity to a nonnull effect in a power analysis is affected by model characteristics, so too are values of fit indices (the target quantity) with optimal sensitivity to misspecification. Just as there is no single efficient sample size ensuring adequate power in all designs, there is no universal fit index value ensuring sensitivity to misspecification in all models.

Binary, Ordinal, and Likert-Type Item Responses

HB's simulation and much of the subsequent methodological work on fit index cutoffs have focused on continuous responses. However, in practice, most research using measurement scales collects discrete responses, often with an ordinal Likert-type scale (Jackson et al., 2009, p. 18). For instance, a review by Flake et al. (2017) reports that 81% of scales in social and personality psychology solicit Likert-type responses. Fit indices behave differently with discrete responses, partly because diagonally weighted least squares (DWLS) or unweighted least squares (ULS) estimators are used to accommodate discrete data (Muthén & Kaplan, 1985) rather than maximum likelihood estimation common with continuous data (Bollen, 1989, pp. 131–134).

Several simulation studies note that traditional cutoffs do not perform well with discrete data (Beauducel & Herzberg, 2006; Monroe & Cai, 2015; Nye & Drasgow, 2011; Savalei, 2021; Shi et al., 2020; Xia & Yang, 2018, 2019). These studies find that—in addition to characteristics noted with continuous data—fit index cutoffs change as a function of the estimator, the number of categories, and the response distribution (e.g., all categories are endorsed evenly vs. most responses falling in a single category). These studies also point out that fit indices with DWLS and ULS are consistently attenuated toward perfect fit (i.e., RMSEA toward 0, CFI toward 1) such that model fit is overly optimistic relative to traditional cutoffs. As a succinct summary of research on evaluating fit of factor models with discrete responses, Xia and Yang (2019) wrote

> Applying the conventional cutoffs to ULS and DWLS can lead in the long run to the accumulation of models with severe misfit that are nonetheless considered acceptable … [fit indices] all appear to be insensitive to model misspecification if Hu and Bentler's cutoff values are applied. (pp. 420–421)

There is therefore a clear—but underappreciated—disconnect between the types of data that psychologists have and the types of data on which methodological recommendations are based. Traditional cutoffs intended for continuous data do not maintain their desirable properties when applied to the discrete data that most researchers possess. The ramification is that researchers are not receiving accurate information about the adequacy of the scales and scores that they use in their analyses, which could undermine conclusions from many studies because it is unclear if scale scores actually represent their intended psychological constructs.

Despite methodological research demonstrating notable issues of applying traditional cutoffs to models with discrete responses, there has been limited guidance on what psychologists should use instead in these circumstances. This perpetuates suboptimal methodological practices because methodologists know the status quo is not great, but there are not many viable alternatives to recommend instead. The next section overviews ideas about alternative methods and discusses how we propose to specifically address problems with discrete responses.

Ideas to Improve Upon Traditional Cutoffs

Despite the abundance of cautions and criticisms of traditional cutoffs, a review of reporting practices by Jackson et al. (2009) noted that they "did not find evidence that warnings about strict adherence to Hu and Bentler's suggestions were being heeded" (p. 18). A more recent review by Zyphur et al. (2023) notes that traditional cutoffs remain the most popular fit index strategy (p. 504). Despite several articulate criticisms of traditional cutoffs, practice has scarcely changed because there have been few proposed alternatives. So, even if psychologists are aware of these criticisms and want to avoid relying on traditional cutoffs, it is unclear what alternative guidance to follow besides returning to heuristics and personal experience. The challenge in this area of research is not to continue identifying weaknesses of traditional cutoffs but instead to devise alternatives that avoid or minimize weaknesses that have already been identified.

Millsap (2007, 2013) was at forefront of this line of thinking and suggested that researchers conduct customized simulations to derive fit index cutoffs with optimal properties for the specific characteristics of the model being evaluated (Kim & Millsap, 2014; Pornprasertmanit et al., 2013). This coincides with thinking about derivation of cutoffs in HB as a power analysis. In power analyses for sample size planning, an analysis is conducted for each individual set of circumstances and there is no one-size-fits-all sample size, where all studies are considered sufficiently powered. For instance, $N = 100$ may be sufficient (but inefficient) for within-subjects designs targeting a large effect but woefully insufficient for between-subjects designs targeting a small effect. However, current fit index interpretation follows a one-size-fits-all approach, whereby a single cutoff (e.g., RMSEA $\leq$ 0.06) from one set of conditions from one simulation is indiscriminately applied. RMSEA $\leq$ 0.06 was the optimal cutoff for the conditions studied by HB, but RMSEA $\leq$ 0.06 can have poor sensitivity to misfit in models with different characteristics.

Millsap's idea was to make fit index cutoffs more closely resemble power analysis such that cutoffs are rederived to ensure optimal classification properties for each model that is evaluated. The premise is that the logic of HB's simulation was reasonable, but the primary issue is generalizability. That is, HB identified cutoffs with optimal properties for the narrow conditions they studied, so researchers could conduct a simulation with conditions based on their model's characteristics to derive improved cutoffs for their model. This removes concerns about inappropriate generalizations because cutoffs are custom-tailored to the model being evaluated.

Millsap's (2007) idea is conceptually alluring but has not been widely adopted given some practical challenges. First, the approach requires Monte Carlo simulation, but many nonstatisticians do not possess expertise in this approach (Arend & Schäfer, 2019; Green & MacLeod, 2016). Second, it takes time to program a Monte Carlo simulation from scratch, so even researchers experienced with simulation methods may find the approach prohibitively time consuming. Third, custom simulation requires researchers to identify a meaningful hypothetical misspecification to which the cutoffs should be sensitive (e.g., the RMSEA cutoff in HB is designed to be sensitive to an omitted cross-loading). This is similar to how researchers must choose an anticipated effect size in a power analysis. However, the process is more complex in a multivariate model like a factor analysis and some researchers may have trouble articulating what they consider a "meaningful" misspecification to which cutoffs should be sensitive.

