Sociological Methods & Research
Volume 36 Number 4
May 2008 462-494

10.1177/0049124108314720
http://smr.sagepub.com
hosted at
http://online.sagepub.com

An Empirical Evaluation of the Use of Fixed Cutoff Points in RMSEA Test Statistic in Structural Equation Models

Feinian Chen
North Carolina State University, Raleigh, NC
Patrick J. Curran
Kenneth A. Bollen
University of North Carolina at Chapel Hill
James Kirby
Agency for Healthcare Research & Quality, Rockville, MD
Pamela Paxton
Ohio State University, Columbus

This article is an empirical evaluation of the choice of fixed cutoff points in assessing the root mean square error of approximation (RMSEA) test statistic as a measure of goodness-of-fit in Structural Equation Models. Using simulation data, the authors first examine whether there is any empirical evidence for the use of a universal cutoff, and then compare the practice of using the point estimate of the RMSEA alone versus that of using it jointly with its related confidence interval. The results of the study demonstrate that there is little empirical support for the use of .05 or any other value as universal cutoff values to determine adequate model fit, regardless of whether the point estimate is used alone or jointly with the confidence interval. The authors' analyses suggest that to achieve a certain level of power or Type I error rate, the choice of cutoff values depends on model specifications, degrees of freedom, and sample size.

Keywords: *RMSEA; SEM; goodness-of-fit; computer simulations*

Structural Equation Modeling (SEM) has been widely used in sociological, psychological, and social science research. One of the appealing attributes of SEM is that it allows for tests of theoretically derived models

against empirical data. For researchers using SEM techniques, evaluation of the fit of a hypothesized model to sample data is crucial to the analysis. A key feature of SEM is the test of the null hypothesis of $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$, also known as the test of exact fit, where $\boldsymbol{\Sigma}$ is the population covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the covariance matrix implied by a specific model, and $\boldsymbol{\theta}$ is a vector of free parameters defined by the model. The model test statistic T enables an asymptotic test of the null hypothesis of H_0: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$. A significant T, often reported as the model chi-square, would suggest misspecification of the model. However, such a test of exact fit of the proposed model is generally unrealistic, as hardly any model using real data is without error (e.g., Browne and Cudeck 1993). A trivial misspecification, particularly with large sample sizes, can lead to rejection of the model even when it may otherwise adequately reproduce the population covariance matrix.

As a result, a variety of goodness-of-fit measures was developed to augment the T statistic. Steiger, Shapiro, and Browne (1985) demonstrated that the T statistic does not follow a central χ^2 distribution under misspecification. Instead, it follows a noncentral χ^2 distribution, with the noncentrality parameter λ (estimated by $T - df$) denotes the degree of misfit in the model. Several baseline fit indices make use of the noncentrality parameter (e.g., Tucker-Lewis index [TLI], relative noncentrality index [RNI], comparative fit index [CFI]) (Bentler 1990, Goffin 1993), which are essentially comparisons between λ_T and λ_B, with λ_T measuring the amount of misfit of the target model and λ_B that of the baseline model. However, such measures are heavily dependent on the baseline null model. In addition, these measures were found to be particularly susceptible to the influence of estimation methods (Sugawara and MacCallum 1993) and do not utilize the feature that the known distribution of the T statistics actually allows for hypothesis testing through construction of confidence intervals around λ.

Originally presented by Steiger and Lind (1980) and popularized by Browne and Cudeck (1993), the root mean square error of approximation (RMSEA) measure is closely tied to the noncentrality parameter λ, which is estimated in a sample as $\hat{\lambda} = T - df$, reflecting the degree of misfit in the proposed model. If $T - df$ is less than zero, then $\hat{\lambda}$ is set to zero. The estimate of RMSEA ($\hat{\varepsilon}$) uses $\hat{\lambda}$ and is given as follows:

$$\hat{\varepsilon} = \sqrt{\max\left(0, \frac{\hat{\lambda}}{df(N-1)}\right)}.$$

It ranges from zero to positive infinity, with a value of zero indicating exact model fit, and larger values reflecting poorer model fit.

A key advantage of the RMSEA is that confidence intervals can be constructed around the point estimate because the RMSEA asymptotically follows a rescaled noncentral χ^2 distribution for a given sample size, degrees of freedom, and noncentrality parameter λ. The confidence interval is as follows:

$$CI = \left(\sqrt{\frac{\hat{\lambda}_L}{df(N-1)}}; \sqrt{\frac{\hat{\lambda}_U}{df(N-1)}} \right),$$

where $\hat{\lambda}_L$ and $\hat{\lambda}_U$ are the specific lower and upper values that define the limits of the desired interval (Browne and Cudeck 1993, Equation 14).

Researchers can use the RMSEA in two ways to assess model fit. The first is simply to examine the point estimate and to compare it with an arbitrary fixed cutoff point. The second is to conduct a more formal hypothesis test, by jointly considering the point estimate and its associated confidence interval. There are three such types of hypothesis tests available (MacCallum, Browne, and Sugawara 1996). The first type of test is the test of exact fit, with the null hypothesis as $H_0: \varepsilon = 0$, where ε is the population value of the RMSEA. The null hypothesis is rejected if the lower CI (confidence interval) is greater than zero. This corresponds to the standard χ^2 test given above. The second type of test is the test of close fit, with the null hypothesis being $H_0: \varepsilon \leq c$, where c is an arbitrary constant. The null hypothesis is rejected if the test statistic exceeds a cutoff value c that defines an area α in the upper tail of the noncentral chi-square distribution (i.e., if the lower CI is greater than c). Retention of the null hypothesis supports the proposed model, while rejection of the null suggests a poor fit of the model. The test of close fit is sometimes considered more realistic than the test of exact fit (MacCallum et al. 1996). MacCallum et al. (1996) proposed a third type of test, the test of not close fit, with the null hypothesis as $H_0: \varepsilon \geq c$. They contended that it was not easy to argue for support of a model with either the test of exact fit or the test of close fit, because failure to reject the null hypothesis (indicating a good model fit) merely suggested the absence of strong evidence against it. In this test, the null hypothesis is rejected if the test statistic is below a value that cuts off an area α in the lower tail of the noncentral chi-square distribution (i.e., if the upper CI is less than or equal to the arbitrary constant c).

Whether the researcher uses the point estimate alone or adopts the hypothesis testing framework by jointly considering the point estimate and its related CI, choosing the optimal cutoff point (c) is of utmost importance in the success of the RMSEA as a measure for goodness-of-fit.

Browne and Cudeck (1993:144) recommended that "a value of the RMSEA of about 0.05 or less would indicate a close fit of the model in relation to the degrees of freedom," and that "the value of about 0.08 or less for the RMSEA would indicate a reasonable error of approximation and would not want to employ a model with a RMSEA greater than 0.1." Similar guidelines were recommended by Steiger (1989). However, both Browne and Cudeck (1993) and Steiger (1989) warned the researchers that these cutoff points were subjective measures based on their substantial amount of experience. Similarly, in their power analysis of different hypothesis tests using the RMSEA, MacCallum et al. (1996) used 0.01, 0.05, and 0.08 to indicate excellent, good, and mediocre fit respectively but clearly emphasized the arbitrariness in the choice of cutoff points. Hu and Bentler (1999) recommended the test of RMSEA > .05 (or .06) as one of the alternative tests for detecting model misspecification, although they noted the test tended to over-reject at small sample size. Marsh, Hau, and Wen (2004) further cautioned researchers about using the cutoff criteria provided by Hu and Bentler (1999) as "golden rules of thumb," particularly due to their limited generalizability to mildly misspecified models. Other researchers echoed the point by suggesting that the use of precise numerical cutoff points for RMSEA should not be taken too seriously (Hayduk and Glaser 2000, Steiger 2000).

