


A literature review of model fit and model comparisons with confirmatory factor analysis: Formalizing the informal in organizational science

Matt C. Howard¹  | Melanie Boudreaux² | Joshua Cogswell² | Kelly G. Manix³ | Matthew T. Oglesby⁴

¹Mitchell College of Business, University of South Alabama, Mobile, Alabama, USA

²Al Danos College of Business Administration, Nicholls State University, Thibodaux, Louisiana, USA

³Jennings A. Jones College of Business, Middle Tennessee State University, Murfreesboro, Tennessee, USA

⁴Sanders College of Business and Technology, University of North Alabama, Florence, Alabama, USA

Correspondence

Matt C. Howard, Mitchell College of Business, University of South Alabama, 5811 USA Drive S., Rm. 337, Mobile, AL 36688, USA.

Email: mhoward@southalabama.edu

Funding information

No funding was received in association with the current work

Abstract

Researchers often stray from recommendations provided by simulation studies when conducting confirmatory factor analysis (CFA), causing unwieldy applications of the analysis and diminished confidence in published results. We introduce three particularly important informal practices associated with (1) alternative interpretations of model fit, (2) the use of inadvisable combinations of fit indices, and (3) the failure to conduct effective model comparisons. We then review over 2000 CFAs in premier organizational science journals. Our results support that researchers widely engage in all three informal practices. To address this tension, we (1) formalize modern interpretations of model fit by providing percentile ranges of indices in

Joshua Cogswell, Kelly Manix, and Matthew Oglesby contributed equally to the manuscript, and their names appear in alphabetical order.

We would like to thank Jeffrey Lovelace, Andrew Hanna, Janaki Gooty, and especially Brett Neely for their feedback on an earlier version of the current article.

published articles, such that researchers can make relative and continuous assessments of model fit. We (2) emphasize the importance of assessing multiple recommended fit indices together to provide complete depictions of model soundness. Lastly, we (3) demonstrate the necessity to perform appropriate model comparisons, including the assessment of more complex models.

KEYWORDS

confirmatory factor analysis, fit indices, methodology, model comparisons, model fit, statistics

INTRODUCTION

When conducting a study, researchers test their developed hypotheses by assessing the relations of observed indicators intended to represent unobserved latent constructs; however, researchers can produce misleading results and inferences if indicators do not represent common latent constructs with properties that reflect their operational definitions and shared meanings (Brown, 2015; Brown & Moore, 2012; Harrington, 2009; Jackson et al., 2009). For this reason, an array of statistical techniques has been developed to assess whether indicators meaningfully relate to latent constructs. Of these, confirmatory factor analysis (CFA) provides particularly powerful evidence, causing it to be among the most frequently used techniques in organizational science (Credé & Harms, 2015; Hurley et al., 1997; Nye, 2022; Williams et al., 2004).

To perform a CFA, researchers must first develop an a priori model, which includes predictions about the number of latent factors represented by a set of indicators and the relations of these latent factors to each indicator (Brown, 2015; Lance & Vandenberg, 2002; Nye, 2022). CFA is then used to determine whether the resultant model estimates adequately reproduce the indicators' covariance, and while all slightly differ in their calculations, model fit indices are used to ascertain whether this goal was achieved (Credé & Harms, 2019; Kenny, 2023; Marsh et al., 2004; Schermelleh-Engel et al., 2003). Researchers also typically identify a set of alternative models with a varying number of specified factors to serve as reasonable relative comparisons for their hypothesized model. If the hypothesized model produces fit indices that meet or exceed suggested guidelines and generally perform better than the alternative models (e.g., based on model fit, factor loadings, and theoretical rationale), then the CFA supports that the indicators sufficiently relate to the latent factor(s) as proposed by the model. When undergoing this process, researchers can falsely claim support for their model if they incorrectly perform their CFA. Such a mistake would potentially result in incorrect tests of hypotheses that typically follow CFAs, as the researcher would draw inferences from indicators that may not accurately represent their underlying constructs of interest. Even if statistically significant results were obtained from these hypothesis tests, the meaning of the observed effects would be unclear.

Due to the importance of CFA, many authors have conducted simulation studies to offer guidance on the analysis (Curran et al., 1996; Heene et al., 2011; Hu & Bentler, 1999;

Koran, 2020; Marsh et al., 1988, 1998; McNeish & Wolf, 2021). Despite direct recommendations, researchers regularly stray from guidelines provided in these articles, which may be driven by skepticism that simulation studies producing widespread CFA guidelines did not include realistic conditions encountered in empirical research (McNeish & Wolf, 2021; Wolf & McNeish, 2023; Yuan et al., 2016). For instance, Nye (2022) in discussing model fit stated,

“The most stringent guidelines have suggested that an RMSEA $\leq .06$, CFI $\geq .95$, TLI $\geq .95$, and SRMR $\leq .06$ generally indicate good approximate fit. However, **in practice, these guidelines are often relaxed** such that an RMSEA $\leq .08$, CFI $\geq .90$, TLI $\geq .90$, and SRMR $\leq .08$ indicate moderate fit. Although the guidelines used to evaluate these fit indices have been based on empirical simulations (Hu & Bentler, 1999), **it is important to remember that their effectiveness is limited to the conditions that were simulated.**”

(p. 11., bold added for emphasis).

As exemplified by this quote, researchers appear to apply informal practices regarding model fit and potential model comparisons when conducting CFA, which poses two primary concerns. First, present norms for CFA may partly differ from article-to-article based on authors', reviewers', and editors' interpretations of best practices, resulting in unwieldy applications and diminished confidence in results. Second and perhaps more concerning, it is unclear whether and when these informal practices result in incorrect analyses and misleading interpretations. Our field may be partially built on improper measurement supported by inappropriate CFA practices, and researchers may be testing – and claiming support for – hypotheses based on results produced by indicators that do not represent their underlying latent constructs of interest.

To investigate the frequency with which such instances are occurring, we introduce and discuss three particularly important and potentially widespread informal practices associated with (1) alternative interpretations of model fit, (2) the use of inadvisable combinations of fit indices, and (3) the failure to conduct effective model comparisons. To guide our investigation, we develop research questions to better understand the prevalence and impact of these informal practices, and we perform a literature review of over 1000 articles that report over 2000 CFAs in premier organizational science journals since 2000. We use our results to offer evidence-based recommendations. Notably, we (1) formalize modern interpretations of model fit by providing percentile ranges of indices in published articles, such that researchers can make relative and continuous assessments of model fit; (2) emphasize the importance of assessing recommended fit indices together that provide complete depictions of model soundness; and (3) demonstrate the necessity to perform appropriate model comparisons, including the assessment of more complex alternative models, rather than nominal assessments that are currently pervasive in the literature.

The current article produces several implications, and we presently highlight three. First, formalizing informal practices can result in more consistent and accurate analyses with clearer interpretations, and curbing inappropriate informal practices altogether can result in more accurate analyses moving forward. As such, our efforts may prompt the reinvestigation of prior CFA findings. Previously supported models may be questioned when interpreted via the lens provided by the current article, as these prior studies may have produced particularly subpar fit indices and/or failed to consider plausible alternative models. These reinvestigations may lead to reinterpretations of established constructs and the development of new theories.

Second, methodologists have continuously called for researchers to move beyond dichotomous assessments of model fit, instead recommending continuous interpretations (Barrett, 2007; Jackson et al., 2009; Kline, 2023; Marsh et al., 2004; McNeish & Wolf, 2023). Even in creating their cutoffs, Hu and Bentler (1999) suggested that their guidelines could distinguish models with and without misspecifications, but researchers should also discuss the strength of their results; however, researchers of organizational science most often focus on whether their model fit met prespecified cutoffs rather than the magnitude of their fit indices. A potential cause of this failure to adopt continuous interpretations may be the lack of specified ranges to consider fit as weak, moderate, and strong. The current article resolves this tension by creating these specified ranges via estimating percentiles of fit indices, such that future researchers can make relative and continuous comparisons of model fit to empirically established guidelines. Therefore, the current article enables future researchers to move beyond dichotomous assessments of fit and adopt the widely suggested recommendation of continuous interpretations.

Third, among the largest developments in the interpretation of model fit in recent decades is the creation of dynamic fit index cutoffs (McNeish & Wolf, 2021; Wolf & McNeish, 2023). This approach uses Monte Carlo simulations to determine cutoffs that are specific to the model being tested. In developing this approach, McNeish and Wolf (2021) chose cutoffs that generally correspond to specifications of misfit considered in extant universal guidelines (e.g., Hu & Bentler, 1999; MacCallum et al., 1996). If organizational science researchers approve of larger misspecifications than recommended in simulation studies (as suggested by the quote above), then these researchers are at a crossroads. These researchers would need to determine whether standards in the organizational sciences need to be elevated to meet other fields, or whether these recent developments need to be modified to meet the standards of organizational science. That is, dynamic fit index cutoffs may need to be relative to the field being investigated, and new simulations may be required to develop field-specific dynamic fit index cutoffs. This consideration may be especially necessary, given that dynamic fit index cutoffs are often stricter than extant universal guidelines that organizational science researchers may already stray from, suggesting that these researchers may be particularly resistant to applying these new cutoffs.

BACKGROUND

Introduction to CFA

Researchers typically craft hypotheses regarding the relations of constructs, such as proposing that job satisfaction relates to job performance (Calder et al., 2021; Stamenkov, 2023). Due to this focus on constructs, analyses that can provide assurances that applied measures appropriately represent constructs of interest are particularly important, which has caused CFA to be an essential analysis in organizational science (and beyond) for several decades (Credé & Harms, 2015; Hurley et al., 1997; Nye, 2022; Williams et al., 2004).

CFA is a model-driven analysis (Brown, 2015; Harrington, 2009; Jackson et al., 2009). The researcher first designs a model of assumed latent factors (i.e., unobserved variables) that correspond to their intended constructs, and they also designate the relation of these latent variables to their studied indicators (i.e., observed measures) and each other latent variable. For instance, a researcher may intend to measure job satisfaction and job performance with a set of four indicators each. They would likely design a model of two latent factors that independently relate to

the two sets of four indicators, and they would correlate these two latent factors. CFA would then be used to statistically test whether this model specified by the researcher adequately represents the shared variance between the indicators within their collected dataset, which would provide support for whether these indicators appropriately relate to the researcher's intended constructs. This model-driven approach of CFA distinguishes it from its closely corresponding analysis, exploratory factor analysis (EFA), which similarly identifies latent factors representing the covariance underlying a set of indicators; however, EFA identifies latent factors and their relations without an a priori model, and it instead identifies the underlying latent factors that best fit the data in an exploratory manner (Howard, 2023; Howard & Henderson, 2023). Thus, CFA is valued because it provides a confirmatory test of the researcher's theoretical proposals.