In summary, custom simulation is an appealing alternative to traditional cutoffs based on analogies to power analysis. Nonetheless, the potential of custom simulation will not be realized until it is more accessible. Analogously, power analysis flourished after programs like G*Power (Erdfelder et al., 1996) lowered barriers to implementation. The DFI approach (McNeish & Wolf, 2023b) and associated software (Wolf & McNeish, 2022) aim to lower barriers to implementing custom simulation approaches to improve accessibility for psychologists, essentially aspiring to become the G*Power of factor model fit. The thought is that if custom simulation were simpler to apply, the method would become a more serious alternative to traditional cutoffs.

Dynamic Fit Indices for Continuous Item Responses

DFI automates the two most challenging hurdles of applying custom simulation methods: defining hypothetical misspecifications to which cutoffs will be sensitive and writing Monte Carlo simulation code. This addresses weaknesses of previous instantiations of this idea by (a) lowering barriers for researchers without programing or Monte Carlo experience, (b) reducing time commitments for those who do have Monte Carlo experience, and (c) internally determining hypothetical misspecifications to relieve researchers of this difficult task. The method can be implemented through the `dynamic` R package (Wolf & McNeish, 2022) or a free web-based, point-and-click Shiny application (Wolf & McNeish, 2021; https://www.dynamicfit.app). A high-level overview focusing on software implementation is available from Wolf and McNeish (2023).

As a general overview of how DFI works, the core idea is that the user's model is the basis for the population model in a simulation. First, data sets are generated from the covariance matrix implied by the user's model—assuming multivariate normality—and the user's model is then fit to all simulated data sets. Fit indices are recorded, forming the equivalent of the dark grey distribution in Figure 1 but tailored to the specific characteristics of the user's model.

Figure 1 also requires a light grey distribution, which is more involved to derive because a hypothetical misspecification to which the cutoffs should be sensitive must be selected. In DFI, an algorithm searches for hypothetical paths (e.g., cross-loadings or residual correlations) and magnitudes of those paths that—if added—would render the fitted model misspecified to a similar magnitude as models considered in HB. For complete details about this algorithmic process, readers are referred to McNeish and Wolf (2023a, 2023b). Once these additional paths have been determined, data sets are simulated—assuming multivariate normality—from the covariance matrix implied by the augmented model consisting of the user's model plus additional hypothetical paths. The user's original model is then fit to these simulated data sets and the fit indices are recorded, forming the light grey distribution in Figure 1—tailored to the model's characteristics—because the user's model is known to be inconsistent with the simulated data.

Once a distribution of fit indices has been determined assuming the user's model is correct and assuming the user's model is misspecified, the same decision rule shown in Figure 1 is applied to determine a customized cutoff for the user's model. The default DFI cutoff decision rule is that 95% of replications from the misspecified model distribution should be rejected while rejecting no more than 5% of replications from the correct model distribution. It is possible that no DFI cutoff may exist if fit indices from the correct and misspecified model distributions are not well differentiated. This occurs when the distributions greatly overlap such that values from the light and dark grey distributions are similar, meaning that the same fit index value could plausibly originate from either distribution. The unavailability of DFI cutoffs is most prevalent in contexts with high sampling variability such as small samples, weak loadings, or few items.

If DFI cutoffs are unavailable, required sensitivity could be reduced (e.g., from 95% to 80%). Alternatively, unavailable DFI cutoffs may indicate that the model and data characteristics may not be amenable to accurately judging global fit. Returning to the power analysis analogy, unavailable DFI cutoffs correspond to conducting an analysis with low power. If power is low, failing to reject the null hypothesis may inconclusively indicate either a null effect or insufficient ability to detect a nonnull effect. Unavailability of DFI cutoffs similarly indicates inconclusive results because the fit of correct models cannot be differentiated from the fit of incorrect models with the researcher's model characteristics. Exact fit tests or local fit assessment (e.g., examining the standardized residual correlation matrix) may be more helpful in these circumstances.

As a sensitivity analysis, this process can also be repeated with increasingly severe misspecifications to derive cutoffs that are sensitive to varying magnitudes of misfit, which are referred to as *levels* in DFI. This is similar to power analysis, where the anticipated effect size can be changed to reflect different expectations about the magnitudes of effects. In DFI, "level" roughly refers to a multiple of the misspecification used in HB such that "Level-1" cutoffs correspond to the misspecification magnitude used in HB, "Level-2" cutoffs will derive a cutoff sensitive to a misspecification about twice as large as the one used by HB, etc. DFI software provides three levels by default for one-factor models, whereas the number of default levels in multifactor models depends on the size of the model.

In summary, DFI generalizes HB's logic by leveraging modern computational advances. Rather than fixed cutoffs coming from fixed simulation conditions decided upon by HB because simulations were far more computationally intensive 25 years ago, DFI dynamically adapts the simulation conditions to expediently replicate HB with new conditions. This creates distributions similar to those shown in Figure 1, but the central tendency and variability of the distributions will dynamically change to reflect the characteristics of the model being evaluated. As a result, the vertical line in Figure 1 shifts left or right to identify the fit index value that optimally separates the correct and misspecified models in the user's specific data and model characteristics.