Nonetheless, the cutoff point of 0.05 has been widely adopted as the "gold standard" in applied research settings. Indeed, various SEM computer programs such as LISREL, AMOS, Mplus, and PROC CALIS (SAS) now offer a test of close fit on the probability of $\varepsilon \leq 0.05$. In addition, despite the known imprecision in using the point estimate alone, it is a popular measure of fit widely adopted by the researchers. How reasonable are these current practices? It is against this backdrop that we conduct the current study. We contend that an empirical evaluation of the choice of fixed cutoff points is essential in the assessment of the success of the RMSEA as a measure of goodness-of-fit. Using data from a large simulation experiment, we first examine whether there is any empirical evidence for the use of a universal cutoff, whether it be 0.05 or any other value. We want to stress that our goal is *not* to develop a new recommended cutoff point; instead, we wish to highlight ranges of values that can be consulted in practice and to provide empirically based information to applied research for the valid and thoughtful use of the RMSEA in practice.

We also examine whether the limitations of using a point estimate are overcome by considering the CI in addition to the fixed cutoff point.

While the theoretical imprecision in using a point estimate alone has been well argued by Browne and Cudeck (1993) and MacCallum et al. (1996), the attraction for researchers can be easily understood because of its parsimony. Alternatively, using the point estimate and its CI is more complicated. For example, one almost always must consider the power of the test within the framework of hypothesis testing, thus making it necessary to take into account sample size, degrees of freedom and other model characteristics (see MacCallum et al. [1996], Nevitt and Hancock [2000], and Hancock and Freeman [2001] for discussions on the power assessment of different tests as well as recommendations for applied researchers). Hence, it is extremely useful for the applied researchers to know empirically the extent of difference between these two approaches in terms of their success in model fit assessment. By directly comparing these two practices, we hope to provide some practical guidance to applied SEM researchers for the optimal use of this fit statistic.

There are a number of well-designed Monte Carlo simulation studies examining the performance of SEM fit indices, including the RMSEA (Hu and Bentler 1998; Fan, Thompson, and Wang 1999; Kenny and McCoach 2003; Nasser and Wisenbaker 2003). However, much attention was paid to finite sampling behavior of the point estimate but not to the corresponding CI or the use of RMSEA in hypothesis testing. An exception is a study by Nevitt and Hancock (2000), which compared the rejection rates of tests of exact fit, close fit, and not close fit using the RMSEA for nonnormal conditions in SEM. While the study provided important information on comparisons of the three hypothesis tests using the RMSEA, our study departs from it in several major ways.

First, Nevitt and Hancock (2000) used the critical value of 0.05 throughout their hypothesis tests. As we argued earlier, we consider this specific cutoff value to be arbitrary and in need of further empirical investigation. In particular, we are interested in how the choice of the cutoff point c affects the performance of the RMSEA, whether it is used as a point estimate alone or used jointly with its related CI. Second, Nevitt and Hancock (2000) used an oblique confirmatory factor analysis model (CFA) as the base underlying population model in their Monte Carlo study. To expand on the external validity, we incorporate a range of general SEM model types that are commonly used in social science research in our simulation study. Third, Nevitt and Hancock (2000) considered one properly specified and one misspecified model. We are interested in how the degree of misspecification can affect the power of the tests. Thus, we studied three types of properly specified models and nine types of

misspecified models. In addition, as Nevitt and Hancock (2000) acknowledged, it was difficult to evaluate the hypothesis tests for misspecified models under nonnormal conditions because the true lack of fit in the population for the misspecified models is due to both misspecification and nonnormality. To avoid the confounding effects of nonnormality, we generate our variables from a multivariate normal distribution. The violation of normality assumption is obviously an important question and occurs often in research, but this is beyond the scope of the current study.

Most important, it is not our goal to directly compare the performance of the test of close fit and not close fit. We are interested in comparing the practice of using the point estimate *alone* versus that of using the point estimate *jointly* with its related CI, when a fixed cutoff point is used in the test. We believe that it is unnecessarily confusing to compare the performance of the test of close fit and not close fit when we are considering both properly specified and misspecified models in the analysis. The meaning of model rejection is the opposite for the test of close fit and not close fit, with the former suggesting a good fit of the model and the latter indicating a poor fit of the model. In addition, the test of close fit is the one that is readily available through various SEM packages, thus making the investigation most relevant to practical researchers. So, in this article, we propose two tests that make use of the CI in a consistent way, using the criteria of lower bound of CI ≤ 0.05 and upper bound of CI ≤ 0.1 as two candidate cutoff values. Rejection of the model thus suggests a poor fit in both tests.

We also want to make a note that the focus of the article is exclusively on how well sample estimates of the RMSEA and its related CIs perform in applied social science research settings. Although the sampling distributions of the RMSEA are known asymptotically, the assumptions of no excess multivariate kurtosis, adequate sample size, and errors of approximations being not "great" relative to errors of estimation are often violated in applied research. Therefore, it is critical to understand the sampling characteristics of the RMSEA point estimates and CIs when the conditions are not met. Indeed, the success of the RMSEA as a measure for model fit also depends on whether the test statistic T indeed follows a noncentral chi-square distribution. Recent research shows that the noncentral χ^2 approximation is conditioned on factors such as sample sizes, degrees of freedom, distribution of the variables, and the degree of misspecification (Olsson, Foss, and Breivik 2004, Yuan 2005; Yuan, Hayashi, and Bentler 2007). Findings from our own research team have indicated that the noncentral chi-square distribution is generally well approximated and that the sample RMSEA values and CIs appear to be unbiased estimates of the corresponding

population values, at least for models with small to moderate misspecification, and when the sample size is reasonably large (Curran et al. 2002; Curran et al. 2003). Future Monte Carlo studies are needed to specifically investigate the conditions under which the noncentral chi-square distributions are not followed.

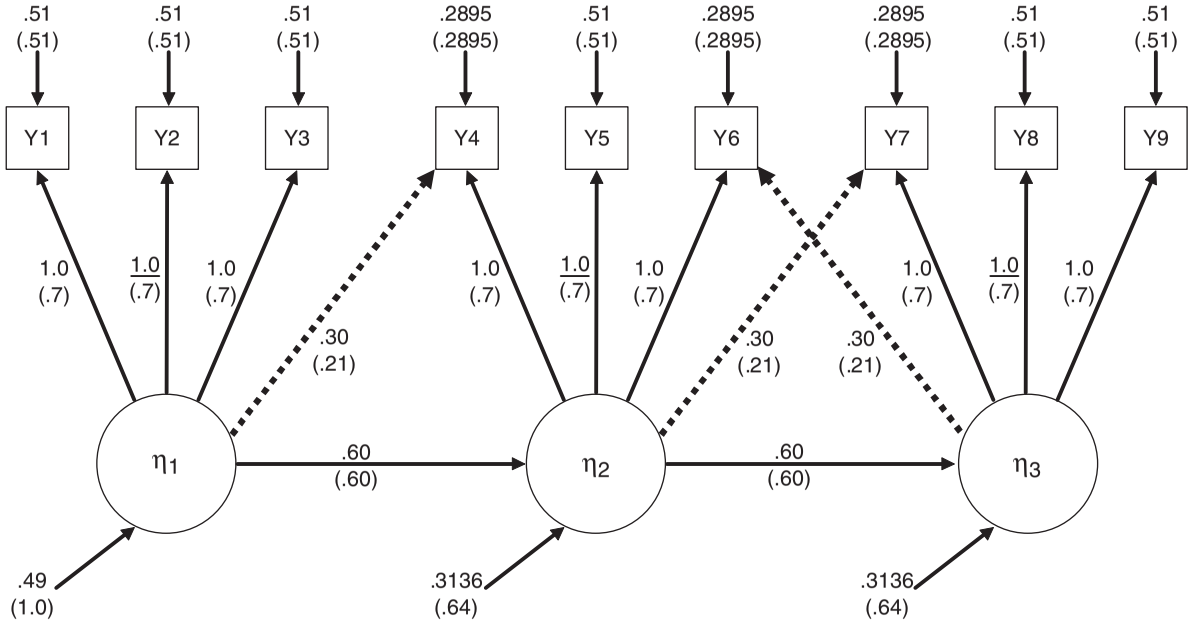
Method

Model Types and Experimental Conditions

We reviewed five years of key journals within several areas of social science research to catalog prototypical model types of SEM applications (see Paxton et al. [2001], Curran et al. [2002], and Chen et al. [2001], for further details for a comprehensive review of our research design). Using this information in combination with our own modeling experience, we carefully selected three general model types that represent features that are commonly encountered in social science research: Model 1 (see Figure 1) contains 9 measured variables and three latent factors with three to four indicators per factor, Model 2 (see Figure 2) has 15 measured variables and three latent factors with 4 to six indicators per factor, and Model 3 (see Figure 3) consists of 13 measured variables with the same form as Model 1 but with the addition of 4 observed and correlated exogenous variables.[1] Furthermore, for each model we use one correct and three incorrect specifications, resulting in a total of 12 individual models.

