2026 ■ Vol. 22 ■ no. 2

An Objective Evaluation of Equivalence Testing for Model Fit: In Response to Marcoulides and Yuan (2025)

James L. Peugh [a,b], Kaylee Litson [c] & David Feldon [d]

[a]Behavioral Medicine Clinical Psychology, Cincinnati Children's Hospital Medical Center

[b]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati

[c]Department of Psychology, University of Houston

[d]Utah State University

Abstract ■ Peugh, Litson, and Feldon (2023) subjected equivalence testing (EQT) to Monte Carlo simulation accuracy and reliability scrutiny with disappointing results. Marcoulides and Yuan (2025) subsequently questioned several aspects of Peugh and colleagues' publication. We engage these claims and establish three conclusions, among others, related to the applied function of EQT. First, missing from Marcoulides and Yuan (2025) is consideration of McNeish's (2023) EQT Monte Carlo study, which reached similar conclusions to Peugh and colleagues. Second, Marcoulides and Yuan (2025) claimed to replicate one of Peugh and colleagues' (2023) simulation conditions, but a closer examination showed the omission of several key details needed to validate their conclusion. Finally, published research cited by Marcoulides and Yuan in support of EQT lacked rigorous Monte Carlo research verifying its accuracy or reliability. Based on results from Peugh and colleagues (2023) and McNeish (2023), and absent additional definitive Monte Carlo evidence establishing the accuracy and reliability of EQT, we assert that the encouragement of EQTs continued use is currently without empirical evidence.

Keywords ■ PutKeywordsHere.

✉ James.Peugh@cchmc.org

10.20982/tqmp.22.2.p071

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer

Introduction

When researchers draw inferences from structural equation models (SEMs), they also assess how well a statistical model (ideally derived from theory) fits their sample data (see Kenny, 2024; Lee & MacCallum, 2015). Accordingly, the tools used to evaluate model fit must be both reliable and valid to ensure that inferences based on model interpretation are reasonably appropriate. Although notoriously widespread, Hu and Bentler (1998, 1999) fit indices (e.g., comparative fit index [CFI], root mean square error of approximation [RMSEA]) have long been known to not meet these criteria (e.g., Davey, 2005; Fan & Sivo, 2005, 2007; Hancock & Mueller, 2011; Heene et al., 2011; Marsh et al., 2004; Millsap, 2007, 2013b, 2013a; Niemand & Mai, 2018).

Marcoulides and Yuan (2017; MY2017 from this point forward) introduced an intriguing alternative to assess model fit: equivalence testing (EQT). While recognizing this proposed EQT as arguably the most significant advance in model fit assessment since 1999, we also noticed that the accuracy and reliability of EQT as a model fit assessment tool had not been established via Monte Carlo simulation testing (Peugh & Feldon, 2020). Peugh, Litson, and Feldon (2023; PLF2023 from this point forward) did subject EQT to Monte Carlo simulation testing under common applied research conditions (both skewed and missing data) with disappointing results (see also Peugh & Feldon, 2024). In short, under conditions common to real world applied research (e.g., modest sample size, nonnormality, missing data), the accuracy and reliability of EQT estimates were

notably poor.

Marcoulides and Yuan (2025; MY2025 from this point forward) subsequently called into question several Monte Carlo simulation design decisions and outcome interpretations from PLF2023. We respond to these comments and specifically focus on four in detail to address what we see as those of consequence. Overall, we focus on EQT's performance as a fit index in its current form, not its theoretical underpinnings or its potential to become a better-informed fit index in future methodological work, as we did in the PLF2023 simulation study. Prior to PLF2023, EQT fit indices had not been subjected to Monte Carlo simulation method scrutiny to ensure that they worked as well as (or better than) standard traditional cut-point fit statistics (Hu & Bentler, 1998, 1999), despite the originators' recommendations for adoption into practice (e.g., see also Marcoulides & Yuan, 2024).