Several pieces of information are used to infer whether the tested CFA model adequately represents the covariance in the collected data. CFA produces factor loadings, which indicate the relation of latent factors to their representative indicators (Brown, 2015). The stronger the factor loading, the more that the latent factor is believed to represent the indicator. A wide range of cutoffs have been proposed to determine whether a latent factor relates to an indicator strongly enough to be considered representative, but common recommendations are .40 and .50 (Brown, 2015; Brown & Moore, 2012). The model also indicates how much variance in each indicator is explained by its corresponding latent factor(s) and how much is not explained (i.e., unique or error variance). If modeled, CFA also estimates the relations between latent factors.

CFA also provides assessments of localized strain, such as residual covariance matrices and modification indices (Perry et al., 2015; Whittaker, 2012). These statistics indicate unmodeled aspects that benefit the model from their inclusion. For instance, a modification index may suggest that two indicators share additional variance not explained by their common latent factor, and therefore they should be covaried. Assessments of localized strain are essential to ensure that the tested model adequately represents each studied indicator, as relying on the other information provided by CFA may not detect these issues (Heene et al., 2011; Whittaker, 2012). We do not discuss localized strain due to the focus of our research questions, but those applying CFA should inspect these statistics to ensure the suitability of their models.

With particular relevance to the current article, CFA also provides assessments of global model fit, which is among the most discussed aspects of CFA (Heene et al., 2011; Koran, 2020; Marsh et al., 1988, 1998). Assessments of global fit, called fit indices, indicate the extent to which the tested model represents the underlying common variance of the indicators and a multitude of fit indices have been developed to make these assessments (Brown, 2015; Brown & Moore, 2012; Nye, 2022). Similarly, authors have proposed a number of cutoffs for each fit index to determine whether the model sufficiently represents the indicators' covariance, and researchers have identified which fit indices must be assessed in tandem to properly assess whether a model sufficiently represents the covariance within the collected dataset. The most popular of these suggestions likely remains those of Hu and Bentler (1999), who suggested that adequate model fit should be determined with the cutoffs of SRMR \leq .08, RMSEA \leq .06, NNFI/TLI \geq .95, CFI \geq .95, and IFI \geq .95. These guidelines also recommend that researchers should interpret SRMR along with RMSEA, NNFI/TLI, IFI, RNI, CFI, and/or CI.

Researchers are recommended to interpret combinations of fit indices because each has their own strengths and weaknesses, which arise due to the manner that they are calculated (Hair et al., 2019; Hu & Bentler, 1999; Kenny, 2023; Kline, 2023). Tables 1 and 2 provide summaries of commonly used fit indices and their calculations. Fit indices fall into various families based on their calculation, and fit indices of a common family are estimated more similarly

TABLE 1 Description of common absolute fit indices.

Fit index	Simplified description of formula	Selected known strengths	Selected known weaknesses
χ^2	Difference between observed and expected covariance matrices.	<ul style="list-style-type: none">Although very unreliable, it does provide a statistical significance test.^{1,2}Assessing ratio of χ^2 to degrees of freedom can partially address biases associated with favoring model complexity or greater number of variables.^{2,3}	<ul style="list-style-type: none">Substantially favors smaller sample sizes.^{1,2,3,4,5,6,7,8}Favors models with fewer indicators.^{1,3,4}Influenced by non-normal variable distributions.^{2,4,5,6,7}Favors greater unique variance in indicators.^{2,4,9,10}Favors complex models.^{2,5}
GFI	Proportion of accounted variance in observed covariance matrix by expected covariance matrix.	<ul style="list-style-type: none">Improvement beyond χ^2, but known weaknesses cause most authors to recommend other fit indices.^{1,2,4}	<ul style="list-style-type: none">Substantially favors larger sample sizes.^{1,5,6,8,11,12,13}Favors models with fewer indicators.^{4,6,11}Influenced by non-normal variable distributions.⁴Favors greater unique variance in indicators.^{14,15}Favors complex models.^{5,6}
AGFI	Proportion of accounted variance in observed covariance matrix by expected covariance matrix while adjusting for degrees of freedom.	<ul style="list-style-type: none">Improvement beyond GFI, but known weaknesses cause authors to recommend other fit indices.^{4,11}Penalizes nonparsimonious models.^{5,6,7}	<ul style="list-style-type: none">Substantially favors larger sample sizes.^{5,6,7,8,12,13}Favors models with fewer indicators.⁴Influenced by non-normal variable distributions.⁴Favors greater unique variance in indicators.^{14,15}
SRMR	Square root of average squared difference between observed and expected covariances.	<ul style="list-style-type: none">Less influenced by non-normal variable distributions.³	<ul style="list-style-type: none">Favors larger sample sizes.^{4,5,6}Favors models with fewer indicators.⁴Favors greater unique variance in indicators.^{5,10,15,16,17}Favors complex models.^{4,5,6}

TABLE 1 (Continued)

Fit index	Simplified description of formula	Selected known strengths	Selected known weaknesses
RMSEA	Chi square of hypothesized model adjusted for degrees of freedom and sample size.	<ul style="list-style-type: none">Penalizes nonparsimonious models.^{4,6,18,19}Known distribution enables to calculation of confidence intervals.^{1,4,6}	<ul style="list-style-type: none">Favors larger sample sizes.^{4,5,11,13,18,20,21}Favors models with more indicators.^{4,18,22}Influenced by non-normal variable distributions.⁴Favors greater unique variance in indicators.^{10,15,16,17}

Note: Superscripts represent the following citations that are included within our references section:

¹Hair et al. (2019), ²Kline (2023), ³Iacobucci (2010), ⁴Kenny (2020), ⁵West et al. (2012), ⁶Hooper et al. (2008), ⁷Schermelleh-Engel et al. (2003), ⁸Marsh et al. (1988), ⁹Browne et al. (2002), ¹⁰Heene et al. (2011), ¹¹Sharma et al. (2005), ¹²Bollen (1990), ¹³Ainur et al. (2017), ¹⁴Shevlin and Miles (1998), ¹⁵Hancock and Mueller (2011), ¹⁶McNeish et al. (2018), ¹⁷Miles and Shevlin (2007), ¹⁸Peugh and Feldon (2020), ¹⁹Hu and Bentler (1998), ²⁰Kenny et al. (2015), ²¹Chen et al. (2008), and ²²Kenny and McCoach (2003). It should not be inferred that the omission of any strength or weakness necessarily suggests that it does not apply to that fit index. It should also be recognized that each fit index is also more or less likely to detect certain types of model misspecifications, causing the interpretation of multiple model fit indices to be necessary regardless of strengths and weaknesses.



TABLE 2 Description of common incremental (or relative) fit indices.

Fit index	Simplified description of formula	Known notable strengths	Known notable weaknesses
NFI	Compares chi-square of hypothesized model and chi-square of null model.	<ul style="list-style-type: none">• Less influenced by non-normal variable distributions.¹	<ul style="list-style-type: none">• Influenced by sample size.^{1,2,3,4,5,6,7}• Favors models with fewer indicators.¹• Varied relation with common variance in indicators.^{5,8}• Favors complex models.^{1,2,9}
IFI	Compares chi-square of hypothesized model and chi-square of null model while adjusting for degrees of freedom.	<ul style="list-style-type: none">• Less influenced by small sample sizes.¹⁰• Less influenced by non-normal variable distributions.¹• Penalizes nonparsimonious models.^{2,11}	<ul style="list-style-type: none">• Somewhat favors smaller sample sizes.²• Favors models with fewer indicators.¹• Varied relation with common variance in indicators.⁸
NNFI/TLI	Compares chi-square divided by degrees of freedom of hypothesized model and chi-square divided by degrees of freedom of the null model.	<ul style="list-style-type: none">• Less influenced by sample size.^{1,2,4,5,6,7}• Less influenced by non-normal variable distributions.¹• Penalizes nonparsimonious models.^{1,2,3,4,11}	<ul style="list-style-type: none">• Somewhat favors larger sample sizes.³• Somewhat favors models with fewer indicators.^{1,12}• Varied relation with common variance in indicators.^{1,8}
CFI	Compares chi-square subtracted by degrees of freedom of hypothesized model and chi-square subtracted by degrees of freedom of the null model.	<ul style="list-style-type: none">• Less influenced by sample size.^{1,2,3,5,7,13}• Less influenced by non-normal variable distributions.¹• Penalizes nonparsimonious models.^{1,2,9,13,14,15}	<ul style="list-style-type: none">• Somewhat favors models with fewer indicators.^{1,12}• Varied relation with common variance in indicators.^{1,5,8,16,17,18}

Note: Superscripts represent the following citations that are included within our references section:

¹Kenny (2020), ²West et al. (2012), ³Hooper et al. (2008), ⁴Schermelleh-Engel et al. (2003), ⁵Ding et al. (1995), ⁶Marsh et al. (1988), ⁷Ainur et al. (2017), ⁸Miles and Shevlin (2007), ⁹Hair et al. (2019), ¹⁰Bollen (1990), ¹¹Hu and Bentler (1998), ¹²Kenny and McCoach (2003), ¹³Peugh and Feldon (2020), ¹⁴Iacobucci (2010), ¹⁵Shi et al. (2022), ¹⁶Heene et al. (2011), ¹⁷McNeish et al. (2018), and ¹⁸Hancock and Mueller (2011). It should not be inferred that the omission of any strength or weakness necessarily suggests that it does not apply to that fit index. It should also be recognized that each fit index is also more or less likely to detect certain types of model misspecifications, causing the interpretation of multiple model fit indices to be necessary regardless of strengths and weaknesses.

than those of differing families. Two primary families are absolute fit indices (e.g., GFI, AGFI, SRMR, and RMSEA) and incremental (or relative) fit indices (e.g., IFI, NFI, NNFI/TLI, and CFI). Absolute fit indices compare the expected covariance matrix generated from the researcher's model to the observed covariance matrix, and a better fit is obtained when the expected matrix more closely matches the observed matrix (Kenny, 2023; Kline, 2023). Incremental fit indices compare the performance of the researcher's model to the performance of a null model wherein all indicators are uncorrelated, and the better fit is obtained when the researcher's model is a greater relative improvement to the null model (see Widaman and Thompson (2003) for a thorough description). This estimation approach is similar to the R^2 statistic in regression (Brown, 2015).