DFI Advantages and Disadvantages

Studies evaluating DFI find that it unambiguously outperforms traditional cutoffs across a range of conditions with continuous item responses. McNeish and Wolf (2023a) evaluated performance of DFI and traditional cutoffs to identify that a one-factor model was inconsistent with multidimensional simulated data. Across various sample sizes, factor loading magnitude, and scale length conditions; DFI cutoffs were consistently sensitive to misspecification and never classified fewer than 90% of replications correctly for either SRMR, RMSEA, or CFI ($M = 96\%$, $SD = 3.1\%$). In the same conditions, average correct classification rates for traditional cutoffs were 0% for SRMR ($SD = 0.3\%$), 69% for RMSEA ($SD = 3.8\%$), and 42% for CFI ($SD = 39.2\%$). These results show consequences of indiscriminately generalizing traditional cutoffs, which often erroneously suggested that one-factor models were often appropriate for multidimensional data while also highlighting potential improvements of customized cutoffs with DFI.

McNeish (2023) found similar performance in the context of the multifactor models when evaluating performance of RMSEA to identify that 25% of items had an omitted 0.30 standardized cross-loading. For three-factor models with various sample sizes, factor loading magnitudes, and scale lengths; DFI cutoffs for RMSEA correctly classified 98% ($SD = 2.1\%$) of misspecified replications, whereas traditional cutoffs for RMSEA correctly classified 69% ($SD = 40.7\%$). In addition to the 29% improvement in average classification, also note the difference in the standard deviation of the classification rates between DFI and traditional cutoffs. DFI is actively adapting to specific model characteristics, so it maintains consistent performance across conditions ($SD = 2.1\%$). Conversely, traditional cutoffs are highly variable ($SD = 40.7\%$), and sensitivity to misfit depends on how closely the model characteristics match the characteristics studied by HB.

Though early appraisals of DFI have been promising, a key detail is that—currently—DFI simulations generate data from a multivariate normal distribution. Just like HB, this assumes that item responses are continuous, even though most psychological data contain discrete responses, which limits the applicability of DFI cutoffs and perpetuates the methodological literature's narrow focus on continuous responses despite most real data being discrete.

In the next section, we propose an extension to DFI that changes how data are simulated so that the resulting DFI cutoffs are applicable to discrete data. Whereas the current version of DFI is intended to produce cutoffs that are tailored to the user's model characteristics, this extension will produce DFI cutoffs that are tailored to the user's model *and* data characteristics.

Accommodating Discrete Responses With Dynamic Fit Index Cutoffs

One important aspect of DWLS or ULS for discrete data is that they are *limited information* estimators that operate on the *polychoric* correlation matrix rather than the observed covariance matrix. In categorical data analysis, a polychoric correlation is calculated by first assuming that there is a latent normal process underlying discrete data such that the discrete responses are a coarse approximation of a more articulate—but unobserved—normal distribution. For example, if participants are asked if they smoked a cigarette, the observed discrete information may be "Yes" or "No," but one could imagine a latent continuous (but harder to measure) normal process like "motivation to smoke" or "nicotine withdrawal" underlying this decision. On the latent normal distribution, there would be some *threshold* that demarcates the tipping point when a person's observed discrete response changes from "No" to "Yes." The polychoric correlation describes associations between these *latent normal distributions* underlying the observed data, rather than associations in the observed discrete data itself.

With this in mind, DFI simulations do not need to completely abandon multivariate normal distributions to accommodate discrete data, there just needs to be an additional step to discretize the simulated normal data. To do this, prior to fitting models to simulated data, DFI will use the estimated thresholds from the fitted model to discretize simulated data.

This will ensure that the simulated data have the same number of categories and the same response frequencies for each category as the observed data. Nothing else about the DFI algorithm needs to change—operations derived for continuous responses remain valid on the polychoric scale.

Practically, this facilitates implementation because thresholds are part of the standard output, so the threshold estimates can just be read in without any additional steps from the user. Another advantage is each item can potentially have a different number of categories (e.g., if some items are binary and others are Likert-type scales). Scales with a mixture of continuous and discrete items can also be accommodated because any item that does not have an estimated threshold will be simulated from a normal distribution without subsequent discretization (i.e., continuous items are a special case with 0 thresholds). In the DFI simulations, the estimator is changed to weighted least squares with mean and variance corrections (WLSMV) by default, although users can choose any estimator in `lavaan` that supports discrete data. This extension can be implemented in the `dynamic` R package with the `catOne` function for one-factor models or `catHB` for multifactor models.[1] A vignette overviewing practical implementation of the method using this software can be found at, https://rpubs.com/dmcneish/1025400.

The goal is to address points raised in the methodological literature where fixed cutoffs for discrete data cannot exist because fit indices scale differently depending on (a) the number of categories and the frequency with which they are endorsed (Monroe & Cai, 2015; Xia & Yang, 2018) and (b) the estimator (Beauducel & Herzberg, 2006; Shi & Maydeu-Olivares, 2020; Xia & Yang, 2019). With the two additional steps of (a) discretizing simulated multivariate normal data based on the model's estimated thresholds and (b) switching the estimator to a method that supports discrete data, DFI cutoffs will be optimally sensitive to the user's model characteristics *and* the exact response scale *and* response frequencies *and* the chosen estimator. The benefit is that researchers no longer need to tepidly apply traditional cutoffs to models—and response scales—to which the cutoffs were never intended to generalize. Correspondingly, fit assessment and scale validation will be more accurate and lead to studies using scores with better psychometric properties. In the next section, we evaluate these claims with a simulation.

Simulation Design

We assess the ability of the proposed DFI extension to classify correct and incorrect models with Likert-type responses, both in absolute terms and relative to traditional cutoffs. Length restrictions result in the main text restricting focus to one-factor models. The simulation design is largely inspired by McNeish and Wolf (2023a), which explored sensitivity to misfit of DFI and traditional cutoffs for one-factor models with continuous responses. DFI cutoffs proposed for discrete responses would ideally mirror results reported in McNeish and Wolf (2023a). The Supplemental Material describes the design and results of a simulation for multifactor models.