Before proceeding, we make an important point of clarification regarding the findings of PLF2023. Possibly lost in the mass of Monte Carlo simulation results were the conditions under which EQT *did* perform adequately. Generally, both the CFI_t and $RMSEA_t$ EQT fit indices performed acceptably (defined as a correct model fit decision "hit rate" of 80% or greater) across all CFA, path analysis, and SEM models, under limited conditions: when the population and data analytic models matched, data were normally distributed, contained no missing values, and if sample sizes were $N \geq 300$. However, these circumstances rarely, if ever, occur in applied research, as we discuss further below. Thus, despite its mathematically elegant form, the current practical utility of the EQT fit indices as a fit assessment function is limited.

MY2025's commentary on the approach and findings stated in PLF2023 can be distilled into a set of interrelated critiques concerning how EQT should be evaluated as a model fit assessment tool. In particular, MY2025 raised specific concerns about 1) the criteria used to judge EQT performance, 2) the role of misspecification size and tolerable error, 3) the use of dichotomized fit classifications, and 4) the interpretation of simulation results relative to conventional fit indices. In the sections that follow, we address these issues by clarifying the inferential standards historically required of any model fit assessment method, examining the accuracy of EQT across applied research conditions, evaluating the logical consistency of MY2025's comments, and situating EQT's performance within the broader empirical model fit evidence base. Throughout, our focus remains on the applied use of EQT alongside its empirical performance and function in simulation research, rather than on its current theoretical and mathematical form or future potential.

What Constitutes a Valid Inferential Test of Model Fit?

To properly assess equivalence testing (EQT) as a means to evaluate model fit, target outcomes must be quantified for a Monte Carlo simulation study (Morris et al., 2019). Sensitivity is defined here as the consistent ability to identify properly and improperly specified models as having acceptable and unacceptable fit, respectively. Inferential accuracy is defined here as the consistent ability to avoid both Type-I errors (judging a properly specified model as having unacceptable fit) and Type-II errors (judging an improperly specified model as having acceptable fit). The key point, we assert, is this: Rather than continuing to place model fit decisions on a continuum of subjectivity, an accurate, reliable, and valid mechanism to adjudicate model fit must consistently identify and *dichotomize* properly specified models as acceptable and misspecified models as unacceptable to a known, useful, and meaningful degree of precision.

In PLF2023, we used confirmatory factor analysis (CFA; Hu & Bentler, 1998, 1999; Marsh et al., 2004), path analysis (Bollen, 1989; McDonald & Clelland, 1984), and structural equation modeling (SEM; Fan & Sivo, 2005, 2007) data generating models that were either consistent with previously published model fit research in the case of the CFI and SEM, or based on path analysis models that have long been used as pedagogical exemplars. More importantly, our determination of "acceptable" or "unacceptable" was also based on previous fit index research: models matching the data generating model were labeled "properly specified," and models missing one or two parameters from the data generating model were labeled "misspecified." Further, misspecification magnitude was controlled quantitatively following the recommendations of Fan and Sivo (2005, 2007). Accordingly, the target outcome in PLF2023 was the frequency with which properly specified models produced "close or better" EQT results and misspecified models produced "fair or worse" results across 5,000 replications per condition. More importantly, as we show below, this dichotomous judgment approach was also utilized by, among others, Marcoulides and Yuan (2017). The point of emphasis here is that these PLF2023 simulation design choices were based on previously published Monte Carlo definitions and operationalizations, not on our own satisfaction criteria as MY2025 (p. 2) assert.

Generalizability and Nonnormality in EQT Model Fit Assessment

By drawing attention to the size of model misspecification provided by equivalence testing (ε_t), MY2025 implicitly make two presumptions that are not clearly supported by published evidence or logical arguments: (1) what constitutes an acceptable versus unacceptable amount of mis-

specification is definitively and empirically established and is completely free of researcher subjectivity; and (2) ε_t generalizes across SEM types regardless of analysis model complexity, sample size, or data characteristics (i.e., non-normal or missing data). EQT Monte Carlo simulation results from both PLF2023 and McNeish (2023) do not support either presumption.