These differing approaches cause absolute and incremental fit indices to pose differing considerations when interpreting model fit. For instance, Kenny (2023) suggests that violations of normality particularly influence absolute measures of fit, but they have a smaller effect on incremental fit indices. However, the nuances of each fit index cause them to have specific strengths and weaknesses beyond their family, of which several are listed in Tables 1 and 2. Namely, fit indices differ on whether they are particularly influenced by sample size, model size, model complexity, factor reliability, and many other attributes of both the studied data and tested models (Iacobucci, 2010; Peugh & Feldon, 2020; Shi et al., 2022; West et al., 2012). Each fit index is also sensitive to certain model characteristics and/or types of misfit (Fan & Sivo, 2007; Heene et al., 2012; Saris et al., 2009). For example, while some are particularly apt at identifying misfit due to omitting necessary paths, others are particularly apt at identifying misfit due to including unnecessary paths. Therefore, modern recommendations not only suggest that multiple fit indices are interpreted but guidelines derived from simulation studies should be followed that specifically recommend combinations of fit indices that explicitly counteract the weaknesses of each other (Hair et al., 2019; Hu & Bentler, 1999; Kenny, 2023; Kline, 2023).

Further, the negative consequences of proceeding with analyses despite evidence of concerning fit should be highlighted. Fit that falls well short of cutoffs could arise when major misspecifications are modeled, such as estimating an incorrect number of latent factors. If ignored, subsequent theorizing would be severely misguided because assumptions would be based on an incorrect interpretation of the construct (Credé & Harms, 2015; Hurley et al., 1997; Nye, 2022; Williams et al., 2004). However, simulation studies have also shown that model misspecification with sizable impacts can have relatively minor effects on model fit (Savalei, 2012; Shi et al., 2018). For instance, Ximénez et al. (2022) showed that omitting a cross-loading of .40 in a model typical of the organizational sciences (primary $\lambda = .50$ and $n = 500$) could alter RMSEA, SRMR, CFI, and GFI by only .01 or .02. In these instances, subsequent analyses could be notably biased by overlooking this cross-loading. The shared variance between the indicator and alternative latent factor could be attributed to the two latent factors if the cross-loading is not modeled, which would cause the two latent factors to appear more strongly related. In turn, this could cause researchers to interpret the two latent factors as meaningfully related, when their association may have instead arisen from the unmodeled measurement error (i.e., cross-loading). Therefore, this case could also produce inappropriate inferences regarding the tested theory.

The detrimental impact of overlooking model fit was the impetus of developing model fit cutoffs (e.g., Hu & Bentler, 1999; MacCallum et al., 1996; McNeish & Wolf, 2021), as prior authors identified points of fit that effectively differentiate models that do and do not include notable misspecifications. This detrimental impact also spurred researchers to call for continuous interpretations of model fit. As stated by McNeish and Wolf (2023),

“Treating fit indices more like the effect sizes they were intended to be protects the sanctity of exact fit tests while allowing researchers who are willing to accept some degree of misspecification in their models a way to quantify misfit more accurately. Separating these approaches to model evaluation by giving fit indices the vocabulary and framework it needs to operate as intended gives each perspective space to operate without encroaching on the mechanisms of the other perspective”

(p. 75).

As evidenced in this quote, researchers can still compare their fit to specified cutoffs, but utilizing continuous interpretations can also enable researchers to use the appropriate language to

stipulate the possibility that their model contains these types of misspecifications. Because impactful misspecifications may only produce relatively minor deviations on model fit, it is necessary for researchers to recognize and specify that gradients of model fit suggest whether a model is more or less likely to contain misspecifications – even beyond whether the model (failed to) met cutoffs. By doing so, more accurate understandings of CFA can be obtained.

Additionally, global model fit enables model comparisons. When conducting a CFA, the researcher should test their hypothesized model, but they should also test reasonable alternatives (Harrington, 2009; Jackson et al., 2009; Nye, 2022; Williams et al., 2004). In the example above, the researcher may also test a model that models three latent factors: one for job satisfaction, one for job performance, and one for organizational commitment. In this case, the researcher may believe it is plausible that some of the items for job satisfaction and/or job performance inadvertently measure organizational commitment, and they could test this alternative model to rule out this concern. In doing so, the researcher intends to demonstrate that their hypothesized model produces a better global fit than the tested alternatives, which could provide significant support that their hypothesized model is a better representation of their indicators than these possible alternative explanations (Brown, 2015; Brown & Moore, 2012). Therefore, the test of alternative models is an essential step to ensure satisfactory CFA results.

Although CFA may seem to be a straightforward process, we contend that researchers regularly engage in problematic practices regarding their interpretation of global fit and model comparisons. We detail our arguments below, which include the provision of specified research questions and hypotheses that are tested via our systematic literature review.

Before continuing, three considerations should be made. First, CFA cannot guarantee that constructs are adequately measured. CFA can indicate that a set of indicators represents a common latent variable, but it cannot identify the construct represented by that latent variable. These inferences must be made by conducting a host of assessments, including whether the latent factor relates as expected to related constructs (concurrent validity) and appears distinct from other constructs (discriminant validity) (Hinkin, 1995, 1998). CFA can perform these assessments, but CFA within itself cannot identify the meaning of a latent variable.

Second, CFA can be generalized to a host of other analyses. CFA is used for tests of measurement invariance, which assesses whether a model performs similarly across multiple groups, such as participants from different cultures (Somaraju et al., 2022). Likewise, CFA is used to assess method effects, such as the method factor technique (Podsakoff et al., 2024). In the current article, we do not discuss these alternative applications of CFA, as our scope is firmly on the use of CFA to assess the psychometric properties of measures. Researchers should be aware of these other uses, as they are invaluable for testing associated research questions.

Third, dynamic model fit is a recent and important development for CFA. Model fit is influenced by more than misspecification alone, and aspects such as model size and complexity systematically strengthen or weaken fit (Iacobucci, 2010; Shi et al., 2022; West et al., 2012). Dynamic model fit utilizes an algorithmic approach to estimate cutoffs for widely recommended fit indices (e.g., SRMR, RMSEA, and CFI) that are specific to the parameters of the model being tested (e.g., model size and complexity), thereby accounting for alternative aspects that may influence model fit. In doing so, this analytical approach aligns with the perspective that universal guidelines should not apply to all applications of CFA, and instead, guidelines catered to the application at hand are preferred. Further, dynamic model fit was created based on the misspecification modeled in earlier simulation studies, such as Hu and Bentler (1999) and MacCallum et al. (1996), and it often recommends cutoffs that are stricter than these universal guidelines. By providing inferences that relate to universal guidelines, our systematic literature

review can also provide considerations for dynamic model fit due to its analytical foundation. Thus, our results can broadly speak to all current approaches for interpreting model fit.

RESEARCH QUESTION AND HYPOTHESIS DEVELOPMENT

A multitude of simulation studies have provided recommendations for interpreting global model fit (Curran et al., 1996; Marsh et al., 1988, 1998), and Hu and Bentler's (1999) recommendations likely remain among the most widespread (Brown, 2015; Brown & Moore, 2012; Heene et al., 2011; Koran, 2020). As evidenced in the quote within our introduction, researchers appear to distrust simulation studies despite their insights into appropriate interpretations of CFA results, and it is possible – if not likely – that researchers are regularly straying from the recommendations produced by these simulation studies when interpreting their model fit. Therefore, we first investigate the frequency which researchers apply alternative interpretations of model fit than recommended in simulation studies for CFA, as it is possible that researchers are not applying the cutoffs derived from the simulations in these articles (e.g., SRMR \leq .08, RMSEA \leq .06, NNFI/TLI \geq .95, CFI \geq .95, and IFI \geq .95).

Applying alternative interpretations of model fit is not an inherently problematic practice. Alternative interpretations would suggest that researchers may have already moved beyond interpreting model fit in a dichotomous manner, but they are unable to explicitly state the strength of their model fit (e.g., weak, moderate, strong) due to the lack of accepted guidelines. For instance, authors often describe indices as approaching cutoffs, indicating that they consider their fit to be acceptable despite not meeting recommended cutoffs. More fully embracing continuous interpretations can enable nuanced readings of CFAs by potentially incorporating attributes of the tested constructs and model(s) that may impact certain model fit indices and these continuous interpretations of fit indices would be welcomed by methodologists (Barrett, 2007; Jackson et al., 2009; Kline, 2023; McNeish & Wolf, 2023). Researchers can, however, misinterpret results by overestimating the extent that model misspecifications have become accepted. For instance, a researcher may presently claim support for a model with inadequate fit by stating that it approached certain cutoffs, but this model may contain severe misspecifications and cause the researcher to misinterpret their construct(s) of interest. Thus, a systematic review to provide empirically based guidelines is essential before these interpretations can be reliable.

Additionally, the acceptance of worse model fit indices would suggest that researchers permit greater model misspecifications than those tested in simulation studies (Goretzko et al., 2024; Hu & Bentler, 1999), and researchers may frequently reconsider whether such strict requirements must be placed in the assessment of model validity. In other words, researchers may have become more lenient with what they consider a good model. Discovering this possibility would suggest that extant research may indeed differ from article-to-article based on authors', reviewers', and editors' interpretations of acceptable model fit, and more discrete guidelines are needed to reduce these unnecessary variations in modern research.

We address these tensions by investigating the following research question and creating formalizing guidelines that are informed by practices in the current literature via our review, such that accepted practices can be more uniformly and soundly applied. We calculate the percentage of published model fit indices that fall below cutoffs recommended by simulations to investigate whether researchers use informal guidelines. We then provide percentile ranges of popular model fit indices, such that researchers can determine whether their fit indices are very

weak (<10th percentile), weak (10th–33rd percentile), moderate (33rd–66th percentile), strong (66th–90th percentile), and very strong (>90th percentile). These percentile ranges and their labels were adopted from similar efforts in determining correlational (Bosco et al., 2015) and exploratory factor analysis (Howard, 2023) benchmarks. By providing these percentiles, we enable future researchers to provide general relative comparisons of their model to previously tested models.