Data Generation Models and Conditions

We simulate data according to one of two models. In the first model, items are generated to be truly unidimensional. In the second model, generated data are truly multidimensional, where 75% of items load on one factor and 25% of items load on a second factor. A one-factor model is subsequently fit to all simulated data. The fitted model is correct and is expected to fit well when applied to the simulated unidimensional data if the model is being evaluated accurately. The fitted model is incorrect for simulated multidimensional data and these replications should be identified as fitting poorly if the model is being evaluated accurately. As described shortly, the main interest is whether different cutoffs accurately classify models.

McNeish and Wolf (2023a) generated data from multivariate normal distributions using covariance matrices implied by these two data generation models. In our simulation, we initially used multivariate normal distributions as well, but we subsequently discretize all items so that observed data resemble Likert-type responses (e.g., similar to the approach in Shi et al., 2020). This mirrors assumptions of polychoric correlations, whereby ordinal data are considered to manifest from an underlying normal distribution. We have two conditions for the number of response categories, 3 and 5. We also have two conditions for the response distribution shape: *balanced* and *skewed*. In the balanced condition, the middle response option was most common and more extreme responses were endorsed less frequently. The upper end and lower end of the scale were equally endorsed in the balanced condition. In the skewed three-category condition, the response frequencies were 10%, 20%, and 70%, respectively. In the skewed five-category condition, the response frequencies were 10%, 10%, 10%, 20%, and 50%, respectively.

We include two sample size conditions (400, 1,000), two conditions for scale length (8, 12), and three conditions for standardized loading magnitude (0.60, 0.75, 0.90). Two hundred replications are conducted for each condition. The somewhat low number of replications is due to the computational nature of DFI because each replication of the simulation contains several thousand subreplications. Data were generated with the `simstandard` R package (Schneider, 2019), and models were fit in the `lavaan` R package (Rosseel, 2012) with the WLSMV or ULSMV estimators.[2] To be clear, we are treating the responses as

[1] Functions can currently be installed in R via the `dynamic` GitHub package, https://github.com/melissagwolf/dynamic. The CRAN package and Shiny app will be updated shortly after final acceptance.

[2] Resplication-level simulation results are available on the Open Science Framework at https://osf.io/2t4zh/.

discrete (i.e., `ordered = TRUE` in `lavaan`), and we are not treating Likert-type responses as continuous.

Justification for Conditions

A review of empirical factor analysis reporting practices by Jackson et al. (2009) found an average sample size of 389 and D'Urso et al. (2022) found a median sample size of 400, which motivated the $N = 400$ condition. The $N = 1{,}000$ condition roughly represents the 85th percentile of the sample size distribution in Jackson et al. (2009) and was included to ensure that DFI cutoffs do not get distorted at large samples (as can happen with traditional cutoffs; Marsh et al., 2004).

Scale length conditions were informed by reviews by Flake et al. (2017), D'Urso et al. (2022), and Jackson et al. (2009) who all report that the average scale length was about 7. We used 8 in the simulation to roughly represent an average scale, using an even number to make it easier to systematically manipulate conditions (described in the next section). Flake et al. (2017) reported that scales one standard deviation above average length were about 13 items, so the 12-item condition represents an above average length, rounded to an even number.

We generated items with three or five categories because five or fewer categories typically require treating the data as discrete to avoid biased estimates (e.g., DiStefano & Morgan, 2014; Muthén & Kaplan, 1992). Previous studies also suggest that response distribution shape affects performance of factor models for discrete data (Rhemtulla et al., 2012). Different estimators were considered because WLSMV is the default in both `lavaan` and M*plus*, but studies have suggested that ULSMV may have better properties (e.g., Shi et al., 2020; Xia & Yang, 2019).

Creating Misspecification in Generated Data

In the data generation model, the loading conditions only applied to certain items to keep the magnitude of misspecification constant throughout the conditions. In the multidimensional data generation conditions, the 75% of items loading on the first factor had their loadings manipulated according to the three conditions. However, the 25% of items loading on the second factor were constant at 0.60 and the factor correlation was constant at 0.50. These values were selected because, when a one-factor model is fit, this would imply factor loadings of 0.30 for the items that were generated to load on the second factor. Loadings of 0.30 may provide possible (but weak) evidence of unidimensionality, so a model with these estimates would not be clearly dismissed merely from the parameter estimates. For the unidimensional data generation condition, 75% of item loadings again were manipulated based on the three conditions and 25% of the items were constant at 0.30. This way, the estimates from the fitted models would be the same regardless of the data generation condition (but misfit would differ between conditions).

Simulation Outcomes

The main interest is whether fit index cutoffs correctly identify that one-factor models applied to unidimensional data fit well while one-factor models applied to multidimensional data do not. We quantify this as the proportion of replications in which (a) fit indices exceed the traditional cutoff established by HB and (b) fit indices exceed the Level-1 DFI cutoff. The Level-1 DFI cutoff is used because it is intended to be sensitive to roughly the same magnitude misspecification as traditional cutoffs, which should foster a relatively fair comparison. We also track the proportion of replications in which DFI cutoffs with at least 90% sensitivity to misspecification are available because availability is not present with traditional cutoffs.

Simulation Results

Unidimensional Data Results

As anticipated, there were few issues when fitting one-factor models to unidimensional data. Traditional cutoffs misclassified 0% of replications as poorly fitting across all conditions. DFI misclassification rates were often 0%, but there were a few conditions where some misclassification was present (at most in 6% of replications, but more often 1%–3%) which occurred when $N = 400$ or with a skewed response distribution. Supplemental Tables S4 and S5 show the full results. Overall, traditional and DFI cutoffs do not indiscriminately classify all models as fitting poorly, which warrants inspecting classification rates for misspecified models.