Equivalence testing requires researchers to declare a non-zero tolerable model misspecification value that is assumed to be free of subjectivity and have inherent substantive meaning. If a tolerable amount of misspecification is not clearly meaningful, universally accepted, and agreed upon, then any non-zero value—such as ε_t—lacks such inherent meaning or practical informative utility, even when accompanied by a confidence interval that quantifies sampling uncertainty. Under such conditions, ε_t has been shown to be unable to meaningfully or accurately adapt to differences in model complexity or sample data characteristics (i.e., skewed or missing data; see McNeish, 2023; Millsap, 2007, 2013b, 2013a; Peugh et al., 2023). Further, placing a 95% CI around a point estimate of questionable value may provide only illusory utility in assessing model fit accurately, because the anchor of the confidence interval is itself selected independently of an objective empirical or consensus basis. This admittedly somewhat skeptical view is consistent with the published evidence thus far, as we also elaborate upon below.

Logical (In)Consistency in Evaluating EQT Model Fit Performance

MY2025 critique PLF2023 by appealing to conventional CFI and RMSEA values and their implied cutpoints. This was done despite repeatedly acknowledging—in their own prior work (e.g., Marcoulides & Yuan, 2017) and in MY2025—the well-documented shortcomings of conventional fit indices. Equivalence testing has been explicitly promoted as an alternative to these indices, with the implication that researchers can avoid both the theoretical subjectivity and empirical shortcomings of traditional fit statistics while still accurately, reliably and validly answering model fit questions.

However, the question of whether EQT was intended to replace conventional Hu and Bentler (1998, 1999) fit indices and their associated cutpoints is, we argue, ultimately inconsequential. If EQT were able to meaningfully and accurately incorporate sampling variation, adjust for analysis model complexity, compensate for less than ideal real-world sample data characteristics (e.g., skewed and missing data), and offer interpretational "bins" capable of correctly assessing model fit without an increase in Type-1 errors (see McNeish, 2023, p. 200), then comparisons of EQT against conventional fit index cutpoints would be irrelevant. Conversely, invoking those same cutpoints to call PLF23 into question and evaluate EQT performance represents an "apples-to-oranges" comparison that is inconsistent with the theoretical logic behind and stated motivation of EQT.

MY2025 further object to the dichotomization procedure for EQT results interpretation used in PLF2023. To this we respond in three ways. First, the *identical* dichotomization classification strategy used by PLF2023 was also used in an additional EQT Monte Carlo simulation research design: McNeish (2023). Specifically, we quote directly from McNeish (2023, p. 205):

> "To make EQT and DFI more comparable, the Excellent and Close bins in EQT were collapsed because the substantive conclusion is similar for either bin. In Table 4, replications classified as having "Good" or "Close" fit are considered accurate classifications whereas replications classified as "Fair", "Mediocre", or "Poor" are considered inaccurate classifications."

Further, it is worth noting that, where the independent variable simulation condition was similar, PLF2023 and McNeish (2023) reached similar EQT conclusions (for example, compare McNeish, 2023, p. 204-206; with Peugh et al., 2023, p. 8-11 & Figure 1). However, a quick check of Marcoulides and Yuan (2025) shows that McNeish (2023) was neither mentioned nor cited. Second, applied researchers would likely agree that the ubiquitous question, "Does your analysis model fit your sample data?" has long implied a requisite binary response. As such, the use of dichotomization strategies has been a standard approach in Monte Carlo simulation evaluation of model fit procedures for decades (e.g., see Beauducel & Wittmann, 2005; Chen et al., 2008; Curran et al., 1996; Davey, 2005; Ding et al., 1995; Fan & Sivo, 2005, 2007; Gerbing & Anderson, 1993; Hu & Bentler, 1995, 1998, 1999; La Du & Tanaka, 1989; Marsh et al., 2004; Sugawara & MacCallum, 1993). Finally, Marcoulides and Yuan (2017, p. 151 & 152): (*italics added*) explicitly invoked dichotomous decision language where adjacent fit labels of "fair" and "close" were used as a boundary condition to assess the "tenability" of models and their fit:

> "The results provided by equivalence testing now *clearly indicate* that the model is in fact *not tenable*, but one that only achieves *fair* fit." (p. 151); "The results provided by equivalence testing also indicate that the model is *tenable* and that it achieves *close* fit." (p. 152).