However, many characteristics of data and models influence the magnitude of model fit indices independent of misspecifications, and researchers may prefer to compare their models to similar models tested in the prior literature for this reason. Among the strongest influences on the magnitudes of model fit indices is the degrees of freedom (df) (Iacobucci, 2010; McNeish & Wolf, 2023; Shi et al., 2022; West et al., 2012; Yin et al., 2023).¹ As seen in Tables 1 and 2, many fit indices include df in their calculation to account for this feature, but many authors have shown that fit indices – even those intended to account for df – are still sensitive to df (Kenny et al., 2015; Shi et al., 2022; Yin et al., 2023). For this reason, we also provide percentile range cutoffs separated by df tertiles, which enables future researchers to compare their fit to previously tested models with more similar characteristics by utilizing these percentiles. By doing so, these percentiles provide more apples-to-apples comparisons that account for a particularly important aspect of models that influences fit independent of misspecifications.

◦ Research Question 1: What are the magnitudes of published fit indices for CFA?

In addition to straying from recommended cutoffs, researchers may also interpret inadvisable combinations of model fit indices, as this is the other aspect of global model fit for which guidance has been provided via simulation studies (Hu & Bentler, 1999; Koran, 2020; Marsh et al., 1988, 1998). For this reason, we also investigate the frequency which researchers assess recommended combinations of model fit indices.

The failure to interpret recommended combinations of model fit is a problematic informal practice for CFA. Each fit index reflects different standards for interpreting model fit, and each has strengths and weaknesses such as being relatively sensitive to model size or complexity (Douma & Shipley, 2023; Jobst et al., 2022; Kline, 2023). Simulation studies have supported that specific combinations of model fit indices broadly provide the most accurate results, as the cumulative strengths of the fit indices reported in these combinations compensate for their weaknesses (Beauducel & Wittmann, 2005; Fan & Sivo, 2007; Hu & Bentler, 1999). Researchers risk misinterpreting results by selectively choosing which fit indices to interpret independent from recommendations provided by these simulation studies.

We propose the following research questions to assess the frequency that model fit indices are reported in published articles and determine whether researchers are using informal guidelines regarding which indices to interpret and report. If we discover that inadvisable combinations are often interpreted, then the current article can provide direct and clear feedback on avenues to improve present applications of CFA. We also provide a correlation matrix of reported fit indices to show that some are partly repetitive when reported together, and we report the correlations of these indices with sample size and df to illustrate that some are relatively sensitive to these aspects of research and model design. Ultimately, the intent of this effort is to reinforce suggestions of simulation studies regarding the interpretation of specific model fit indices together, curbing this ongoing informal practice and promoting extant formal guidelines.

- Research Question 2: Which fit indices are reported for CFA in published studies?
- Research Question 3: What is the relation of popularly reported fit indices for CFA with each other, sample size, and degrees of freedom?

Because researchers may be straying from recommended interpretations of model fit, we suggest that they may fail to abide by a closely related and recommended practice in conducting CFA – the assessment of reasonable alternative models. For this reason, we lastly assess the extent that researchers perform model comparisons and, when model comparisons are made, whether researchers compare their hypothesized model to more complex alternatives (i.e., fewer df).² While researchers are interested in whether their hypothesized model produces adequate fit, modern guides recommend that plausible alternatives should be tested because multiple models may produce adequate fit, and the model that best represents the data may be overlooked by only testing an adequately fitting hypothesized model (Tomarken & Waller, 2003, 2005). Thus, it is essential to test alternative models, and it would be problematic to not engage in this practice.

Plausible alternative models should have some type of theoretical rationale to justify their assessment (Tomarken & Waller, 2003, 2005). Unfortunately, it is common for researchers to test alternative models with a number of factors in the hypothesized model collapsed together, such as testing an alternative one-factor model to the hypothesized six-factor model. Alternative models with collapsed factors are often a weak comparison with little expectation of producing an improved fit, and it is problematic for researchers to rely on testing these relatively inefficient alternative models alone. The best-fitting model may again be overlooked by solely testing weak alternative models, which would be a problematic informal practice.

Given these considerations, we propose the following research question to assess the frequency of model comparisons and a number of tested alternative models (Jackson et al., 2009; Nye, 2022; Williams et al., 2004). We also assess whether researchers test alternative models with fewer df (i.e., more complex) than the hypothesized model and whether the best fitting model was the tested model with the fewest df (i.e., most complex tested model). By identifying the frequency that more complex models are tested, our results can speak to whether researchers have developed an informal practice to only nominally test alternative models, such that hypothesized models are tested against alternatives expected to produce a poorer fit. Moreover, if the model with the fewest df (i.e., most complex) is most often the best fitting, our results may also suggest that many better-fitting models have been overlooked due to neglecting more complex alternatives. Thus, we uncover the nature of this informal practice to encourage researchers to make more robust comparisons and test more complex alternative models.

- Research Question 4: Are researchers testing plausible alternative models with CFA?

METHOD

Supplemental Material A includes our systematic literature review database.

Article retrieval and coding

We performed a literature review of articles within premier journals of organizational science, which were determined by the TAMUGA journal list. This list is comprised of *Academy of Management Journal*, *Academy of Management Review*, *Administrative Science Quarterly*, *Journal of*

Applied Psychology, *Organizational Behavior and Human Decision Processes*, *Organization Science*, *Personnel Psychology*, and *Strategic Management Journal*. We performed Google Scholar searches using Publish or Perish 8 for the term “Confirmatory Factor Analysis” within these eight journals in March of 2024, with the results restricted from the year 2000 to the present. These searches resulted in an initial list of 1679 articles.

We then coded these articles in two phases. For both phases, two coders coded sets of 20 articles until sufficient interrater agreement was met (Cohen's κ or ICC $\geq .80$). Once an agreement was met, the two coders coded sources independently, conferring on any unclear decisions. In the first phase, the coders recorded how many CFAs were reported in each article. We did not include multilevel CFAs, meta-analytic CFAs, or multi-group CFAs. Guidelines for these analyses differ from standard CFAs, and including these would produce misleading findings. We also did not include CFAs that were solely used as a precursor to structural equation modeling (SEM) (i.e., measurement model), which was considered a CFA on all the same constructs as the subsequent SEM.³ Lastly, we did not include example CFAs in methodology-focused articles, as authors often purposely report poor results in these articles.⁴ From this first coding phase, we discovered 2422 CFAs reported in 1105 articles.

The same coders then reviewed each CFA for the characteristics detailed below. The attributes of the final selected model by the original authors were coded for the characteristics of model fit, sample size, and *df*. This was almost always the best-fitting model, but it could also be a more parsimonious model if multiple models produced a largely equivalent fit.

Model fit

We recorded all reported model fit indices, but we only discussed those reported for 80 or more CFAs. These were: SRMR, RMSEA, NNFI/TLI, IFI, CFI, GFI, NFI, RMSR, and AGFI. The criteria of 80 CFAs represented a natural breaking point in the frequency of reporting. The least frequently reported index that met this cutoff was reported 81 times (AGFI), whereas the most frequent index that did not meet this cutoff was reported only 23 times (RFI). Because few sources reported these indices, reviewing them would provide few benefits.

Sample size

We recorded the sample size for each CFA. To reduce the inflated impact of CFAs conducted with especially large samples, we rescaled extreme sample sizes (Aguinis et al., 2013). One study had a particularly extreme sample size ($n = 754,856$), which skewed the calculation of z-scores. We rescaled this study to the next largest sample size ($n = 60,602$) and then calculated z-scores for the sample size of each CFA. We then proceeded to rescale each sample size with a z-score larger than 6.00 to the largest sample size with a z-score smaller than 6.00. This resulted in the rescaling of 10 sample sizes to a value of 14,260.

Degrees of freedom

We recorded the *df* reported for each CFA. As detailed in our limitations section, other model aspects produced concerns due to inconsistent reporting. For instance, it was often unclear

whether reported fit indices reflected a model with all indicators and factors detailed in the methods section or whether unreported alterations were conducted, such as parceling (see Cortina et al., 2017); however, it could be more reliably assumed that the reported *df* reflected the tested model, as *df* is often provided by software with fit indices. Thus, *df* did not produce coding concerns as authors can be straightforward in their reporting.

Model comparisons

We recorded four model comparison aspects: whether alternative models were tested, the number of alternative models tested, whether an alternative model had fewer *df* than the hypothesized model, and whether the best-fitting model had the fewest *df*.

RESULTS

Primary results

We first assessed outliers regarding the number of CFAs reported per article. The average number of reported CFAs was 2.19, but the highest number of reported CFAs were 99 (z-score = 21.84), 81 (z-score = 17.78), 27 (z-score = 5.60), 25 (z-score = 5.14), and 24 (z-score = 4.92). To prevent articles with unusually large numbers of reported CFAs from having a disproportionate influence, we removed two articles with z-scores greater than 6.00 regarding the number of reported CFAs, resulting in a final sample of 1103 articles and 2242 CFAs used to calculate our results. All inferences are consistent between alternative analyses including these articles and our primary analyses reported below, supporting the robustness of our results.

To determine whether researchers adhere to the recommendations of simulation studies, we utilized the guidelines of Hu and Bentler (1999), which are still among the most applied CFA guidelines in organizational science (Nye, 2022). These guidelines were published before the timeframe of our review, making it justifiable to apply these standards to all included articles. Hu and Bentler recommend the following for cutoffs of adequate model fit: SRMR \leq .08, RMSEA \leq .06, NNFI/TLI \geq .95, CFI \geq .95, and IFI \geq .95. These guidelines also recommend that researchers should interpret SRMR along with RMSEA, NNFI/TLI, IFI, RNI, CFI, and/or CI.

We assessed the frequency that reported model fit indices fell short of cutoffs provided by simulation studies (Research Question 1), and Table 3 provides percentiles of reported model fit indices. Of studies that provided the fit index, we found that 95% of CFAs reported a SRMR of .08 or below ($k = 856$), 49% reported a RMSEA of .06 or below ($k = 698$), 48% reported a NNFI/TLI of .95 or above ($k = 358$), 59% reported a CFI of .95 or above ($k = 1010$), and 56% reported a IFI of .95 or above ($k = 137$). As close to half of reported CFAs fell short of widely used cutoffs for most indices, researchers are straying from provided recommendations. To formalize practices in the current literature, we provide percentiles of these fit indices for all models in Table 3. We also provide percentiles of fit indices separated by *df* tertiles in Table 4. Our discussion details recommended approaches to interpret model fit with these percentiles.