Eight-Item Multidimensional Data Results

Table 1 shows the proportion of replications in which cutoffs correctly classified a one-factor model with WLSMV as misfitting when applied to eight-item multidimensional data. Numbers outside parentheses are the proportion of correct classification for replications in which DFI cutoffs were available. Numbers within parentheses are the percent of replications for which DFI cutoffs were available. If no parenthetical value is present, DFI cutoffs were available for all replications in the corresponding condition. Results for the eight-item conditions with ULSMV are similar and are provided in Supplemental Table S6 (traditional cutoffs perform slightly better with ULSMV).

Traditional cutoffs in Table 1 highlight properties for which they have been criticized previously. First, the traditional SRMR cutoff had no utility to identify this misspecification and correctly identified 0% of replications as

Table 1

Percentage of Replications Correctly Classified as Misfitting When a One-Factor Model Is Applied to Multidimensional Data With Eight Items (Higher Values Are Better)

			Three categories						Five categories					
			Balanced			Skewed			Balanced			Skewed		
N	Loadings	Cutoffs	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI
400	0.90	DFI	99	96	97	91	91 (73)	92 (62)	97	96	97	96	96	96
		HB	0	81	0	0	57	0	0	97	0	0	87	0
	0.75	DFI	99	97	97	(1)	(2)	(1)	97	96	97	93	94	93
		HB	0	70	11	2	31	13	0	91	13	0	71	14
	0.60	DFI	94 (56)	93 (73)	93 (59)	(0)	(0)	(0)	95	97	96	91 (64)	91 (82)	92 (71)
		HB	1	51	62	4	17	55	0	80	73	0	53	65
1,000	0.90	DFI	99	98	99	98	97	97	99	98	97	97	96	97
		HB	0	96	0	0	60	0	0	100	0	0	99	0
	0.75	DFI	97	97	97	97	96	96	99	99	99	97	97	97
		HB	0	83	3	0	27	4	0	100	3	0	87	6
	0.60	DFI	99	97	98	97	97	98	99	98	99	99	99	99
		HB	0	58	71	0	9	60	2	91	86	0	61	75

Note. Cells with numbers in parenthesis indicate that DFI cutoffs were not available for all 200 replications and the number in parentheses indicates the proportion of replications for which DFI cutoffs could be calculated. *N* = sample size. Fit indices correspond to a model estimated with diagonally weighted weight squares with mean and variance corrections. SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation; CFI = comparative fit index; DFI = dynamic fit index; HB = Hu and Bentler.

misfitting in most conditions, suggesting that the traditional SRMR cutoff is too lenient for one-factor models with discrete items. Second, the traditional RMSEA cutoff was mixed and confounded model characteristics with misfit. In some conditions (e.g., balanced response distributions with more response categories and strong loadings), correct classification rates were near 100%. However, these rates deteriorated as loadings became weaker, fewer response options were provided, or responses favored extreme categories and correct classification rates fell as low as

Table 2

Percentage of Replications Correctly Classified as Misfitting When a One-Factor Model Is Applied to Multidimensional Data With 12 Items (Higher Values Are Better)

			Three categories						Five categories					
			Balanced			Skewed			Balanced			Skewed		
N	Loadings	Cutoffs	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI
400	0.90	DFI	100	100	100	100 (93)	100 (89)	99 (77)	100	100	100	100	100	100
		HB	0	92	0	1	50	0	0	98	0	0	96	0
	0.75	DFI	100	100	100	100 (21)	100 (75)	100 (64)	100	100	100	100	100	100
		HB	0	89	19	3	27	14	0	96	32	0	84	27
	0.60	DFI	100	99	100	(1)	96 (40)	96 (27)	100	100	100	99	99	99
		HB	1	56	91	9	10	76	0	87	95	0	58	91
1,000	0.90	DFI	99	98	99	98	97	97	99	98	97	97	97	97
		HB	0	96	0	0	60	0	0	100	0	0	99	0
	0.75	DFI	100	100	100	100	100	100	100	100	100	100	100	100
		HB	0	94	15	0	24	10	0	100	34	0	94	27
	0.60	DFI	100	100	100	100	100	100	100	100	100	100	100	100
		HB	0	68	98	0	3	88	0	97	100	0	66	98

Note. Cells with numbers in parenthesis indicate that DFI cutoffs were not available for all 200 replications and the number in parentheses indicates the proportion of replications for which DFI cutoffs could be calculated. *N* = sample size. Fit indices correspond to a model estimated with diagonally weighted weight squares with mean and variance corrections. SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation; CFI = comparative fit index; DFI = dynamic fit index; HB = Hu and Bentler.

9%. Third, the traditional CFI cutoff was typically poor (e.g., Xia & Yang, 2019, p. 421), but improved as loadings became weaker. Overall, traditional cutoffs rarely provided accurate or consistent fit evaluation in these conditions.

DFI cutoffs fared much better. Classification rates were high across all conditions, for all three indices, with correct classification rates never dropping below 91%. This provides some evidence that DFI had success in unconfounding model characteristics from misfit such that misspecification can be consistently detected regardless of the model characteristics, response distribution, or estimator. A caveat is that DFI cutoffs were often unavailable in some conditions. This was particularly prevalent for conditions with three response options, weaker loadings, and a skewed response distribution. In some combinations of conditions, DFI cutoffs were unavailable for all replications.[3] This indicates that the sampling variability of fit index distributions was large, and a single value could not reliably distinguish correct from misspecified models. In other words, the data did not contain enough information to confidently evaluate fit with at least 90% sensitivity to misspecification. Predictably, excessive sampling variability precluding availability of DFI cutoffs was more prevalent with discrete data than in the simulation by McNeish and Wolf (2023a) for continuous data given that discrete data convey less information. Overall, DFI cutoffs consistently and accurately classified model fit with discrete data, provided that sampling variability was not too high to preclude derivation of DFI cutoffs.