Further, additional statements dichotomizing model fit conclusions based on the implicit EQT boundary condition described above can be found elsewhere (see Jiang et al.,

2017, p. 4; Marcoulides & Yuan, 2020, p.07; Marcoulides & Yuan, 2024, p.3426).

Finally, in support of their assertions, MY2025 offered a partial replication of a single PLF2023 Monte Carlo simulation condition (under normally distributed and no missing data conditions, more on the importance of this issue below). However, in addition to again invoking traditional fit indices as a comparison, their procedural description omitted several critical methodological details. Specifically, and in contrast to PLF2023 (see p. 5-6), MY2025 do not provide: 1) documentation of the software used for data generation, 2) 2) a clear rationale for why non-normal and missing data, ubiquitous in applied research, was not considered, 3) Monte Carlo simulation generated data accuracy verification via triangulation (data generation accuracy was confirmed in PLF2023 via three separate published methods; see Collins et al., 2001; Moshagen & Auerswald, 2018, p. 322; Muthén & Muthén, 2002; Paxton et al., 2001), or 4) reported quantitative evidence of simulation accuracy. Further, by not specifying the statistical software package used, MY2025 created uncertainty around their results because different software packages define and specify "baseline" models differently. To illustrate this point, Peugh and Feldon (2020, see p. 6 & footnote 6) demonstrated that the EQT CFIt statistic showed diametrically opposite model fit assessment statistics and conclusions for the same sample data and analysis model depending on whether EQT was used with `AMOS` or `Mplus` results, respectively (see also Sobel & Bohrnstedt, 1985; Widaman & Thompson, 2003). MY2025 shows no mention of the statistical software package used to generate their data, no quantitative evidence offered in support of the accuracy of their generated data, and no mention of how their statistical software package of choice defined and specified a "baseline" model. As such, MY2025's claims that their simulation replication results, "do not completely resemble" those of PLF2023 are difficult to validate or replicate.

What the Existing Evidence Does and Does Not Support Regarding EQT as a Fit Index

MY2025 state (p. 1-2): "Numerous empirical studies (e.g., Counsell et al., 2020; Finch & French, 2018; Marcoulides & Yuan, 2024; Peugh & Feldon, 2020; West et al., 2023; Yuan & Chan, 2016) have also confirmed the practical benefits of using the equivalence testing approach for the assessment of model fit in structural equation applications." However, a closer examination of the studies Marcoulides and Yuan cited in support of the continued use of EQT reveals a common denominator: When studies that generated normally distributed data without any missing values (i.e., Counsell et al., 2020; Finch & French, 2018; Montoya & Edwards, 2021; Yuan & Chan, 2016) were removed from consideration, the two remaining studies rely on the use of EQT for model fit judgments with input data based either on a model-reproduced covariance matrix ("The covariance matrices for the synthetic data set … were generated in Mplus"; Marcoulides & Yuan, 2020, p. 8) or data analytic model summary statistics ("For all models examined, the likelihood ratio χ^2 test statistic and its degrees of freedom (df), the CFI, the RMSEA, the sample size (n), the number of observed variables (p), and resulting statistical significance level were provided in the published articles"; Marcoulides & Yuan (2024, p. 3427)). Decades of methodological work have demonstrated that non-normality (e.g., Micceri, 1989 and subsequent publications) and missing data (e.g., Rubin, 1976; Peugh & Enders, 2004) are realities to be expected in empirical sample datasets, and the detrimental effects of both on model fit assessments have been well-established (e.g., Finch & French, 2018; Marcoulides & Yuan, 2020, 2024; Montoya & Edwards, 2021; Yuan & Chan, 2016). As shown in PLF2023, when non-normality, missing data, and uncertainty in model specification (i.e., correct vs. misspecified) are introduced—conditions that are well documented in applied structural equation modeling literature—EQT performance becomes both inaccurate and unreliable.