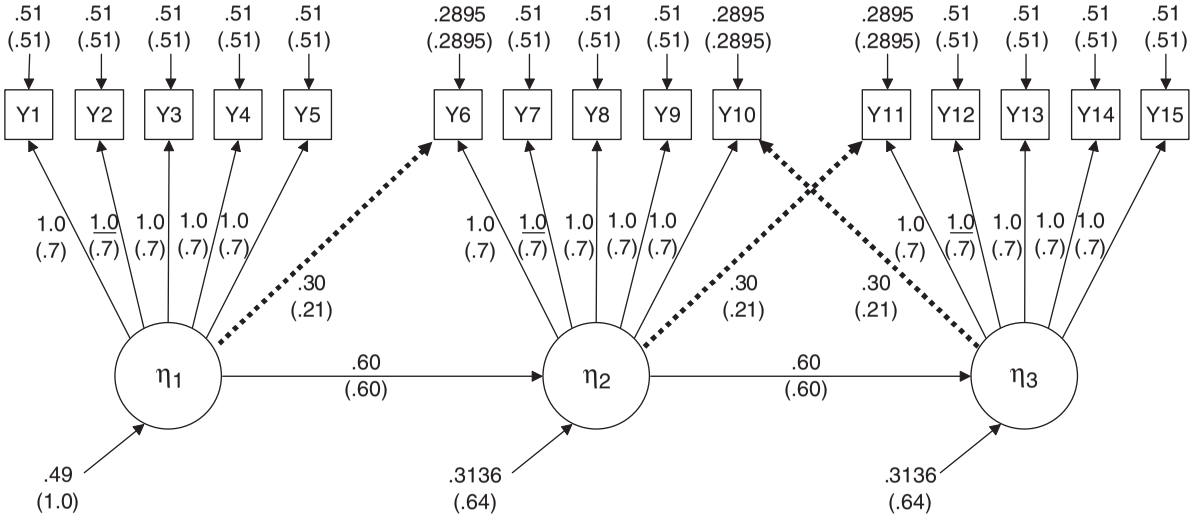
For Model 1, the model with the smallest misspecification omits the complex loading linking item 7 with Factor 2; the model with moderate misspecification additionally omits the complex loading linking item 6 with Factor 3; the model with the largest misspecification additionally removes the complex loading linking item 4 with Factor 1. For Model 2, we first omit the complex loading linking item 11 with Factor 2, then omit the complex loading linking item 10 with Factor 3, and finally remove the complex loading linking item 6 with Factor 1. For Model 3, the degree of misspecification changes in the following order: The model with the smallest misspecification *jointly* omits the set of three complex factor loadings (item 7 with Factor 2, item 6 with Factor 3, and item 4 with Factor 1); the model with moderate misspecification omits the set of four regression parameters (Factor 2 regressed on predictor 1, Factor 3 regressed on predictor 1, Factor 2 regressed on predictor 3, and Factor 3 regressed on predictor 3); the model with the largest misspecification omits the set of three factor loadings *and* the set of four regression parameters).

Figure 1
Target Population Model 1



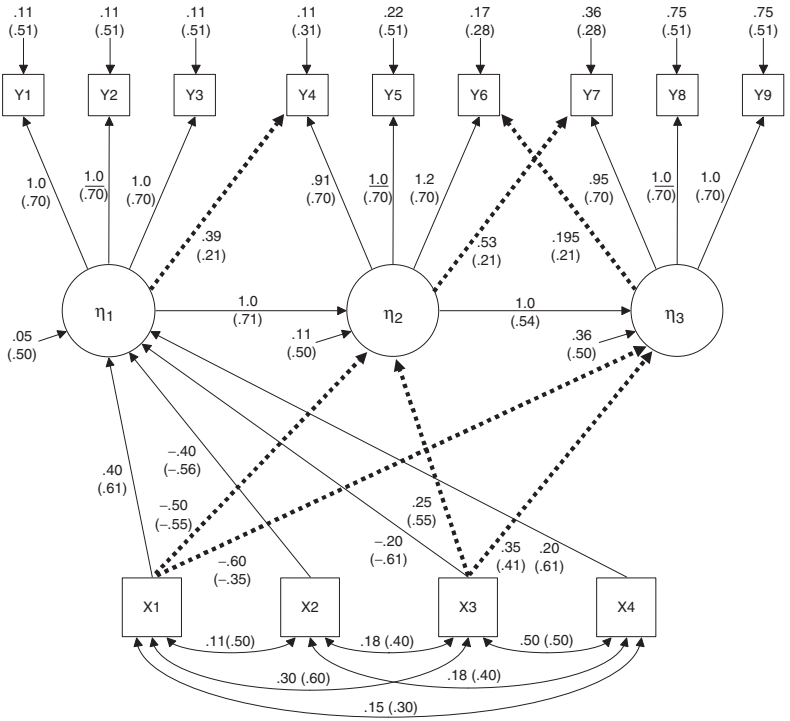
Note: Numbers shown are unstandardized parameter values with standardized values in parentheses; solid and dashed lines represent the population model structure, and dashed lines represent omitted parameters under model misspecification.

Figure 2
Target Population Model 2



Note: Numbers shown are unstandardized parameter values with standardized values in parentheses; solid and dashed lines represent the population model structure, and dashed lines represent omitted parameters under model misspecification.

Figure 3
Target Population Model 3



Note: Numbers shown are unstandardized parameter values with standardized values in parentheses; solid and dashed lines represent the population model structure, and dashed lines represent omitted parameters under model misspecification.

Model parameterization. For all three model types, parameter values were carefully selected to result in a range of effect sizes (e.g., communalities and R^2 values ranging from 49 percent to 72 percent), and for the misspecified conditions to lead to both a wide range of power to detect the misspecifications (e.g., power ranging from .07 to 1.0 across all sample sizes) and a range of bias in parameter estimates (e.g., absolute bias ranging from 0 percent to 37 percent).[2] We believe this parameterization reflects values commonly encountered in applied research and that the omission of one or more parameters would result in meaningful impacts on parameter estimation and overall model fit.

Sample size. We chose seven sample sizes to represent those commonly encountered in applied research and these range from very small to large: 50, 75, 100, 200, 400, 800, and 1,000.

Data generation and estimation. We used the simulation feature in Version 5 of EQS (Bentler 1995) to generate the raw data and EQS's maximum likelihood estimation to fit the sample models. Population values for each parameter were used as initial start values, and a maximum of 100 iterations was allowed to achieve convergence.

Distribution. We generated data from a multivariate normal distribution.

Replications. There were a total of 84 experimental conditions (three models, four specifications, and seven sample sizes), and we generated up to 500 replications for each condition (see below discussion for exceptions).

Convergence. We eliminated any replication that failed to converge within 100 iterations or did converge but resulted in an out-of-bounds parameter estimate (e.g., "Heywood Case") or a linear dependency among parameters. We adopted this strategy because the research hypotheses were directly related to proper solutions in SEM, and the external validity of findings would be threatened with the inclusion of improper solutions (see Chen et al. [2001] for detailed discussion).[3] To maintain 500 replications per condition, we generated an initial set of up to 650 replications. We then fit the models to the generated data and selected the first 500 proper solutions, or selected as many proper solutions as existed when the total number of replications was reached. This resulted in 500 proper solutions for all properly specified and most misspecified experimental conditions, but there were several misspecified conditions that resulted in fewer than 500 proper solutions.[4]

Outcome measures. The outcome measures studied here are the RMSEA point estimates and their associated 90 percent CIs. These values were computed in SAS Version 8 using the computational formulae presented in Browne and Cudeck (1993) based on the maximum likelihood fit function minima calculated by EQS.