12-Item Multidimensional Data Results

Table 2 shows the proportion of replications in which cutoffs correctly classified a one-factor model with WLSMV as misfitting when applied to 12-item multidimensional data. Results for the 12-item conditions with ULSMV were again similar and are provided in Supplemental Table S7. The general pattern of results is similar to Table 1. With traditional cutoffs, SRMR had no ability to detect that the model was inconsistent with the data; RMSEA performed worse when there were fewer categories, more skewed responses, or weaker loadings; and CFI was insensitive to misspecification with strong loadings but improved with weaker loadings.

Conversely, DFI cutoffs had a 100% classification rate when $N = 1{,}000$ or when there were five categories. Even with three categories and $N = 400$, DFI cutoffs corrected identified at least 96% of misspecified replications. The additional information provided by 12 versus eight items also decreased sampling variability and improved the percentage of replications for which DFI cutoffs were available in the $N = 400$, skewed three-category conditions. As intended, the DFI cutoffs appear able to maintain desired sensitivity to misspecification for a wide range of model characteristics.

Brief Follow-Up Simulation With Mixed Loading and Response Distributions

The previous subsections simulated all items to have equal loadings and response distributions to explore systematic patterns. In this section, we provide a more realistic picture of expected performance by mixing loading and response conditions within the same model. That is, we retain the same conditions for the number of items, sample size, and number of response categories; however, we simulated data where half the items have a balanced response distribution and the other half have a skewed distribution. Additionally, 25% of the items have 0.60 loadings, 25% have 0.75 loadings, 25% have 0.90 loadings, and 25% have loadings based on the misspecification condition (0.60 for multidimensional data generation, 0.30 for unidimensional data generation). This is intended to more closely match empirical scales that have items of varying quality. The main interest is the same where the goal is to determine if traditional and DFI cutoffs can accurately identify if the fitted model is correct or misspecified.

There were again no issues when fitting one-factor models to unidimensional data, so Table 3 only shows the ability to identify misspecified models for different conditions with WLSMV. ULSMV results are shown in Supplemental Table S8. Similar patterns emerge as in Tables 1 and Table 2. Namely, with traditional cutoffs, SRMR and CFI perform rather poorly, whereas RMSEA is highly sensitive to the model characteristics. Conversely, DFI cutoffs exhibit high accuracy and consistency for all three indices. As in the previous simulations, DFI cutoffs with at least 90% sensitivity were not always available in the upper left portion of Table 3 where data have less information (smaller samples, fewer response categories) because sampling variability is high in this context. Overall, DFI cutoffs seem to continue to provide improved classifications over traditional cutoffs in conditions that more closely mirror characteristics of empirical data.

Discussion and Limitations

Many psychologists are under the impression that understanding of factor model fit is more developed than it truly is. This is not their fault since methodological training often emphasizes fixed, universally applicable cutoffs. In reality, we have an idea which fit index values indicate good fit for a narrow subset of possible models with continuous responses estimated with maximum likelihood. However, many data sets with which psychologists are working do not

[3] To maintain comparability to traditional cutoffs, we only considered DFI cutoffs with at least 90% sensitivity. DFI software will report cutoffs with sensitivity below 90% if required. In practice, researchers with these conditions may still compute DFI cutoffs if they are comfortable with sensitivity below 90%. Supplemental Table S3 explores results if different minimum sensitivity values were used.

Table 3

Percentage of Replications Correctly Classified as Misfitting When a One-Factor Model Is Applied to Multidimensional Data Generated With a Mix of Factor Loadings and Item Responses Distributions (Higher Values Are Better)

			Eight items			12 Items		
N	Categories	Cutoffs	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI
400	3	DFI	89 (39)	80 (33)	80 (20)	100	99	99
		HB	0	47	3	1	55	4
	5	DFI	97	95	96	100	100	100
		HB	0	84	4	0	87	6
1,000	3	DFI	96	95	96	96	95	96
		HB	0	46	0	0	46	0
	5	DFI	97	97	97	97	97	97
		HB	0	96	0	0	96	0

Note. Cells with numbers in parenthesis indicate that DFI cutoffs were not available for all 200 replications and the number in parentheses indicates the proportion of replications for which DFI cutoffs could be calculated. SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation; CFI = comparative fit index; DFI = dynamic fit index; HB = Hu and Bentler.

have continuous responses and instead have discrete responses requiring other estimators, so much of the traditional guidance of interpreting fit indices may not be broadly applicable.

In this article, we proposed a potential remedy by augmenting the recently proposed DFI framework to accommodate discrete data. Simulation studies for one-factor models showed that the method produces cutoffs with desirable properties across a range of model characteristics, response distributions, response category options, and estimators for discrete data. The method is also accessible via an R package or a free Shiny app. Our hope is that this may provide psychologists with a low-barrier alternative to traditional cutoffs that can improve the accuracy of conclusions in common circumstances, where responses are discrete.

In general, these simulation results should be considered preliminary evidence supporting the potential for DFI cutoffs to help evaluate models with categorical outcomes rather than definitive and comprehensive evidence. For instance, as one limitation, the simulation only considered the ordinal case where items have three or five response categories but did not include a condition with only two response categories. The proposed method readily extends to binary items as a special case where there is a single threshold, but further simulation evidence would be needed to explore performance specifically with binary responses. In particular, the availability of DFI cutoffs may be impacted given that binary items contain less information.