We next assessed the frequency that researchers are using certain model fit indices and combinations of indices (Research Question 2), which is reported in Table 5. SRMR was reported for 40% ($k = 902$); RMSEA was reported for 64% ($k = 1432$); and NNFI/TLI, IFI, CFI, and/or RNI was reported for 79% of CFAs ($k = 1773$). As recommended by Hu and Bentler (1999),



TABLE 3 Percentiles of reported model fit indices.

	SRMR	RMSEA	NNFI/TLI	IFI	CFI	GFI	NFI	RMSR	AGFI	χ^2/df
n	902	1432	746	245	1707	310	185	96	81	1525
10th percentile	.08	.09	.88	.90	.90	.87	.88	.08	.80	5.61
20th percentile	.07	.08	.90	.92	.92	.89	.90	.07	.82	3.95
25th percentile	.06	.08	.91	.92	.93	.90	.91	.06	.84	3.40
30th percentile	.06	.08	.92	.93	.93	.91	.92	.06	.85	3.00
33rd percentile	.06	.07	.92	.93	.94	.91	.92	.05	.87	2.79
40th percentile	.06	.07	.93	.94	.94	.91	.93	.05	.88	2.55
50th percentile	.05	.07	.94	.95	.95	.92	.94	.05	.89	2.22
60th percentile	.05	.06	.95	.96	.96	.93	.95	.04	.90	1.97
66th percentile	.04	.06	.96	.96	.97	.94	.96	.03	.91	1.81
70th percentile	.04	.06	.96	.97	.97	.95	.96	.03	.91	1.75
75th percentile	.04	.05	.97	.97	.98	.95	.96	.03	.92	1.65
80th percentile	.04	.05	.97	.97	.98	.96	.97	.02	.93	1.57
90th percentile	.03	.04	.98	.99	.99	.97	.98	.02	.94	1.35

Note: The top row indicates the number of CFAs included in the calculation of percentiles for the respective column. The following rows indicate the percentile values indicated by the left label for the statistic indicated by the top label. Bolded and underlined values indicate model fit indices that meet or exceed the cutoff recommendations of Hu and Bentler (1999).

TABLE 4 Percentiles of reported model fit indices.

		SRMR	RMSEA	NNFI/TLI	IFI	CFI	GFI	NFI	RMSR	AGFI	χ^2/df
	n	795	1197	623	209	1410	231	140	71	62	1497
Low DF (<51)	10th	.07	.11	.91	.92	.93	.90	.91	.08	.84	1.17
	33rd	.05	.08	.94	.95	.96	.93	.94	.06	.89	1.72
	Median	.04	.07	.96	.97	.97	.95	.95	.05	.90	2.27
	66th	.03	.06	.98	.98	.99	.96	.97	.03	.92	3.01
	90th	.02	.03	.99	1.00	1.00	.98	.98	.02	.94	7.07
Med DF (51–163)	10th	.08	.10	.89	.90	.90	.85	.87	.06	.81	1.43
	33rd	.06	.08	.93	.94	.94	.90	.91	.05	.85	1.87
	Median	.05	.07	.95	.95	.95	.91	.93	.05	.89	2.32
	66th	.05	.06	.96	.96	.96	.93	.94	.03	.91	2.90
	90th	.03	.04	.98	.97	.98	.96	.96	.00	.94	5.29
High DF (>163)	10th	.08	.08	.83	.88	.89	.71	.86	.07	.63	1.44
	33rd	.06	.07	.90	.91	.92	.87	.91	.05	.70	1.82
	Median	.06	.06	.92	.92	.93	.90	.93	.04	.81	2.14
	66th	.05	.06	.94	.94	.95	.91	.95	.03	.82	2.62
	90th	.04	.04	.97	.97	.98	.94	.97	.02	.89	4.95

Note: The top row indicates the number of CFAs included in the calculation of percentiles for the respective column. The following rows indicate the percentile values indicated by the left label for the statistic indicated by the top label. Bolded and underlined values indicate model fit indices that meet or exceed the cutoff recommendations of Hu and Bentler (1999).

TABLE 5 Number and percentage of CFAs that reported model fit indices and combinations.

	k (%)
1.) SRMR	902 (40%)
2.) RMSEA	1432 (64%)
3.) NNFI/TLI, IFI, CFI, and/or RNI	1773 (79%)
4.) 1 and 2	676 (30%)
5.) 1 and 3	885 (39%)
6.) 2 and 3	1400 (62%)
7.) 1 and 2 or 3	895 (40%)
8.) 1, 2, and 3	666 (30%)
9.) None ¹	421 (19%)

Note: Hu and Bentler (1999) recommendation is represented in Row 7.

¹These figures represent the number of CFAs that did not have any model fit indices reported whatsoever, including those beyond the recommendations of Hu and Bentler (1999).

SRMR was reported with RMSEA, NNFI/TLI, IFI, CFI, or RNI in 39% of CFAs ($k = 885$). SRMR and RMSEA with NNFI/TLI, IFI, CFI, and/or RNI were reported in 30% of CFAs ($k = 666$). No fit indices were reported in 19% of CFAs ($k = 421$). Because only 39% of CFAs reported model fit indices recommended by Hu and Bentler (1999), researchers appear to have developed informal guidelines for which indices to interpret.

To emphasize similarities and differences in model fit indices (Research Question 3), Table 6 presents their correlations. SRMR produced an average absolute correlation with the other fit indices of $|.48|$. RMSEA produced a smaller average absolute correlation, which was $|.25|$. χ^2/df produced the smallest average absolute correlation with the other fit indices of $|.07|$. NNFI/TLI, IFI, CFI, GFI, and NFI produced a very strong average absolute intercorrelation with each other, which was $|.73|$; however, their average absolute correlation with SRMR, RMSEA, and χ^2/df was $|.32|$. Of the two least commonly reported statistics, AGFI was extremely strongly correlated with GFI ($r = .97, p < .01$), and RMSR produced a small average absolute correlation with other fit indices ($r = .11|$); however, not enough studies reported RMSR with its most similar fit index, SRMR, to calculate a correlation, suggesting that both of these infrequently reported indices may be repetitive with more commonly reported indices.

To explore the association of sample size and df with the fit indices, we calculated their intercorrelations (Table 6). Sample size produced larger than a small correlation only with AGFI ($r = .36, p < .01$) and χ^2/df ($r = .33, p < .01$), whereas df produced larger than a small correlation with SRMR, NNFI/TLI, IFI, CFI, GFI, NFI, and AGFI ($r = .19|$ to $|.57|$). The largest of these correlations was between df and AGFI ($r = -.57, p < .01$).

We assessed model comparison practices (Research Question 4). Sixty-three percent of CFAs include model comparisons ($n = 1420$). Of CFAs that compared models, 50% compared two models ($n = 707$), 16% compared three models ($n = 233$), 10% compared four models ($n = 140$), 8% compared five models ($n = 107$), 3% compared six models ($n = 39$), 3% compared seven models ($n = 38$), and 5% compared eight or more models ($n = 75$). Six percent reported conducting model comparisons but did not specify how many models were compared ($n = 81$). Of CFAs that compared models, only 6% reported testing a model with fewer df , and the best fitting model in 94% of cases was reported to be the tested model with the fewest df .

TABLE 6 Correlation of sample size, degrees of freedom, and model fit indices.

	n	Df	SRMR	RMSEA	NNFI/TLI	IFI	CFI	GFI	NFI	RMSR	AGFI	χ^2/df
1.) n	-											
2.) df	.02 (1517)	-										
3.) SRMR	-.11** (900)	.25** (795)	-									
4.) RMSEA	-.01 (1431)	-.11** (1197)	.24** (676)	-								
5.) NNFI/TLI	-.12** (745)	-.25** (623)	-.49** (332)	-.46** (658)	-							
6.) IFI	.09 (244)	-.34** (209)	-.63** (75)	-.23** (136)	.83** (87)	-						
7.) CFI	.01 (1705)	-.19** (1410)	-.60** (837)	-.34** (1345)	.88** (683)	.98** (231)	-					
8.) GFI	.14* (309)	-.39** (231)	-.71** (86)	-.33* (209)	.51** (77)	.57** (52)	.57** (269)	-				
9.) NFI	.02 (184)	-.25** (140)	-.57** (42)	-.07 (114)	.84** (63)	.68** (54)	.77** (165)	.58** (77)	-			
10.) RMSR	-.08 (95)	.01 (71)	- (2)	-.02 (35)	.08 (34)	-.16 (24)	.02 (90)	-.10 (44)	-.05 (28)	-		
11.) AGFI	.36** (80)	-.57** (62)	-.62** (17)	-.51** (60)	.25 (20)	.13 (12)	.56** (67)	.97** (70)	.66** (29)	-.35 (15)	-	
12.) χ^2/df	.33** (1524)	.01 (1497)	.01 (796)	.02 (1213)	-.09* (622)	-.10 (213)	-.07* (1421)	.02 (240)	-.03 (141)	-.07 (75)	.24 (65)	-

Note: First number in each cell represents the correlation coefficient. The second number (in parentheses) represents the number of CFA results used to calculate the associated correlation coefficient.

Sensitivity analysis results

Our primary findings were estimated based on a review of premier organizational science journals. These outlets were specifically targeted because articles within premier journals are known to have a disproportionate influence on the field, and they are also typically believed to uphold higher standards than alternative outlets. By performing a systematic review of these outlets, the present article could also potentially produce a disproportionate influence on the field, and any concerning practices in these outlets would be likely to be present in alternative outlets – if not even more severe. These factors together would also encourage subsequent authors to adopt our recommendations. These assumptions cannot be guaranteed, however, and we, therefore, perform a sensitivity analysis wherein we replicate our primary findings in two alternative high-quality outlets in the organizational sciences that are outside our initial scope.

We replicated all search, coding, and analytical procedures to investigate CFA practices in the outlets, the *Journal of Organizational Behavior* and *Journal of Business and Psychology*. We chose these two journals because they are respected within the organizational sciences, but they are typically considered a different tier than the journals included in our primary analyses (e.g., A* vs. A in ABDC list). If our results replicate in these two outlets, they can be more confidently claimed to generalize to a broader range of research within organizational science.