Analytic Plan

The goal of the project is to empirically evaluate the performance of the RMSEA point estimate as well as its joint use with its related CI in assessing model fit. We address three specific research questions: (a) Is

there any empirical evidence supporting the use of 0.05 or 0.10 as universal cutoff points? (b) Is there empirical evidence for the use of *any* universal cutoff values? (c) Does the joint use of the point estimate with its associated CI improve model assessment relative to the use of the point estimate alone? To empirically evaluate these questions, we start the analysis with a simple examination of the proportion of models rejected using 0.05 or 0.10 cutoff values of the RMSEA point estimate across different model types, sample sizes, and model specifications. We hypothesize that these cutoff points are arbitrary and expect that the performance of the test will vary by sample sizes, degrees of freedom, and model specifications. Next, we depart from the common approach by examining model rejection rates when the cutoff points used in the test range from 0 to 0.15 with an increment of 0.005. The purpose of this element of the analysis is *not* to propose a new set of cutoff points, but rather to systematically explore the sensitivity of the RMSEA test statistic to the choice of varying cutoff values. To maintain equal model rejection rates, we predict that varying "correct" cutoff points are necessary for models with different experimental characteristics.

Finally, we examine model rejection rates based on the RMSEA point estimate and its related CI by using the lower and upper bound of the CI in combination with 0.05 and 0.10 as the cutoff values. Although it is widely recognized that the joint use of the point estimate and the CIs provides a more informative view of the RMSEA statistic, we illustrate limitations of this approach when universal cutoff points are applied. Specifically, we utilize two tests: (a) a given model is rejected when the *lower bound* of the CI is *greater than* 0.05 and (b) a given model is rejected when the *upper bound* of the CI is *greater than* 0.10. The first test corresponds to a test of close fit. The second test does *not* correspond to a test of not close fit, which would have rejected a model if the upper bound of the CI is less than or equal to a cutoff point. However, the two tests we consider are consistent in that the rejection of a model indicates a poor fit.

Results

Throughout the analysis, we use the model rejection rate, or proportion of the models rejected when the point estimate or the lower/upper bound of its related CI is larger than the specified cutoff point, as an empirical

gauge of the performance of the tests. However, it is worth noting that the meaning of the model rejection rate varies from one specification to the other. When a model is rejected, it implies that the model does not achieve acceptable RMSEA model fit. Thus, for a properly specified model, rejection of the model is incorrect and the empirical rejection rates reflect Type I error. For an improperly specified model, rejection of the model is correct and the empirical rejection rates reflect the power of the test.[5]

Test Using the Point Estimate and Fixed Cutoff Value

We begin our description of the analytical results with Table 1, where we present the proportion of models rejected based on the RMSEA point estimate, using 0.05 and 0.10 as the cutoff point, respectively. As expected, the results clearly reflect that the performance of the tests vary tremendously by sample sizes and model specification. First, for correctly specified models (with a population RMSEA = 0), the test of whether the RMSEA point estimate is $\leq$0.05 only works well for larger samples ($n \geq 200$), where the rejection rates are below 10 percent across different model types. The model rejection rate with this cutoff virtually approaches zero when $n = 800$ and $n = 1{,}000$. We also observe differences in rejection rates by model types. For example, for a correctly specified Model 1, with $n = 200$, the model rejection rate is 8 percent, while for a correctly specified Model 2, the model rejection rate is 0.8 percent, a tenfold difference. This is striking, considering that these two models are identical except that an extra indicator is added to each factor in Model 2. For smaller samples, the cutoff of 0.05 is too conservative. For example, for a correctly specified Model 3, with $n = 50$, almost half of the models are *incorrectly* rejected. As expected, the rejection rate declines as sample sizes increase. For example, for a correctly specified Model 2, the model rejection rate is as high as 29.4 percent with $n = 75$, but quickly declines to 16.6 percent with $n = 100$ and less than 1 percent with $n = 200$.

Similarly, for incorrectly specified models, using 0.05 as the cutoff does not work consistently well across different models and sample sizes. As we noted earlier, the rejection rates for misspecified models reflect the statistical power of the test. However, comparing results across models with different levels of misspecifications, it is clear that rejection rates are generally quite low except in the case of Model 3 with moderate and the largest misspecification, where the rejection rates are above 90 percent

Table 1
Proportion of Models Rejected Based on RMSEA Point Estimate

			RMSEA ≤ 0.05 (null)							RMSEA ≤ 0.1 (null)				
			N							*N*				
Model	Population RMSEA	*df*	50	75	100	200	400	800	1,000	50	75	100	200	400
Model 1														
Correct specification	0.000	22	0.388	0.310	0.260	0.078	0.002	0.000	0.000	0.112	0.020	0.010	0.000	0.000
Smallest misspecification	0.027	23	0.431	0.345	0.349	0.164	0.040	0.004	0.002	0.140	0.028	0.016	0.000	0.000
Moderate misspecification	0.040	24	0.486	0.421	0.446	0.326	0.214	0.100	0.056	0.143	0.035	0.018	0.000	0.000
Severest misspecification	0.061	25	0.640	0.598	0.651	0.688	0.818	0.916	0.936	0.236	0.110	0.089	0.012	0.000
Model 2														
Correct specification	0.000	85	0.578	0.294	0.166	0.008	0.000	0.000	0.000	0.032	0.000	0.000	0.000	0.000
Smallest misspecification	0.021	86	0.629	0.362	0.220	0.034	0.000	0.000	0.000	0.044	0.000	0.000	0.000	0.000
Moderate misspecification	0.031	87	0.659	0.428	0.310	0.084	0.002	0.000	0.000	0.058	0.000	0.000	0.000	0.000
Severest misspecification	0.040	88	0.739	0.540	0.440	0.204	0.078	0.006	0.000	0.088	0.004	0.000	0.000	0.000
Model 3														
Correct specification	0.000	50	0.498	0.328	0.162	0.024	0.000	0.000	0.000	0.090	0.008	0.000	0.000	0.000
Smallest misspecification	0.049	53	0.700	0.587	0.550	0.504	0.406	0.422	0.392	0.192	0.030	0.006	0.000	0.000
Moderate misspecification	0.084	54	0.936	0.956	0.945	0.996	1.000	1.000	1.000	0.524	0.317	0.204	0.078	0.008
Severest misspecification	0.097	57	0.964	0.988	0.984	1.000	1.000	1.000	1.000	0.666	0.515	0.466	0.404	0.288

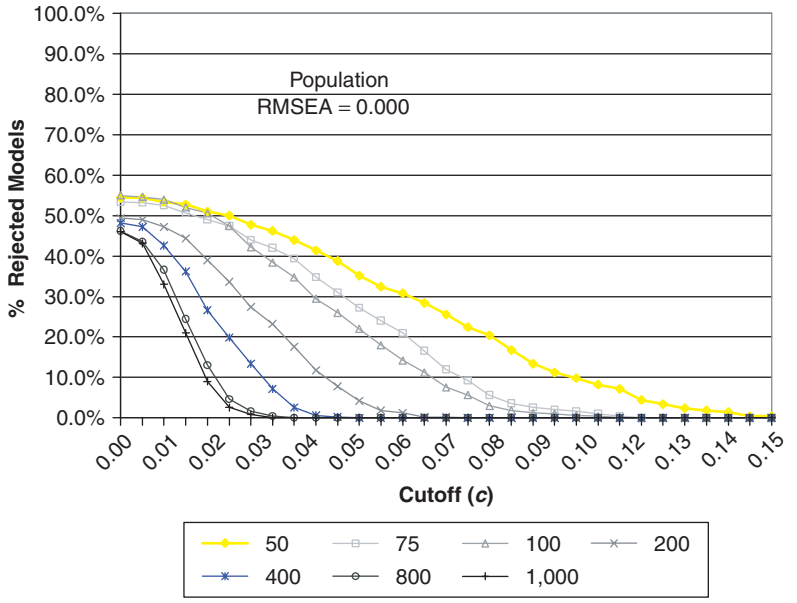
Note: RMSEA = root mean square error of approximation.

across all sample sizes. This is not surprising, given that the population RMSEA values for these two model specifications are 0.084 and 0.097, respectively. For all other misspecified models, particularly those where the population RMSEA are below 0.05, the rejection rates are expectedly and strikingly low, given that the RMSEA test in conjunction with nonzero cutoffs is designed to overlook some degrees of error in the model structure. For example, for Model 2, the population RMSEA for the model with the smallest misspecification is 0.02 and that for the model with moderate misspecification is 0.03. As a result, we observe *higher* rejection rates with *decreasing* sample sizes (e.g., rejection rates between 60 percent–70 percent with $n = 50$), and *lower* rejection rates with *increasing* sample sizes (rejection rates between 20 percent–40 percent at $n = 100$), and rejection rates converging to zero at sample sizes of 800 and above. Even in Model 2 with the largest misspecification, almost all models are incorrectly accepted, at $n = 800$ and 1,000. Although this finding is predictable (i.e., given that the population RMSEA is less than 0.05), the power is zero to detect the most misspecified version of Model 2 at the largest sample sizes. It is worth noting that the degrees of misspecification in these models are not trivial. For example, for Model 2 with moderate misspecification, the biases in many of the estimated parameters are as high as 35 percent, yet 100 percent of these models are deemed acceptable using the .05 criterion.