As a second limitation, the very idea of cutoffs has faced criticism and dissenters contemplate whether a complex matter like model fit can be productively distilled into a single number. We, of course, agree that nuanced and holistic evaluation of factor models and validity evidence is preferable and do not wish to give the impression that relying solely on global fit is sufficient. Complementing any global fit assessment with local fit assessment to identify potential areas of local strain is a wise strategy (McNeish & Wolf, 2023b, p. 85), as is considering validity evidence from sources beyond internal structure. However, given that psychologists continue to heavily rely on fit index cutoffs and that there is no sign that this approach is waning, we feel that there is value in working to provide researchers with improved cutoffs that refine the original intent and logic of global fit indices. That is, even if cutoffs are not the ideal approach, cutoffs remain practically influential, so it is worthwhile to meet researchers where they are and try to improve frameworks in which empirical researchers are comfortable and accustomed.

Put plainly, as *prescriptively* noted in the methodological literature, the current “rules” for fit assessment have many shortcomings (e.g., Marsh et al., 2004). Nonetheless, despite these shortcomings, researchers continue to abide by these rules. Quixotic calls to alter practice have yielded little change, so DFI attempts to address shortcomings of fit assessment *descriptively* and views the issue as a constrained optimization problem. Methodologists have pointed out that the “rules” of fit assessment are silly, but the field is generally averse to changing these deeply engrained rules, so DFI is intended to be the smartest way to operate according to silly rules.

Methods journals are filled with insightful points that have potential to improve practice, but much of this work fails to generate buy-in from empirical researchers, either because it is too technical to be broadly accessible, it lacks of software support, or there is concern that new methods may not be well understood by readers, reviewers, and editors ultimately evaluating the research. By altering nothing except about the status quo except the numbers themselves, the goal of DFI is to facilitate buy-in for a new method, which will hopefully facilitate moving away from traditional cutoffs and toward improved methods, whether that is DFI or whatever comes next.

Third, our simulations only explored performance when using DWLS or ULS and treating the data as discrete. Two-step approaches exist where the polychoric correlation matrix of discrete data is calculated first, followed by maximum likelihood using the polychoric correlation as input (Jöreskog, 1990; Savalei, 2021). Future studies could compare the performance of this approach to discrete data, both with traditional and DFI cutoffs, especially to determine if this might help more frequent availability of DFI cutoffs at smaller sample sizes.

In closing, measurement and psychometrics serve a central role in psychological research because conclusions are only as strong as the scores used to represent focal psychological constructs. Much of our knowledge on evaluating factor analyses assumes continuous responses even though most responses in psychological data are discrete. This article contains one idea to improve evaluations of factor models in the context of discrete data along with some preliminary supporting evidence, but more research in this area is sorely needed given that relatively few studies exist in this space and those that exist do not receive attention commensurate with the prevalence of this type of data. We hope that this article helps to reveal the longstanding disconnect between the methodological literature and the type of data that psychologists often have. Additionally, we hope to inspire more interest in developing and applying more appropriate methods to evaluate the factor analyses that have massive unstated and unappreciated influence on our collective knowledge of psychological phenomena.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Psychological Association.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods*, *24*(1), 1–19. https://doi.org/10.1037/met0000195

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815–824. https://doi.org/10.1016/j.paid.2006.09.018

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., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606. https://doi.org/10.1037/0033-2909.88.3.588

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley. https://doi.org/10.1002/9781118619179

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511490026

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Sage 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

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, *36*(4), 462–494. https://doi.org/10.1177/0049124108314720

Crutzen, R., & Peters, G. Y. (2017). Scale quality: Alpha is an inadequate estimate and factor-analytic evidence is needed first of all. *Health Psychology Review*, *11*(3), 242–247. https://doi.org/10.1080/17437199.2015.1124240

D'Urso, D., Maassen, E., van Assen, M., Nuijten, M., De Roover, K., & Wicherts, J. (2022, July 29). *The dire disregard of measurement invariance testing in psychological science*. PsyArXiv. https://doi.org/10.31234/osf.io/n3f5u

Davey, A. (2005). Issues in evaluating model fit with missing data. *Structural Equation Modeling*, *12*(4), 578–597. https://doi.org/10.1207/s15328007sem1204_4

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling*, *21*(3), 425–438. https://doi.org/10.1080/10705511.2014.915373

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments & Computers*, *28*(1), 1–11. https://doi.org/10.3758/BF03203630

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

Flake, J. K. (2021). Strengthening the foundation of educational psychology by integrating construct validation into open science reform. *Educational Psychologist*, *56*(2), 132–141. https://doi.org/10.1080/00461520.2021.1898962

Flake, J. K., Davidson, I. J., Wong, O., & Pek, J. (2022). Construct validity and the validity of replication studies: A systematic review. *American Psychologist*, *77*(4), 576–588. https://doi.org/10.1037/amp0001006

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological & Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Furr, R. M. (2021). *Psychometrics: An introduction*. Sage Publications.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Greiff, S., & Heene, M. (2017). Why psychological assessment needs to start worrying about model fit. *European Journal of Psychological Assessment*, *33*(5), 313–317. https://doi.org/10.1027/1015-5759/a000450

Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement*, *71*(2), 306–324. https://doi.org/10.1177/0013164410384856

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, *16*(3), 319–336. https://doi.org/10.1037/a0024917

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

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., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*(2), 351–362. https://doi.org/10.1037/0033-2909.112.2.351

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

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183–202. https://doi.org/10.1007/BF02289343

Jöreskog, K. G. (1990). New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity*, *24*(4), 387–404. https://doi.org/10.1007/BF00152012

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

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

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

McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64–82. https://doi.org/10.1037/1082-989X.7.1.64

McNeish, D. (2022). Limitations of the sum-and-alpha approach to measurement in behavioral research. *Policy Insights From the Behavioral and Brain Sciences*, *9*(2), 196–203. https://doi.org/10.1177/23727322221117144

McNeish, D. (2023). Generalizability of dynamic fit index, equivalence testing, and Hu & Bentler cutoffs for evaluating fit in factor analysis. *Multivariate Behavioral Research*, *58*(1), 195–219. https://doi.org/10.1080/00273171.2022.2163477