The results of these sensitivity analyses are provided and fully discussed in Supplemental Material B, and they were remarkably similar to our primary analyses. Researchers strayed from model fit cutoffs to a similar extent, and the percentiles of model fit indices were similar to our primary analyses. Researchers were also shown to interpret inadvisable combinations of model fit indices, and they engaged in inappropriate model comparison practices. Therefore, the strong similarity between our primary analyses and these sensitivity analyses supports that our findings are robust and our recommendations can generalize across organizational science more broadly.

DISCUSSION

We proposed that researchers presently apply informal guidelines in organizational science to perform and interpret CFAs, particularly regarding model fit and comparisons. This practice is problematic, as the use of informal guidelines may cause standards for CFA to differ between authors, reviewers, and editors, producing an unwieldy, confusing, and potentially inaccurate field of research. To resolve this tension, we conducted a systematic literature review to determine the extent to which organizational science researchers use informal CFA practices. We aimed to use our findings to provide new guidelines for interpretations of model fit and model comparisons, which could formalize the informal practices that are potentially pervasive in the present literature. Below, we detail how our systematic literature review produced many insightful findings that lead to important recommendations to improve modern CFA practices in organizational science, providing guidelines to formalize the informal. We also discuss several directions for future empirical and methodological research in organizational science.

Model fit cutoffs and continuous interpretations

Perhaps most surprising, almost 20% of researchers did not report any model fit indices. This finding is very unexpected, and it perhaps most visibly demonstrates the need for significant

changes to modern reporting practices for CFA. In these cases, readers are entirely unable to determine whether models are valid representations of studied indicators when model fit indices are not reported, and almost 20% of reported CFAs provide little – if any – assurances for the psychometric properties of the applied measures. Future researchers should strongly consider the reinvestigation of prior models without any reported model fit indices provided in our Supplemental Material A. Results obtained in these studies may be particularly misleading, as the measures may be inappropriate representations of the studied constructs.

Our results also showed that, when reported, approximately half of the published CFAs fall short of the cutoffs provided by Hu and Bentler (1999) for most model-fit indices. This finding suggests that researchers have developed informal guidelines for the interpretation of model fit that do not directly correspond to prior simulation studies, and larger model misspecifications are considered appropriate than those considered in these simulation studies. This finding may also imply that researchers are not assessing fit on firm pass or fail criterion, as they are permitting their published models to fall short of accepted cutoffs (Hu & Bentler, 1999; MacCallum et al., 1996). Instead, they may partially interpret model fit in a continuous manner. While permitting larger misspecifications and interpreting model fit continuously is not inherently problematic, it is concerning that established guidelines are missing from organizational science.

To address this issue, we provided percentiles of all model fit indices published in the current literature (Table 3), and we also provided percentiles of model fit indices separated by *df* tertiles (Table 4). Both can be used as formal guidelines to interpret model fit in a continuous manner, such that researchers can use the former table for general interpretations and the latter table for interpretations that are more specific to the model being tested. Further, we provide percentile ranges and labels that correspond to similar efforts in determining correlational (Bosco et al., 2015) and exploratory factor analysis (Howard, 2023) benchmarks. Researchers can consider model fit indices below the 10th percentile to be very weak, between the 10th and 33rd percentiles to be weak, between the 33rd and 66th percentiles to be moderate, between the 66th and 90th percentiles to be strong and above the 90th percentile to be very strong. In interpreting these ranges, researchers do not need to make purely dichotomous assessments of meeting or failing to meet certain standards. Instead, researchers can interpret their results on a spectrum. Researchers can report their model fit indices as weak, moderate, or strong, providing a more accurate assessment of the psychometric evidence produced by their CFAs. By doing so, the current results provide a new approach to interpreting model fit and satisfies the calls of prior methodologists, but our results can also develop two primary directions for future research.

Researchers should consider reinvestigating prior applications of CFA, even beyond prior investigations that did not report model fit altogether. It is common for studies to fall short of cutoffs but use vague language to support their models, such as claiming that their models approached cutoffs or compensated for falling short. Some models, however, fell excessively short of cutoffs compared to other models in the current literature, such as the nearly 10% of articles in our review that reported an RMSEA above .10. Future researchers can utilize our literature review database provided in Supplemental Material A to identify CFAs that produced particularly problematic results, and they could reanalyze measures used in these articles. It is possible – if not likely – that incorrect interpretations regarding the substantive nature of latent constructs were made via these analyses, and widely studied constructs may have significantly different properties than presently assumed, such as a differing number of dimensions. By conducting these reinvestigations, organizational science can be based on more solid theoretical and empirical foundations with more accurate conceptualizations and operationalizations.

Future researchers of organizational science should also (re)consider the extent of model misspecification that is permissible. With the development of dynamic model fit cutoffs, these researchers appear to be at a crossroads. Dynamic model fit cutoffs were created based on the misspecification modeled in earlier simulation studies, such as Hu and Bentler (1999) and MacCallum et al. (1996). Because our results demonstrated that researchers often stray from cutoffs provided by these earlier simulation studies, future researchers may be resistant to apply dynamic model fit cutoffs, especially because these model-specific cutoffs are often stricter than earlier recommendations for model fit. Researchers of organizational science need to determine whether future studies must meet the standards of simulation studies and dynamic model fit cutoffs, or they need to determine whether domain-specific cutoffs need to be developed for dynamic model fit. That is, dynamic model fit identifies cutoffs that are specific to the model being tested (McNeish & Wolf, 2021; Wolf & McNeish, 2023), but it could be modified further to also consider the domain being studied. Future researchers could identify guidelines that both produce more accurate insights into model fit and more closely adhere to the informal (now formalized by our review) standards that are pervasive in organizational science. Therefore, the percentiles provided in the current article can be presently used to interpret model fit, but they may also be essential to future evolutions in creating new cutoffs for model fit indices.

Interpretation of model fit indices

We showed that researchers are not reporting widely recommended combinations of model fit indices, as less than half of the authors followed the recommendations of Hu and Bentler (1999) regarding which indices to interpret and report. These authors recommended reporting SRMR with RMSEA, NNFI/TLI, IFI, CFI, and/or RNI, but it was much more common for authors to report a combination of RMSEA, NNFI/TLI, IFI, CFI, and/or RNI without SRMR. This practice is concerning. Each model fit index has its relative strengths and weaknesses. Our results showed that NNFI/TLI, IFI, CFI, and RNI produce strong interrelations, indicating that each produces few insights beyond the other. These fit indices are also particularly sensitive to *df*, as evidenced in our correlation matrix and their notable percentile differences across *df* tertiles. By only reporting fit indices from one family (e.g., NNFI/TLI, IFI, CFI, and RNI), any supportive results may arise from biasing attributes rather than substantive relations of the indicators. For this reason, researchers should pay close attention to their fit indices and report those suggested by Hu and Bentler (1999), as these indices provide accurate assessments of model soundness when interpreted together. Researchers should also monitor developments for any new collection of indices that provides more complete and accurate information, as researchers are continuously conducting simulations that investigate a wider range of scenarios.

We also encourage reporting more model fit indices via the interpretation of indices in a continuous manner using the guidelines above. It is possible – if not likely – that some authors do not report suggested model fit indices because some fail to meet recommended cutoffs. By interpreting model fit indices on a spectrum, researchers may not feel such a strong tension when model fit indices fail to meet or approach cutoffs. Instead, researchers can signal which of their fit indices fell within the ranges of small, moderate, and/or large. In doing so, these researchers could (hopefully) specify that most of their model fit indices fell within the moderate and/or large ranges, but they could also acknowledge when some may fall within the small range. By doing so, these researchers could note that their model was largely supported by the

fit indices, but they could also more fully recognize that not all fit indices may be entirely supportive of their model – nor do they need to be based on the recommendations of prior authors (Brown, 2015; Brown & Moore, 2012; Hu & Bentler, 1999; Nye, 2022). In other words, interpreting model fit indices in a continuous manner discourages the pass-or-fail thinking that is persuasive with the dichotomous interpretation of model fit. Thus, the application of our recommendations can result in both more precise and more comprehensive reporting of CFA results.

Model comparisons

Our results lastly showed that about two-thirds of CFAs included model comparisons. Of these, CFAs were split between comparing two and more than two models. This result is not entirely problematic within itself. It would be preferable for more researchers to include model comparisons and to test all possible plausible alternative models; however, most applications of CFA may only involve one or two plausible alternative models, and researchers may have assessed all plausible alternative models when comparisons were made.

It is more concerning, however, that the vast majority of CFAs did not test more complex models than the hypothesized model, and the best-fitting model in the vast majority of CFAs was also the most complex tested model. This suggests that researchers are not testing all plausible alternative models, and model comparison practices may be largely nominal in current research. For instance, it is common for researchers to compare their hypothesized model to an alternative model with most – if not all – latent factors merged together, often providing little justification as to why this alternative model is plausible or expected to potentially produce a better fit than the hypothesized model. Then, the researcher claims support for their hypothesized model because it produced the best model fit (or produced an equivalent fit but aligned with the applied theoretical rationale), despite very little expectation for the alternative model(s) to provide a better model fit. Researchers should cease this practice and instead test alternative models including those that are more complex and less complex. Researchers should also provide sound justifications as to why these alternative models may be reasonable explanations for the covariance of their indicators, such that readers and reviewers can understand that the alternative models are robust tests of theory. By doing so, researchers can have greater assurances that their results represent their studied latent constructs, which cannot be provided for the vast majority of published CFAs.

Future directions for interpreting model fit

Utilizing our provided percentiles to perform relative comparisons of model fit represents a significant advancement in the current literature on CFA, but it should be recognized that this approach may still pose certain concerns of its own. Perhaps most notably, performing relative comparisons to the percentiles constructed from all CFAs enables authors to make general interpretations of model fit, whereas performing relative comparisons to percentiles matched for *df* enables interpretations of model fit that are more specific to the model being tested; however, some researchers may desire an approach that is even more catered to the data and model being tested, such as model fit cutoffs that are specific to the exact sample size, *df*, and other relevant characteristics of the model. Deriving such an approach is a separate endeavor from our

intent, but it should be emphasized that the current article advances researchers toward this endeavor.