The above findings also showed that 0.10 is too liberal to be used as the cutoff point in the test. For all the model specifications we examined in this study, the population RMSEA are all less than 0.1. As expected, all correctly specified models, regardless of sample size, have very low rejection rates (i.e., low Type I error rates). However, for the misspecified models, the rejection rates are also exceedingly low, suggesting low power of the test. For example, for the most misspecified condition of Model 1, the model rejection rate is 23.6 percent with $n = 50$ and quickly diminishes to a low 1.2 percent with $n = 200$. Even for a severely misspecified Model 4, where the population RMSEA $= 0.097$, the model rejection rate ranges from 25.2 percent to 66.6 percent across all sample sizes.

The above results clearly demonstrate that there is no empirical support for the use of 0.05 or 0.10 as universal cutoff values to determine adequate model fit. The means of the sampling distributions of the RMSEA are related to the size of the sample, the type of the model, and the degree of misspecification (see also Curran et al. 2003). As such, maintaining a fixed cutoff for all models and all sample sizes will lead to different decisions regarding model fit that varies as a direct function of sample size and model type.

Figure 4
Model Rejection Rates by Sample Size,
H_0: RMSEA $\leq c$ Model 1, Correct Specification

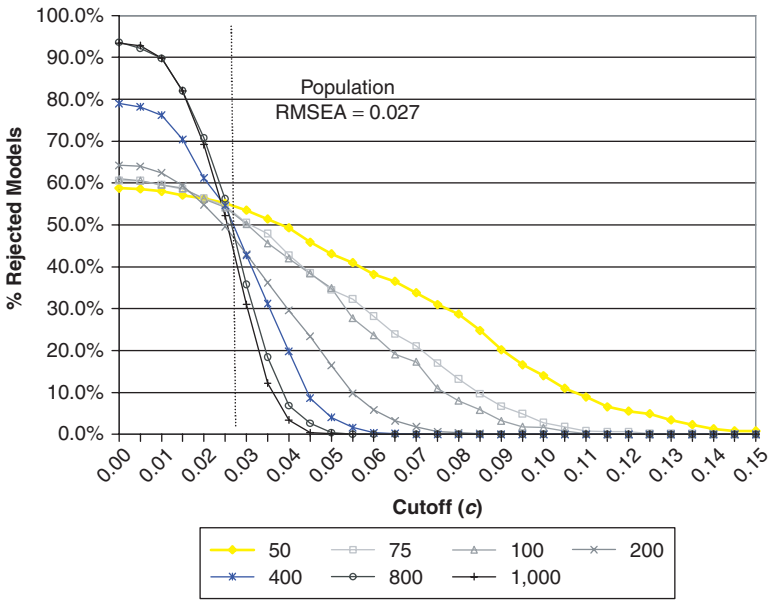


Note: RMSEA = root mean error of approximation.

We next examine model rejection rates across different model specifications and samples sizes, using RMSEA cutoff points ranging from 0 to 0.15, with an increment of 0.005. To facilitate presentation, these results are presented in a series of graphs in Figures 4 to 15.

Figure 4 shows that for a correctly specified Model 1, model rejection rates range from 46 percent to 54 percent across sample sizes when $c = 0$. Given that this is a properly specified model and that the rejection of the model is incorrect, this suggests a very high Type I error rate. As the value of the cutoff point increases, the model rejection rates decrease. The rate of decline is sharper for larger samples than smaller samples. For example, with $n = 1{,}000$, the model rejection rate declines below 10 percent when c reaches 0.02. For smaller samples, changes in model rejection rates are much slower with every increment in the cutoff value. With $n = 100$, the

Figure 5
Model Rejection Rates by Sample Size,
H_0: RMSEA $\leq c$ Model 1, Smallest Misspecification

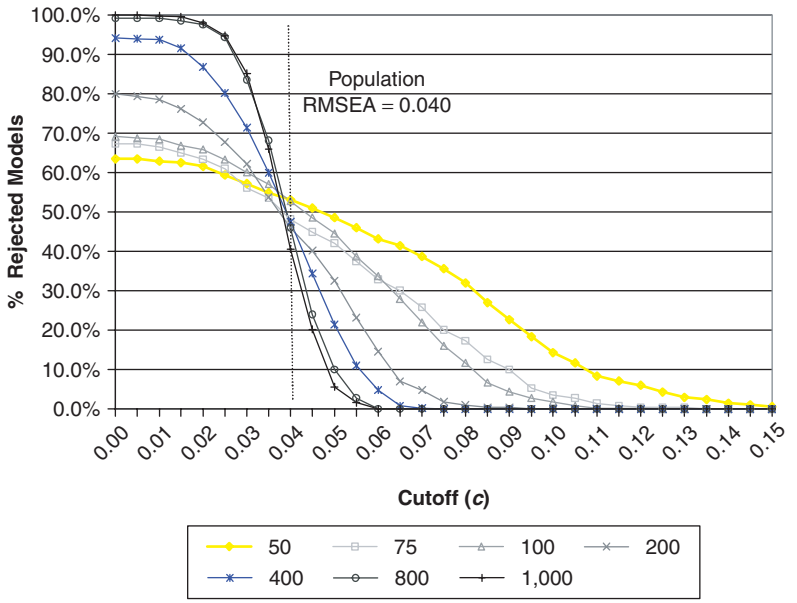


Note: RMSEA = root mean error of approximation.

model rejection rate does not reach below 10 percent until c is below 0.07. Overall, it seems that the test performs well (i.e., low Type I error rate) when c is reasonably large, at least for larger sample sizes.

When we move to a Model 1 with the smallest misspecification, as expected the model rejection rate declines as the cutoff increases (see Figure 5). However, the implication is different because the model is misspecified. Some researchers might find this level of misspecification as acceptable while others would not. For those in the latter group, the model rejection rate can be regarded as the power of the test. For larger sample sizes, the model rejection rate starts high at $c = 0$ but quickly diminishes with every small increment of the critical value, and the rate of decrease varies with sample size. For example, with $c = 0$, the model rejection rate for $n = 1{,}000$ is as high as 93 percent, but drops to a low 3 percent with $c = 0.04$, and almost approaches zero with $c = 0.05$. In contrast, for smaller

Figure 6
Model Rejection Rates by Sample Size,
H_0: RMSEA $\leq c$ Model 1, Moderate Misspecification