McNeish, D., An, J., & Hancock, G. R. (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *Journal of Personality Assessment*, *100*(1), 43–52. https://doi.org/10.1080/00223891.2017.1281286

McNeish, D., & Wolf, M. G. (2023a). Dynamic fit index cutoffs for one-factor models. *Behavior Research Methods*, *55*(3), 1157–1174. https://doi.org/10.3758/s13428-022-01847-y

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

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. https://doi.org/10.3102/0013189X018002005

Miles, J., & Shevlin, M. (2007). A time and a place for incremental fit indices. *Personality and Individual Differences*, *42*(5), 869–874. https://doi.org/10.1016/j.paid.2006.09.022

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 fit. In M. Edwards & R. MacCallum (Eds.), *Current issues in the theory and application of latent variable models* (pp. 165–182). Routledge.

Monroe, S., & Cai, L. (2015). Evaluating structural equation models for categorical outcomes: A new test statistic and a practical challenge of interpretation. *Multivariate Behavioral Research*, *50*(6), 569–583. https://doi.org/10.1080/00273171.2015.1032398

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of nonnormal Likert variables. *British Journal of Mathematical & Statistical Psychology*, *38*(2), 171–189. https://doi.org/10.1111/j.2044-8317.1985.tb00832.x

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical & Statistical Psychology*, *45*(1), 19–30. https://doi.org/10.1111/j.2044-8317.1992.tb00975.x

Nunnally, J. C. (1967). *Psychometric theory*. McGraw-Hill.

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

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, *8*(2), 287–312. https://doi.org/10.1207/S15328007SEM0802_7

Pornprasertmanit, S., Wu, W., & Little, T. D. (2013). Using a Monte Carlo approach for nested model comparisons in structural equation modeling. In R. Millsap, L. van der Ark, D. Bolt, & C. Woods (Eds.), *New developments in quantitative psychology* (pp. 187–197). Springer. https://doi.org/10.1007/978-1-4614-9348-8_12

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

Ropovik, I. (2015). A cautionary note on testing latent variable models. *Frontiers in Psychology*, *6*, Article 1715. https://doi.org/10.3389/fpsyg.2015.01715

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

Savalei, V. (2021). Improving fit indices in structural equation modeling with categorical data. *Multivariate Behavioral Research*, *56*(3), 390–407. https://doi.org/10.1080/00273171.2020.1717922

Schneider, W. J. (2019). *simstandard: Generate Standardized Data*. R package (Version 0.3.0). The Comprehensive R Archive Network. https://cran.r-project.org/web/packages/simstandard/

Shaw, M., Cloos, L. J., Luong, R., Elbaz, S., & Flake, J. K. (2020). Measurement practices in large scale replications: Insights from Many Labs 2. *Canadian Psychology*, *61*(4), 289–298. https://doi.org/10.1037/cap0000220

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement*, *79*(2), 310–334. https://doi.org/10.1177/0013164418783530

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

Shi, D., Maydeu-Olivares, A., & Rosseel, Y. (2020). Assessing fit in ordinal factor analysis models: SRMR vs. RMSEA. *Structural Equation Modeling*, *27*(1), 1–15. https://doi.org/10.1080/10705511.2019.1611434

Sivo, S., Fan, X., Witta, E., & Willse, J. (2006). The search for "optimal" cutoff properties: Fit index criteria in structural equation modeling. *Journal of Experimental Education*, *74*(3), 267–288. https://doi.org/10.3200/JEXE.74.3.267-288

Slaney, K. (2017). *Validating psychological constructs: Historical, philosophical, and practical dimensions*. Macmillan. https://doi.org/10.1057/978-1-137-38523-9

Tanaka, J. S. (1987). "How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. *Child Development*, *58*(1), 134–146. https://doi.org/10.2307/1130296

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with "well fitting" models. *Journal of Abnormal Psychology*, *112*(4), 578–598. https://doi.org/10.1037/0021-843X.112.4.578

West, S., Wu, W., McNeish, D., & Savord, A. (2023). Model fit in structural equation modeling. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (2nd ed., pp. 184–205). Guilford.

Wolf, M. G., & McNeish, D. (2021). *Dynamic model fit* (Version 1.1.0) [Computer software]. https://www.dynamicfit.app

Wolf, M. G., & McNeish, D. (2022). *dynamic: DFI cutoffs for latent variables models* (version 1.1.0) [Software]. The Comprehensive R Archive Network. https://cran.r-project.org/web/packages/dynamic

Wolf, M. G., & McNeish, D. (2023). dynamic: An R package for deriving dynamic fit index cutoffs for factor analysis. *Multivariate Behavioral Research*, *58*(1), 189–194. https://doi.org/10.1080/00273171.2022.2163476

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

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, Article e1. https://doi.org/10.1017/S0140525X20001685

Zhang, M. F., Dawson, J. F., & Kline, R. B. (2021). Evaluating the use of covariance-based structural equation modelling with reflective measurement in organizational and management research: A review and recommendations for best practice. *British Journal of Management*, *32*(2), 257–272. https://doi.org/10.1111/1467-8551.12415

Zhang, X., & Savalei, V. (2020). Examining the effect of missing data on RMSEA and CFI under normal theory full-information maximum likelihood. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(2), 219–239. https://doi.org/10.1080/10705511.2019.1642111

Zyphur, M. J., Bonner, C. V., & Tay, L. (2023). Structural equation modeling in organizational research: The state of our science and some proposals for its future. *Annual Review of Organizational Psychology and Organizational Behavior*, *10*(1), 495–517. https://doi.org/10.1146/annurev-orgpsych-041621-031401

Received April 12, 2023
Revision received May 30, 2023
Accepted June 7, 2023 ■