Specifically, authors have developed regression-based approaches to identifying model fit cutoffs specific to the data and model being tested, which are associated with various levels of model misspecification. For instance, Yuan et al. (2016) developed adjusted fit index cutoffs for equivalence testing specific to the sample size and df of the model at hand. Future researchers could adapt these approaches to incorporate our systematic literature review database and develop a procedure that provides specific cutoffs while accounting for misspecification permitted in the current literature, which could provide cutoffs based on relative comparisons that are specific to the data and model being tested. In doing so, our systematic literature review database includes information on the reporting of 10 model fit indices, each of which could be included in this novel approach. As recommended above, future researchers could similarly incorporate our observations into the development of dynamic fit dynamic model fit cutoffs that are specific to organizational science, which would also represent a more catered approach to the model being tested. Therefore, while developing these alternative approaches are outside the scope of the current article, our systematic literature review may be essential in developing these novel methods for identifying model-specific cutoffs relative to the field being studied.

Limitations

We provided correlations to depict the relations of our studied fit indices, but this is a relatively simplified approach to showing their similarities and differences. A more complete understanding can be obtained by discussing the mathematical calculation of these fit indices and testing their performance via simulation studies. Due to the apparent distrust of simulation studies (Nye, 2022), we chose our approach to understanding the relations of fit indices via correlations, but future researchers should consider more in-depth mathematical discussions and simulation studies to provide a more complete depiction of their similarities and differences.

The present article focused on model fit and model comparisons due to their central importance, but several other aspects are involved in conducting CFAs. In general, we could not provide insights into these aspects because most authors do not report them, preventing their inclusion in a literature review. For instance, most authors do not report assessments of localized strain. Researchers should consider alternative investigations to draw attention to these other aspects. While researchers appear hesitant to fully adhere to the results of simulation studies, this approach may be the only viable option for investigating infrequently reported aspects of CFA.

Further, Cortina et al. (2017) showed that many researchers do not accurately report their model for CFA or SEM by comparing the expected df based on model descriptions (e.g., indicators and latent factors) to the reported df . We initially attempted to assess the extent that the presently reviewed articles incorrectly reported their model characteristics; however, authors' descriptions of their CFA models were often vague, making it impossible to determine whether the expected df matched the reported df . We similarly attempted to determine the extent that researchers utilized error covariances and/or parceling, but our df comparisons made it apparent that many researchers neglected to report these practices. As a result, it is difficult to replicate many reported CFAs, leaving it unknown whether they were appropriately conducted.

It should also be noted that this discrepancy caused the present article to assess the relations of model fit indices with df rather than the number of indicators or latent factors. It is believed

that researchers report the correct df from their analyses, although they may not fully disclose all model modifications that produced those df (Cortina et al., 2017). By analyzing the number of reported indicators and latent factors, we would run the risk of assessing these aspects with model fit indices produced from an entirely separate model; however, by analyzing df , we provide more assurances that we analyzed the correct figures associated with the tested model.

Taken together, many of these limitations speak to the importance of complete and consistent reporting of results. Coupled with increased attention to open science practices (Banks et al., 2019; Vicente-Saez & Martinez-Fuentes, 2018), it is critical to improve these practices. This need is particularly evident as our investigation focused on premier outlets that would be expected to have the highest standards for statistical analyses and reporting. Our supplemental analyses supported that similar practices were seen across all the studied premier outlets (Supplemental Material C), and our sensitivity analyses supported that similar practices were seen in two high-quality outlets outside of our primary scope. By providing these additional analyses, our results support that our findings are robust and able to broadly speak to the organizational sciences. We also found that the publication year explained only modest variance ($\leq 12\%$) in the aspects of CFA studied in the current article (e.g., magnitude and reporting of fit indices), indicating that researchers are improving in these manners but at a slow rate (Supplemental Material D). It is hoped that the current article will hasten these trends.

Lastly, any review is limited to the extant literature. We were unable to review alternative fit indices, such as AIC and BIC, because they were too rarely reported to meaningfully interpret. We could not discuss novel CFA techniques, such as Bayesian CFA, because they are too rarely conducted in our reviewed outlets. Researchers should monitor novel developments to identify when atypical CFA approaches and techniques may be ideal for their research questions. It should be recognized, however, that our results were consistent between the sources studied in the primary text and those studied in Supplemental Material B. This suggests that our observations are not specific to any one category or tier of outlet, and it provides assurances that our inferences may speak toward a wider range of researchers and research contexts. This finding also corresponds to recent authors who have conducted similar analyses to show consistency across studying tiers of outlets (e.g., Howard et al., 2024), which together suggests that the analytical practices of broader outlets generally correspond to those of premier outlets. Thus, inferences obtained from investigating these outlets may be more broadly generalizable.

CONCLUSION

The goal of our review was to create formal guidelines for ongoing informal practices regarding model fit and model comparisons in modern organizational research by assessing current CFA practices and considering their (mis)match with widespread recommendations. Our results showed that certain current practices have little correspondence to the recommendations provided by simulation studies, indicating that researchers are indeed utilizing informal practices. We produced formalizing guidelines for model fit and model comparisons from our results, which include the following actionable recommendations for current research:

1. Interpret and report model fit indices recommended by Hu and Bentler (1999) and more contemporary simulation studies. Hu and Bentler (1999) recommend reporting SRMR with RMSEA, NNFI/TLI, IFI, CFI, and/or RNI.

2. In addition to considerations about meeting cutoffs of simulation studies, researchers should use percentiles in Tables 3 and 4 to interpret model fit. Fit indices should be described as either very weak (<10th percentile), weak (10th–33rd percentile), moderate (33rd–66th percentile), strong (66th–90th percentile), and very strong (>90th percentile).
3. Researchers should test alternative models that are reasonable relative comparisons, including both less complex models and more complex models. Sound justification should be provided for the rationale of testing each alternative model.

By following these guidelines, researchers can move beyond dichotomous assessments of meeting or missing cutoffs, reinvestigate prior results to ensure adequate psychometric evidence, and consider future evolutions in domain-specific model fit assessments. Together, the current article addresses many of the tensions surrounding modern uses of CFA, and it opens many directions for future empirical and methodological research.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to disclose.

ETHICS STATEMENT

Because the current study did not include the collection of primary data, no IRB approval was sought.

DATA AVAILABILITY STATEMENT

All data associated with the current article are provided in the supplemental materials.

PERMISSION TO REPRODUCE MATERIAL

The current article does not reproduce any materials.

ORCID

Matt C. Howard  <https://orcid.org/0000-0002-2893-0213>

ENDNOTES

¹ Readers should refer to prior sources on the statistical calculation of df with CFA (Cortina et al., 2017; Rigdon, 1994), but the practical meaning of df should be presently considered. When comparing models with the same number of indicators, such as nested models, df is often considered a proxy of model complexity (Goretzko et al., 2024; Kenny, 2023; Preacher, 2003). This is because the inclusion of an estimated parameter reduces df by one, such as including an additional cross-loading. When the number of indicators is the same, models with fewer df include more estimated parameters, and models with fewer df are therefore typically considered more complex models. Alternatively, when comparing models with a differing number of indicators, df is often considered a proxy of model size (Kenny et al., 2015; Shi et al., 2022; Yin et al., 2023). This is because the inclusion of additional indicators has a greater effect than the inclusion of additional estimated parameters. For instance, a CFA model including two latent factors with five items each has a df of 34. Adding an additional cross-loading would cause the df to become 33, but adding an additional indicator to one of the factors would cause the df to become 43.

For these reasons, comparing df across different studies is typically more representative of model size (as studies include differing indicators), and df has been used as a proxy for model size in research drawing inferences across studies (Kenny et al., 2015; Shi et al., 2022; Yin et al., 2023). Because our percentiles draw inferences across studies, df could be considered an indicator of model size in this application, as done in prior research. On the other hand, we also use df in analyzing researchers' model comparison practices below. In this

case, we are inspecting the models that researchers compare when analyzing their collected set of indicators within a single study. In this case, the use of df does not compare across studies, and it instead only considers the df of models tested on the same set of indicators within studies. df has been used as a proxy for model complexity in prior research drawing inferences within individual studies (Goretzko et al., 2024; Kenny, 2023; Preacher, 2003). For this reason, df could be considered a proxy of model complexity in these analyses, as done in prior research. Thus, although both sets of analyses utilize df to draw inferences, the practical meaning differs between the two uses.

² See Footnote 1 for justification of df as a proxy for model complexity.

³ A CFA on all the same constructs as a subsequent SEM is typically called a measurement model, but some authors refer to the assessment of the measurement portion of a SEM as a CFA. We did not include measurement models for two primary reasons. First, measurement models provide support to analyze the relations of latent constructs via a subsequent structural model. As measurement models are part of SEM, it would be inappropriate to include them in a focused review of CFA. Second, researchers are more likely to perform modifications to measurement models than they would for CFAs. Because measurement models are precursors to structural models, researchers may be more likely to remove items and/or add error covariances to achieve appropriate fit and move from the measurement model to the structural model. On the other hand, authors may be less likely to perform such alterations with CFA, as the purpose of this analysis is to provide a close investigation of the psychometric properties of measures independent of subsequent analyses. It cannot be readily claimed that researchers treat measurement models and CFAs the same, which would support the exclusion of measurement models in an investigation of CFAs.

⁴ In these articles, authors often provide examples of poor results for illustrative purposes. In articles focused on CFA, it is common for researchers to report models with poor fit to show the potential of misinterpretations by ignoring fit. As these models are used as illustrations, they would be inappropriate to include in our review.