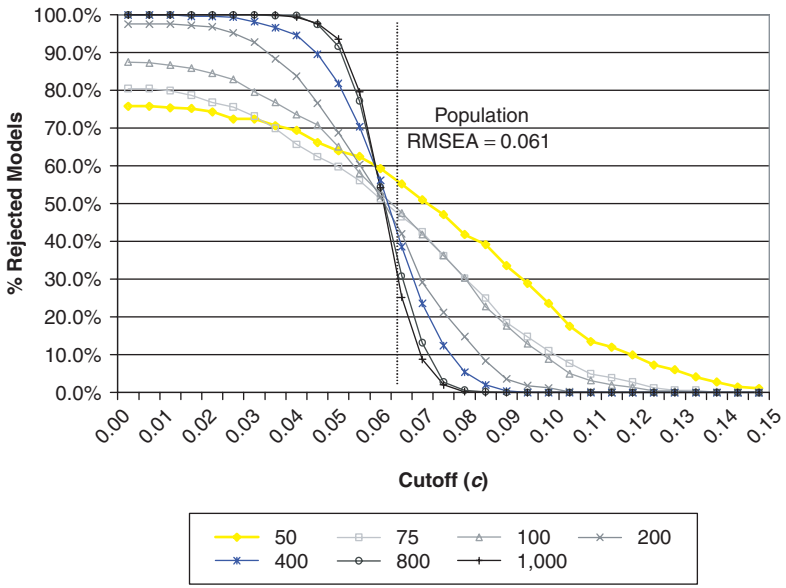


Note: RMSEA = root mean error of approximation.

samples, the power curve starts at a much lower level and declines gradually with increments of the cutoff. For example, the model rejection rate for $n = 100$ at $c = 0$ is 61 percent, and it declines to roughly 38 percent at $c = 0.05$. Interestingly, when $c < 0.027$ (the population RMSEA), the larger sample sizes have higher model rejection rates; but when $c > 0.027$, the smaller samples have higher model rejection rates, although the power is never as high as 60 percent at any sample size. This effect is predicted given that the sample estimates are converging on the population value at larger sample sizes. Because of the small population RMSEA, a cutoff value even much smaller than 0.05 does not yield high model rejection rates.

We observe a similar pattern for the moderate misspecification of Model 1. As in the other figures, the rejection rate curves decline quickly with increases in the critical cutoff (see Figure 6). With $c = 0$, the rejection

Figure 7
Model Rejection Rates by Sample Size,
H_0: RMSEA $\leq c$ Model 1, Severest Misspecification

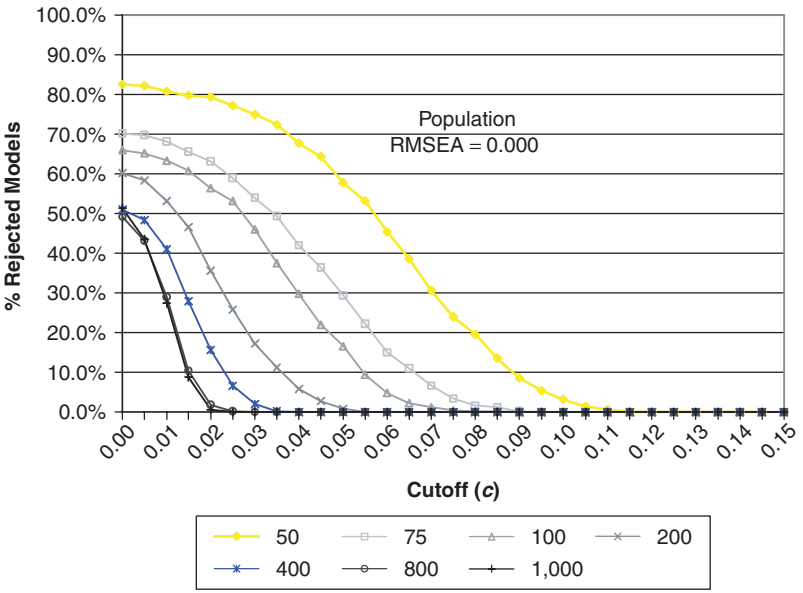


Note: RMSEA = root mean error of approximation.

rate is close to 100 percent for $n = 800$ and is 67 percent for $n = 75$, indicating high power of the test for large samples and lower power of the test for small samples. However, the gap disappears when larger critical values are used. Since larger samples have a steeper power curve, we observe a convergence of curves at $c = 0.04$ (population RMSEA), where rejection rates are around 50 percent for all sample sizes. When $c < 0.04$, the test performs relatively better for bigger samples (higher power of the test) than smaller samples. When $c > 0.04$, the power of the test is low across all sample sizes. For example, if we use the population cutoff of 0.05 (a value greater than the population RMSEA), the rejection rate is approximately 10 percent with $n = 800$, indicating very low power of the test at a rather large sample size.

Similarly, for Model 1 with the largest misspecification, the model rejection rate is extremely sensitive to changes in the cutoff points, particularly

Figure 8
Model Rejection Rates by Sample Size,
H_0: RMSEA $\leq c$ Model 2, Correct Specification

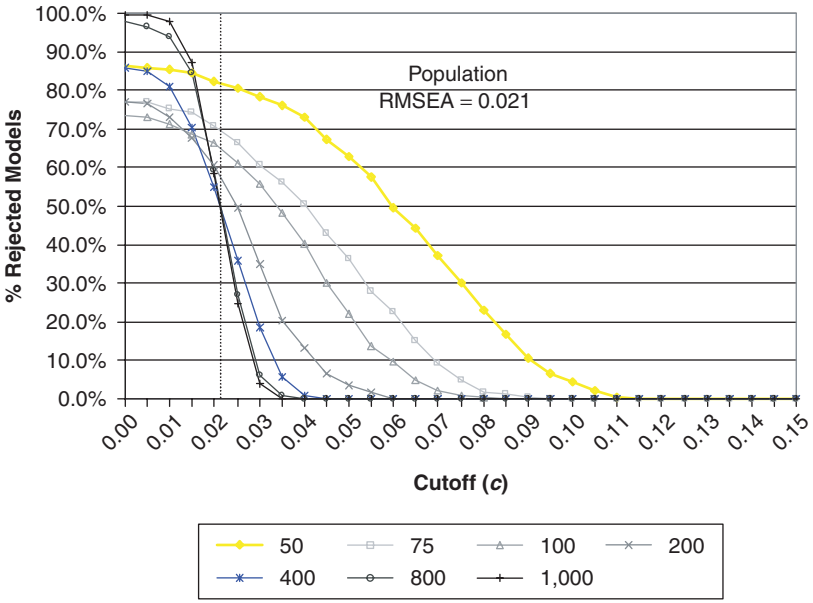


Note: RMSEA = root mean error of approximation.

for the larger samples (see Figure 7). For example, with $c \leq 0.04$, model rejection rates are almost 100 percent with $n = 1{,}000$, suggesting extremely high power of the test. Within the same range of c, the power of the test is lower for smaller samples but still higher than those in a Model 1 with moderate misspecification. For example, the model rejection rate is approximately 60 percent with $c = 0.03$ and with $n = 100$ for a moderately misspecified Model 1. For the most misspecified Model 1, the model rejection rate improves to 80 percent with the same cutoff point and sample size. Again we observe a convergence at around $c = 0.061$, the population RMSEA. Because it is larger than the typical cutoff of 0.05, we observe that the test performs quite well in Table 1 (also see Figure 7).

Moving to the more complex model types, Models 2 and 3, we find the patterns described above are supported in general (see Figures 8 through 15). For the properly specified models, with large samples, there are virtually no

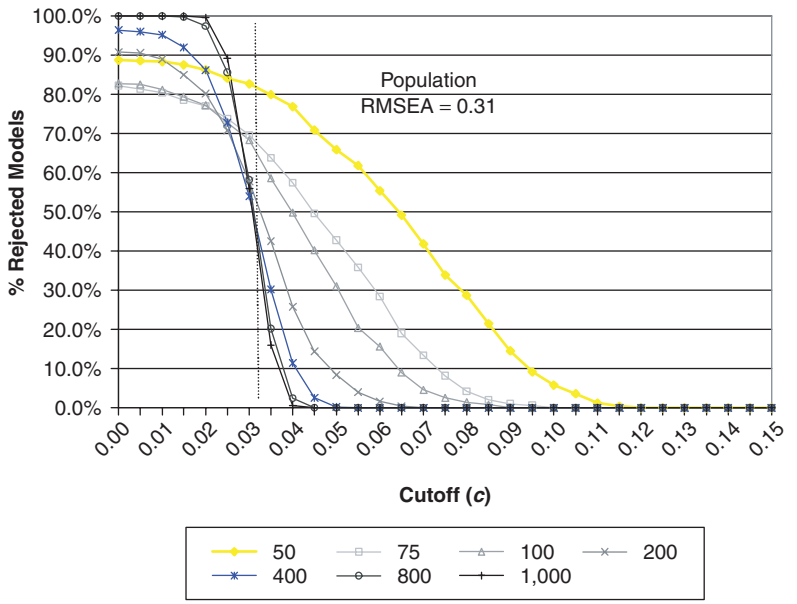
Figure 9
Model Rejection Rates by Sample Size,
H_0: RMSEA $\leq c$ Model 2, Smallest Misspecification