MY2025 do state that PLF2023's findings are consistent with prior work demonstrating that likelihood-based test statistics (e.g., χ^2) and their associated fit indices are sensitive to violations of distributional assumptions (Yuan & Bentler, 1999; Millsap, 2007, 2013b, 2013a). To be fair, MY2025 could hypothetically further assert that results from PLF2023 merely confirm limitations that have been acknowledged in earlier EQT publications (e.g., Marcoulides & Yuan, 2017; Yuan & Chan, 2016), and that EQT *could possibly* perform as intended if missing and non-normal data are essentially side-stepped by using model-reproduced covariance matrices (Marcoulides & Yuan, 2020) or data analysis model summary statistics (Marcoulides & Yuan, 2024) as EQT input data. However, as of this writing, thorough searches of Google Scholar and PubMed databases yielded no published Monte Carlo simulation studies that: 1) generated non-normal and missing data from known population structural equation models, 2) analyzed the generated data assuming either properly specified or misspecified models, 3) used the resulting model-reproduced covariance matrices or data analysis summary statistics as EQT input data for model fit assessment, or 4) showed EQT could accurately and reliably distinguish between properly specified or misspecified models based on the fit assessment results (but see Beribisky & Cribbie, 2025, Whittaker et al., 2026).

Conclusion

We assert that the point of inferential statistics in general—and model fit determinations in particular—is to permit unbiased inferences to be drawn from real-world sample data. Such data can be expected to be non-normally distributed, have missing values, and analyzed using a theoretical model which may be misspecified (despite a researcher's best efforts to perfectly model phenomena that are not yet fully known and understood). We do not argue with the mathematically sophisticated form of EQT in general or the ε_t statistic in particular. However, we do argue that existing evidence fails to support the assertion that EQT will provide valid and reliable fit assessments under common real-world research conditions. Because the fundamental goal of empirical research is to either support or reject hypothesized models of real-world phenomena (with the inherent uncertainties and inconveniences of real-world data), tools to support such goals should only be relied upon when their robustness to such conditions is rigorously empirically established. Based on results from both PLF2023 and McNeish (2023), and absent any published Monte Carlo simulation evidence of the accuracy and reliability of EQT when summary statistics or model-reproduced covariance matrices from real-world sample data are used as input, we find that MY2025's continued encouragement of EQT's routine use as a fit assessment function in applied research is currently lacking a rigorous empirical justification.

References

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. doi: 10.1207/s15328007sem1201_3.

Beribisky, N., & Cribbie, R. A. (2025). Equivalence testing based fit index: Standardized root mean squared residual. *Multivariate Behavioral Research*, *60*(1), 138–157. doi: 10.1080/00273171.2024.2386686.

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

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. doi: 10.1177/0049124108314720.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351. doi: 10.1037/1082-989X.6.4.330.

Counsell, A., Cribbie, R. A., & Flora, D. B. (2020). Evaluating equivalence testing methods for measurement invariance [Article 1633617]. *Multivariate Behavioral Research*, *2019*, 55. doi: 10.1080/00273171.

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. doi: 10.1037/1082-989X.1.1.16.

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

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*, 119–143. doi: 10.1080/10705519509540000.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling*, *12*, 343–367. doi: 10.1207/s15328007sem1203_1.

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

Finch, W. H., & French, B. F. (2018). A simulation investigation of the performance of invariance assessment using equivalence testing procedures. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 673–686. doi: 10.1080/10705511.2018.1431781.

Gerbing, D. W., & Anderson, J. C. (1993). Monte carlo evaluations of goodness-of-fit-indices for structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 40–65). Sage Publications.

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

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

Hu, L. T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). Sage Publications.

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

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

Jiang, G., Mai, Y., & Yuan, K. H. (2017). Advances in measurement invariance and mean comparison of latent variables: Equivalence testing and a projection-based approach. *Frontiers in Psychology*, *8*, 1823. doi: 10.3389/fpsyg.2017.01823.

Kenny, D. A. (2024, October 6). *Measuring model fit*. Retrieved October 6, 2024, from https://davidakenny.net/cm/fit.htm

La Du, T. J., & Tanaka, J. S. (1989). Influence of sample size, estimation method, and model specification on goodness-of-fit assessments in structural equation models. *Journal of Applied Psychology*, *74*(4), 625–635. doi: 10.1037/0021-9010.74.4.625.