REFERENCES

- Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, 16(2), 270–301. <https://doi.org/10.1177/1094428112470848>
- Ainur, A. K., Sayang, M. D., Jannoo, Z., & Yap, B. W. (2017). Sample size and non-normality effects on goodness of fit measures in structural equation models. *Pertanika Journal of Science & Technology*, 25(2), 575–586.
- Banks, G. C., Field, J. G., Oswald, F. L., O'Boyle, E. H., Landis, R. S., Rupp, D. E., & Rogelberg, S. G. (2019). Answers to 18 questions about open science practices. *Journal of Business and Psychology*, 34, 257–270. <https://doi.org/10.1007/s10869-018-9547-8>
- 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., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling*, 12(1), 41–75. https://doi.org/10.1207/s15328007sem1201_3
- Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin*, 107(2), 256–259. <https://doi.org/10.1037/0033-2909.107.2.256>
- Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, 100(2), 431–449. <https://doi.org/10.1037/a0038047>
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.
- Brown, T. A., & Moore, M. T. (2012). Confirmatory factor analysis. In *Handbook of structural equation modeling* (2nd ed., Vol. 361). (p. 379). Guilford Press.
- 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>
- Calder, B. J., Brendl, C. M., Tybout, A. M., & Sternthal, B. (2021). Distinguishing constructs from variables in designing research. *Journal of Consumer Psychology*, 31(1), 188–208. <https://doi.org/10.1002/jcpy.1204>

- 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>
- Cortina, J. M., Green, J. P., Keeler, K. R., & Vandenberg, R. J. (2017). Degrees of freedom in SEM: Are we testing the models that we claim to test? *Organizational Research Methods*, 20(3), 350–378. <https://doi.org/10.1177/1094428116676345>
- Crédé, M., & Harms, P. (2019). Questionable research practices when using confirmatory factor analysis. *Journal of Managerial Psychology*, 34, 18–30. <https://doi.org/10.1108/JMP-06-2018-0272>
- Crédé, M., & Harms, P. D. (2015). 25 years of higher-order confirmatory factor analysis in the organizational sciences: A critical review and development of reporting recommendations. *Journal of Organizational Behavior*, 36(6), 845–872. <https://doi.org/10.1002/job.2008>
- 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>
- Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling: A Multidisciplinary Journal*, 2(2), 119–143. <https://doi.org/10.1080/10705519509540000>
- Douma, J. C., & Shipley, B. (2023). Testing model fit in path models with dependent errors given non-normality, non-linearity and hierarchical data. *Structural Equation Modeling: a Multidisciplinary Journal*, 30(2), 222–233. <https://doi.org/10.1080/10705511.2022.2112199>
- 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>
- Goretzko, D., Siemund, K., & Sterner, P. (2024). Evaluating model fit of measurement models in confirmatory factor analysis. *Educational and Psychological Measurement*, 84, 123–144. <https://doi.org/10.1177/00131644231163813>
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis*. Pearson.
- 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>
- Harrington, D. (2009). *Confirmatory factor analysis*. Oxford University Press.
- 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>
- Heene, M., Hilbert, S., Freudenthaler, H. H., & Bühner, M. (2012). Sensitivity of SEM fit indices with respect to violations of uncorrelated errors. *Structural Equation Modeling*, 19(1), 36–50. <https://doi.org/10.1080/10705511.2012.634710>
- Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, 21(5), 967–988. <https://doi.org/10.1177/014920639502100509>
- Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, 1(1), 104–121. <https://doi.org/10.1177/109442819800100106>
- Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modeling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, 6(1), 53–60.
- Howard, M. C. (2023). A systematic literature review of exploratory factor analyses in management. *Journal of Business Research*, 164, 113969. <https://doi.org/10.1016/j.jbusres.2023.113969>
- Howard, M. C., & Henderson, J. (2023). A review of exploratory factor analysis in tourism and hospitality research: Identifying current practices and avenues for improvement. *Journal of Business Research*, 154, 113328. <https://doi.org/10.1016/j.jbusres.2022.113328>
- Howard, M. C., Boudreaux, M., & Oglesby, M. (2024). Can Harman's single-factor test reliably distinguish between research designs? Not in published management studies. *European Journal of Work and Organizational Psychology*, 33(6), 790–804. <https://doi.org/10.1080/1359432X.2024.2393462>
- Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3(4), 424–453. <https://doi.org/10.1037/1082-989X.3.4.424>

- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: a Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Hurley, A. E., Scandura, T. A., Schriesheim, C. A., Brannick, M. T., Seers, A., Vandenberg, R. J., & Williams, L. J. (1997). Exploratory and confirmatory factor analysis: Guidelines, issues, and alternatives. *Journal of Organizational Behavior*, 18, 667–683. [https://doi.org/10.1002/\(SICI\)1099-1379\(199711\)18:6<667::AID-JOB874>3.0.CO;2-T](https://doi.org/10.1002/(SICI)1099-1379(199711)18:6<667::AID-JOB874>3.0.CO;2-T)
- Iacobucci, D. (2010). Structural equations modeling: Fit indices, sample size, and advanced topics. *Journal of Consumer Psychology*, 20(1), 90–98. <https://doi.org/10.1016/j.jcps.2009.09.003>
- Jackson, D. L., Gillaspay, 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>
- Jobst, L. J., Auerswald, M., & Moshagen, M. (2022). The effect of latent and error non-normality on measures of fit in structural equation modeling. *Educational and Psychological Measurement*, 82(5), 911–937. <https://doi.org/10.1177/00131644211046201>
- Kenny, D. (2023). Measuring model fit. *DavidAKenny.net*. Retrieved from: <https://davidakenny.net/cm/fit.htm>
- Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, 44(3), 486–507.
- Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling*, 10(3), 333–351. https://doi.org/10.1207/S15328007SEM1003_1
- Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford Publications.
- Koran, J. (2020). Indicators per factor in confirmatory factor analysis: More is not always better. *Structural Equation Modeling: a Multidisciplinary Journal*, 27(5), 765–772. <https://doi.org/10.1080/10705511.2019.1706527>
- Lance, C. E., & Vandenberg, R. J. (2002). Confirmatory factor analysis. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 221–254). Jossey-Bass/Wiley.
- 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>
- Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, 103(3), 391–410. <https://doi.org/10.1037/0033-2909.103.3.391>
- Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, 33(2), 181–220. https://doi.org/10.1207/s15327906mbr3302_1
- 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
- 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. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*, 28(1), 61.
- 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>
- 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>
- Nye, C. D. (2022). Reviewer resources: Confirmatory factor analysis. *Organizational Research Methods*, 26(4), 608–628.
- Perry, J. L., Nicholls, A. R., Clough, P. J., & Crust, L. (2015). Assessing model fit: Caveats and recommendations for confirmatory factor analysis and exploratory structural equation modeling. *Measurement in Physical Education and Exercise Science*, 19(1), 12–21. <https://doi.org/10.1080/1091367X.2014.952370>

- Peugh, J., & Feldon, D. F. (2020). "How well does your structural equation model fit your data?": Is Marcoulides and Yuan's equivalence test the answer? *CBE—Life Sciences Education*, 19(3), es5. <https://doi.org/10.1187/cbe.20-01-0016>
- Podsakoff, P. M., Podsakoff, N. P., Williams, L. J., Huang, C., & Yang, J. (2024). Common method bias: It's bad, it's complex, it's widespread, and it's not easy to fix. *Annual Review of Organizational Psychology and Organizational Behavior*, 11, 17–61. <https://doi.org/10.1146/annurev-orgpsych-110721-040030>
- Preacher, K. J. (2003). *The role of model complexity in the evaluation of structural equation models*. The Ohio State University.
- Rigdon, E. E. (1994). Calculating degrees of freedom for a structural equation model. *Structural Equation Modeling: A Multidisciplinary Journal*, 1(3), 274–278. <https://doi.org/10.1080/10705519409539979>
- Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications. *Structural Equation Modeling*, 16(4), 561–582. <https://doi.org/10.1080/10705510903203433>
- 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>
- Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2), 23–74.
- Sharma, S., Mukherjee, S., Kumar, A., & Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research*, 58(7), 935–943. <https://doi.org/10.1016/j.jbusres.2003.10.007>
- Shevlin, M., & Miles, J. N. (1998). Effects of sample size, model specification and factor loadings on the GFI in confirmatory factor analysis. *Personality and Individual Differences*, 25(1), 85–90. [https://doi.org/10.1016/S0191-8869\(98\)00055-5](https://doi.org/10.1016/S0191-8869(98)00055-5)
- Shi, D., DiStefano, C., Maydeu-Olivares, A., & Lee, T. (2022). Evaluating SEM model fit with small degrees of freedom. *Multivariate Behavioral Research*, 57(2–3), 179–207. <https://doi.org/10.1080/00273171.2020.1868965>
- Shi, D., Maydeu-Olivares, A., & DiStefano, C. (2018). The relationship between the standardized root mean square residual and model misspecification in factor analysis models. *Multivariate Behavioral Research*, 53(5), 676–694. <https://doi.org/10.1080/00273171.2018.1476221>
- Somaraju, A. V., Nye, C. D., & Olenick, J. (2022). A review of measurement equivalence in organizational research: What's old, what's new, what's next? *Organizational Research Methods*, 25(4), 741–785. <https://doi.org/10.1177/10944281211056524>
- Stamenkov, G. (2023). Recommendations for improving research quality: Relationships among constructs, verbs in hypotheses, theoretical perspectives, and triangulation. *Quality & Quantity*, 57(3), 2923–2946. <https://doi.org/10.1007/s11135-022-01461-2>
- 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>
- Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology*, 1, 31–65. <https://doi.org/10.1146/annurev.clinpsy.1.102803.144239>
- Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, 88, 428–436. <https://doi.org/10.1016/j.jbusres.2017.12.043>
- West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In *Handbook of structural equation modeling* (Vol. 1(1) (pp. 209–231). Guilford Press.
- Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *The Journal of Experimental Education*, 80(1), 26–44. <https://doi.org/10.1080/00220973.2010.531299>
- Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, 8(1), 16–37. <https://doi.org/10.1037/1082-989X.8.1.16>
- Williams, L. J., Ford, L. R., & Nguyen, N. (2004). Basic and advanced measurement models for confirmatory factor analysis. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology* (pp. 366–389). Blackwell Publishing.

- Wolf, M. G., & McNeish, D. (2023). dynamic: An R package for deriving dynamic fit index cutoffs for factor analysis. *Multivariate Behavioral Research*, 58, 189–194. <https://doi.org/10.1080/00273171.2022.2163476>
- Ximénez, C., Revuelta, J., & Castañeda, R. (2022). What are the consequences of ignoring cross-loadings in bifactor models? A simulation study assessing parameter recovery and sensitivity of goodness-of-fit indices. *Frontiers in Psychology*, 13, 923877. <https://doi.org/10.3389/fpsyg.2022.923877>
- Yin, Y., Shi, D., & Fairchild, A. J. (2023). The effect of model size on the root mean square error of approximation (RMSEA): The nonnormal case. *Structural Equation Modeling: A Multidisciplinary Journal*, 30(3), 378–392. <https://doi.org/10.1080/10705511.2022.2127729>
- 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: A Multidisciplinary Journal*, 23(3), 319–330. <https://doi.org/10.1080/10705511.2015.1065414>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Howard, M. C., Boudreaux, M., Cogswell, J., Manix, K. G., & Oglesby, M. T. (2025). A literature review of model fit and model comparisons with confirmatory factor analysis: Formalizing the informal in organizational science. *Applied Psychology*, 74(1), e12592. <https://doi.org/10.1111/apps.12592>