Note: RMSEA = root mean error of approximation.

differences among the three model types. For example, with $n = 1{,}000$, model rejection rates start at around 50 percent with $c = 0$ and quickly declines to zero when $c > 0.02$. For smaller samples, the rejection rates are slightly lower for Model 1 with a given cutoff value, suggesting even lower Type I error rate. For misspecified models, we observe a similar split between the bigger and smaller sample sizes. Again, we observe a convergence of the rejection rate curves around the population RMSEA value. The above analysis focuses on the use of the RMSEA point estimate alone in assessing model fit. We find that model rejection rates vary substantially by the choice of cutoff points, model specification, and sample size, and that there is no empirical support for the use of 0.05 or any other value as a universal cutoff. This then begs the question: Is the choice of 0.05 as a cutoff point better justified when the RMSEA point estimate is used jointly with its CI?

Figure 10
Model Rejection Rates by Sample Size,
H_0: RMSEA $\leq c$ Model 2, Moderate Misspecification

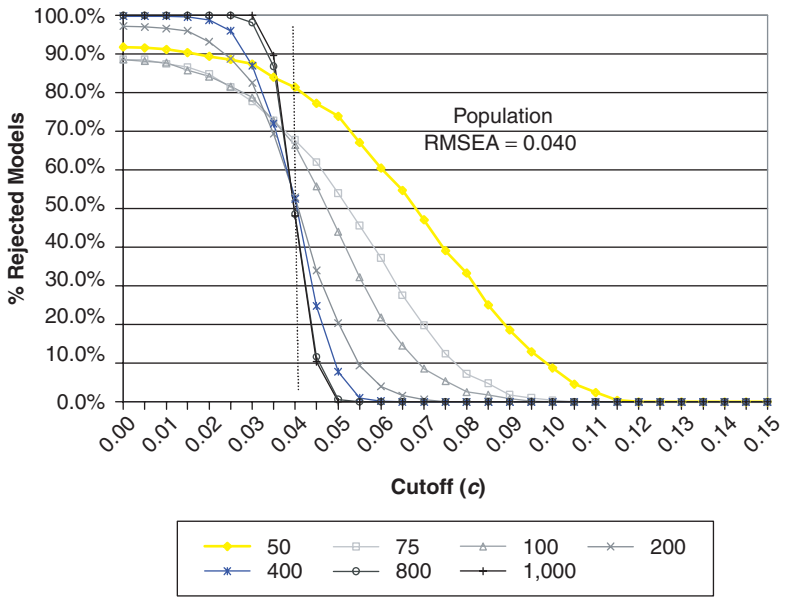


Note: RMSEA = root mean error of approximation.

Test Using the Point Estimate and CI

Certainly the use of CI requires the use of an arbitrary critical value, the use of which we just claimed was unsupported. However, because it is important to consider the empirical behavior of these CIs, we adopt the typical 0.05 and 0.10 values to focus our subsequent presentation. In Table 2, we present the results of two tests, whether the *lower* bound of the CI is *greater* than 0.05, and whether the *upper* bound of the CI is *greater* than 0.1. The first test, a test of close fit, performs extremely well for properly specified models, reflected by similar and low model rejection rates across all sample sizes and different models. For example, with $n = 50$, the model rejection rate ranges from 4 percent to 7 percent from Model 1 to 3. With $n \geq 200$, model rejection rates for all properly specified models approaches zero. This is an improvement over the test of using the point estimate alone, where we

Figure 11
Model Rejection Rates by Sample Size,
H_0: RMSEA $\leq c$ Model 2, Severest Misspecification

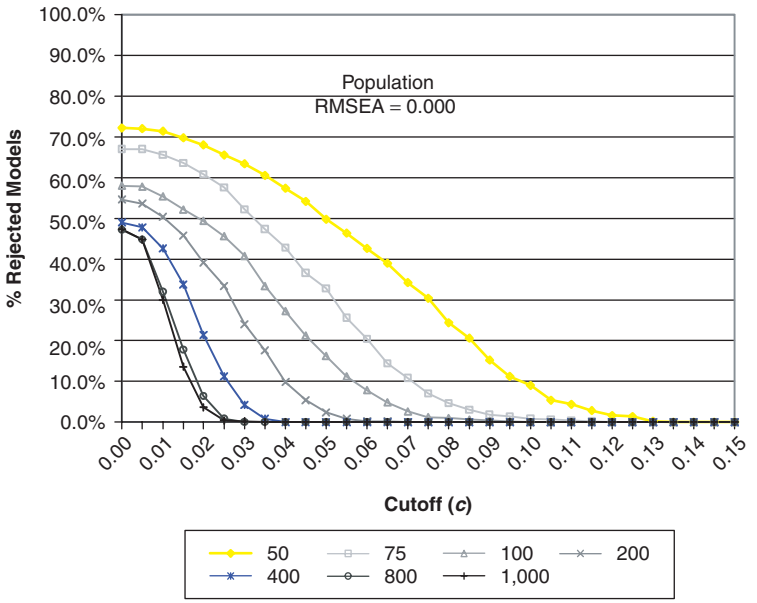


Note: RMSEA = root mean error of approximation.

observe higher model rejection rates (i.e., higher Type I error rate) and more variations across different models with $n \leq 200$.

However, for misspecified models, the test of lower CI of RMSEA $\leq$ 0.05 does not achieve higher power than the check of whether the RMSEA $\leq$ 0.05. In contrast, model rejection rates for all the misspecified models are consistently lower than those numbers shown in Table 1. For example, with $n = 1{,}000$, for Model 1 with the largest misspecification, the model rejection rate is 94 percent when evaluating models based on RMSEA $\leq$ 0.05 (see Table 1). In comparison, the model rejection rate is only 49 percent when the lower bound of the CI is considered. The difference is even more dramatic for mildly or moderately misspecified models. For example, with $n = 100$, for Model 1 with the smallest misspecification, the model rejection rate is 35 percent when we use the point estimate alone (see Table 1) but drops to 2 percent when we use the lower bound of

Figure 12
Model Rejection Rates by Sample Size,
H_0: RMSEA $\leq c$ Model 3, Correct Specification

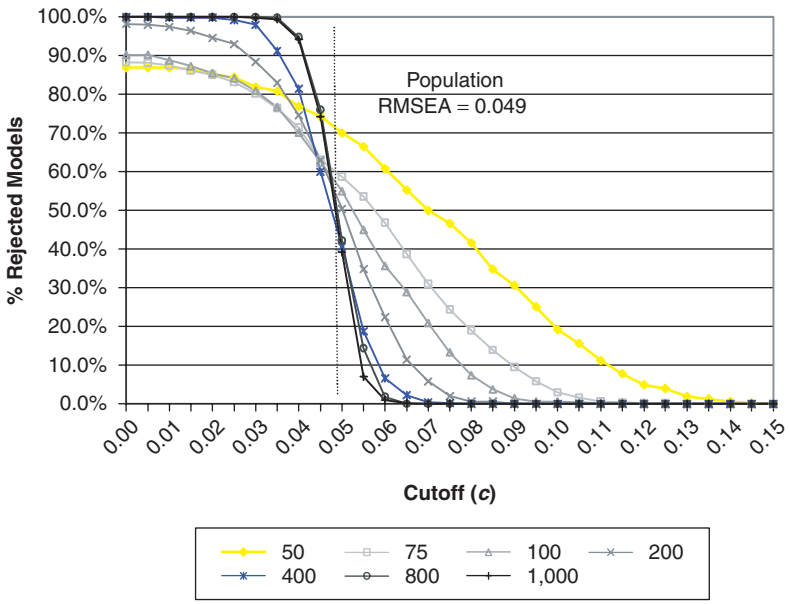


Note: RMSEA = root mean error of approximation.

the CI (see Table 2). Similarly, for Model 3 with moderate and the largest misspecification, when the point estimate is used in the test, model rejection rates are above 90 percent for all sample sizes. When the lower bound of the CI is used, model rejection rates are above 90 percent for $n \geq 200$, but for the smaller samples, model rejection rates range from 51 percent to 85 percent.