Lee, T., & MacCallum, R. C. (2015). Parameter influence in structural equation modeling. *Structural Equation Modeling*, *22*, 102–114. doi: 10.1080/10705511.2014.935255.

Marcoulides, K. M., & Yuan, K. H. (2025). A note on comparing equivalence testing with conventional criteria in evaluating the overall fit of structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–6. doi: 10.1080/10705511.2025.2508735.

Marcoulides, K. M., & Yuan, K.-H. (2017). New ways to evaluate goodness of fit: A note on using equivalence testing to assess structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 148–153. doi: 10.1080/10705511.2016.1225260.

Marcoulides, K. M., & Yuan, K.-H. (2020). Using equivalence testing to evaluate goodness of fit in multilevel structural equation models. *International Journal of Research & Method in Education*, *43*, 431–443. doi: 10.1080/1743727X.2020.1795113.

Marcoulides, K. M., & Yuan, K.-H. (2024). Testing structural equation model fit in psychological studies: A replication study using equivalence testing. *Quality & Quantity*, *58*, 3417–3433. doi: 10.1007/s11135-023-01796-4.

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*, 320–341. doi: 10.1207/s15328007sem1103_2.

McDonald, J. A., & Clelland, D. A. (1984). Textile workers and union sentiment. *Social Forces*, *63*, 502–521. doi: 10.1093/sf/63.2.502.

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

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166. doi: 10.1037/0033-2909.105.1.156.

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

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

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

Montoya, A. K., & Edwards, M. C. (2021). The poor fit of model fit for selecting number of factors in exploratory factor analysis for scale evaluation. *Educational and Psychological Measurement*, *81*, 413–440. doi: 10.1177/0013164420942899.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. doi: 10.1002/sim.8086.

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

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*(4), 599–620. doi: 10.1207/S15328007SEM0904_8.

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

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

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*, 525–556. doi: 10.3102/00346543074004525.

Peugh, J. L., & 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*, es5. doi: 10.1187/cbe.20-01-0016.

Peugh, J. L., & Feldon, D. F. (2024). The answer is “no”: A comment on peugh and feldon (2020). *CBE—Life Sciences Education*, *23*(4), le1. doi: 10.1187/cbe.24-07-0182.

Peugh, J. L., Litson, K., & Feldon, D. F. (2023). Equivalence testing to judge model fit: A monte carlo simulation. *Psychological Methods*, *30*(4), 888–925. doi: 10.1037/met0000591.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. doi: 10.1093/biomet/63.3.581.

Sobel, M. E., & Bohrnstedt, G. W. (1985). Use of null models in evaluating the fit of covariance structure models. In N. B. Tuma (Ed.), *Sociological methodology* (pp. 152–178). Jossey-Bass.

Sugawara, H. M., & MacCallum, R. C. (1993). Effect of estimation method on incremental fit indexes for covariance structure models. *Applied Psychological Measurement*, *17*(4), 365–377. doi: 10.1177/014662169301700405.

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

Whittaker, T. A., Khojasteh, J., & Park, S. H. (2026). The performance of equivalence-based fit indices with bootstrapped confidence intervals and varied tolerable cutoffs. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–13. doi: 10.1080/10705511.2026.2686107.

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, *8*, 16–37.

Yuan, K. H., & Chan, W. (2016). Measurement invariance via multigroup sem: Issues and solutions with chi-square-difference tests. *Psychological Methods*, *21*, 405–426. doi: 10.1037/met0000080.

Yuan, K.-H., & Bentler, P. M. (1999). On normal theory and associated test statistics in covariance structure analysis under two classes of nonnormal distributions. *Statistica Sinica*, *9*, 831–853.

Citation

Peugh, J. L., Litson, K., & Feldon, D. (2026). An objective evaluation of equivalence testing for model fit: In response to Marcoulides and Yuan (2025). *The Quantitative Methods for Psychology*, *22*(2), 71–77. doi: 10.20982/tqmp.22.2.p071.

Received: 29/05/2026 ~ Accepted: 30/07/2026