When we use upper CI of RMSEA ≤ 0.1 as the test of model fit, the performance of the test is generally worse than the test of lower CI of RMSEA ≤ 0.05. Unlike the previous test, which achieved low Type I error rates for properly specified models across all sample sizes, model rejection rates are still quite high for smaller samples when this test is used. For example, for the correctly specified Model 1, model rejection rate is as high as 71 percent with $n = 50$. For misspecified models, the test performs poorly. For example, for a moderately misspecified Model 3, with $n = 1{,}000$,

Figure 13
Model Rejection Rates by Sample Size,
H_0: RMSEA $\leq c$ Model 3, Smallest Misspecification

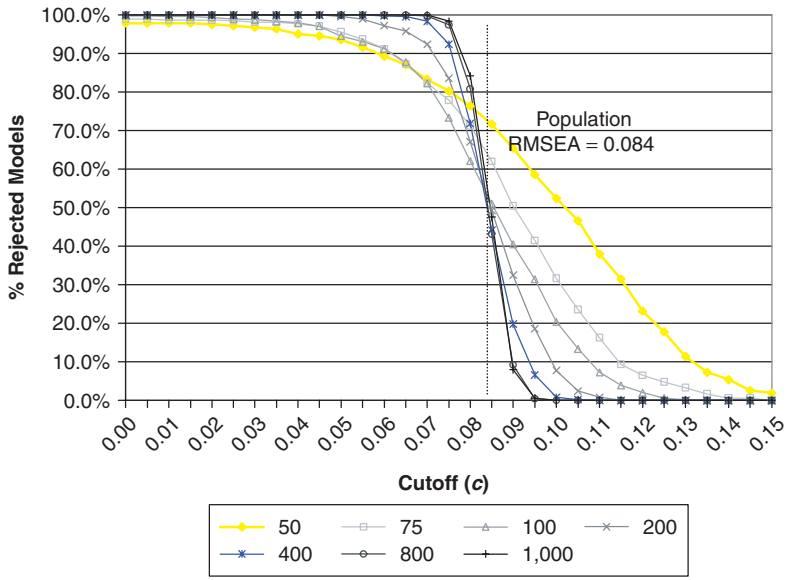


Note: RMSEA = root mean error of approximation.

the model rejection rate is a strikingly low 2 percent (i.e., very low power), compared with 100 percent when the test of lower CI of RMSEA ≤ 0.05 is used.

Directly comparing Tables 1 and 2, it is clear that evaluating model fit based on the joint use of the RMSEA estimate and CI suffers problems similar to using the RMSEA point estimate alone. The CI has the advantage of revealing the uncertainty in our estimates. But when it is combined with fixed cutoffs, it, like the cutoff for point estimates, can lead to incorrect decisions such as too frequently rejecting true models or too infrequently rejecting misspecified models. We believe that the problem does not lie with the logic of the test, but lies with the arbitrary choice of the fixed cutoff. If 0.05 does not work well as the cutoff in the test using the point estimate alone (which our earlier results unambiguously indicated),

Figure 14
Model Rejection Rates by Sample Size,
H_0: RMSEA $\leq c$ Model 3, Moderate Misspecification



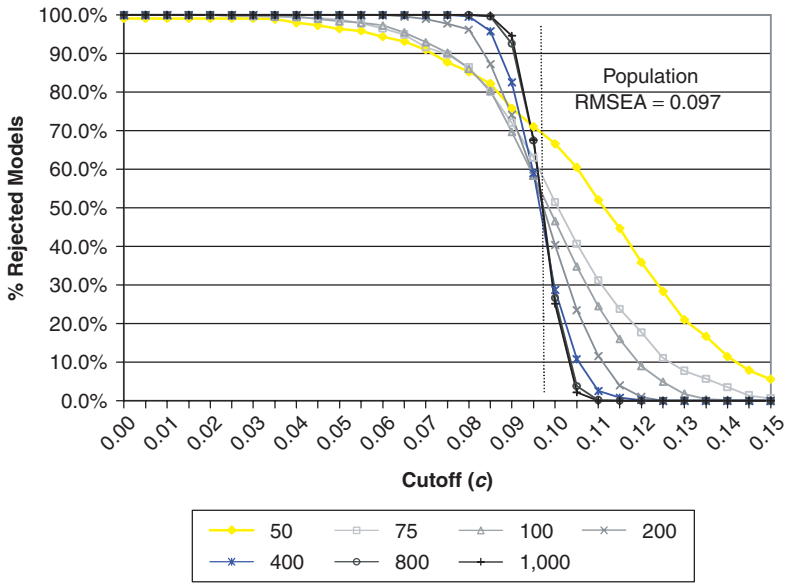
Note: RMSEA = root mean error of approximation.

then using the CI in conjunction with the point estimate does not help to overcome this limitation.

Conclusions

Current popularity of the RMSEA is partially driven by the availability of easy-to-use cutoff values (e.g., 0.05) and the possibility of forming CIs for it. Our simulation results raise questions about the use of a universal cutoff of 0.05, or any other specific value, to evaluate model fit. This is true whether the point estimate is used alone or used with the CI. For properly specified models, a 0.05 cutoff value of the RMSEA rejects too many valid models in small sample sizes ($n \leq 100$), albeit performing better in larger sample sizes (although it tends to overaccept at $n \geq 800$).

Figure 15
Model Rejection Rates by Sample Size,
H_0: RMSEA $\leq c$ Model 3, Severest Misspecification



Note: RMSEA = root mean error of approximation.

For example, with 0.05 as the cutoff for a point estimate and for a properly specified Model 1, 40 percent of the models are incorrectly rejected at $n = 50$ but virtually none at $n = 400$. On the other hand, in some models the same cutoff value for RMSEA (≤ 0.05) too frequently accepts misspecified models (particularly models with moderate and larger misspecifications) to be models of good fit. For example, for Model 1 with moderate misspecification, only 6 percent of the models are rejected at $n = 1{,}000$ and 49 percent of the models are rejected at $n = 50$. The power of the test using the RMSEA is generally low except in the cases of Model 3 with moderate and severe misspecification. This indeed suggests poor performance of the test, because the degrees of bias in some of the parameter estimates are quite noteworthy in these misspecified models.

Table 2
Proportion of Models Rejected Based on the RMSEA Confidence Interval

	Lower CI ≤ 0.05 (null)							Upper CI ≤ 0.1 (null)						
	N							*N*						
Model	50	75	100	200	400	800	1,000	50	75	100	200	400	800	1,000
Model 1														
Correct specification	0.040	0.016	0.012	0.000	0.000	0.000	0.000	0.710	0.490	0.290	0.004	0.000	0.000	0.000
Smallest misspecification	0.055	0.020	0.018	0.004	0.000	0.000	0.000	0.730	0.548	0.378	0.018	0.000	0.000	0.000
Moderate misspecification	0.060	0.030	0.034	0.012	0.004	0.000	0.000	0.758	0.579	0.450	0.044	0.000	0.000	0.000
Severest misspecification	0.105	0.091	0.139	0.162	0.284	0.410	0.490	0.831	0.726	0.645	0.266	0.028	0.002	0.000
Model 2														
Correct specification	0.068	0.012	0.004	0.000	0.000	0.000	0.000	0.520	0.066	0.004	0.000	0.000	0.000	0.000
Smallest misspecification	0.088	0.016	0.010	0.000	0.000	0.000	0.000	0.567	0.094	0.008	0.000	0.000	0.000	0.000
Moderate misspecification	0.122	0.038	0.024	0.000	0.000	0.000	0.000	0.594	0.128	0.022	0.000	0.000	0.000	0.000
Severest misspecification	0.168	0.058	0.048	0.014	0.002	0.000	0.000	0.645	0.194	0.038	0.000	0.000	0.000	0.000
Model 3														
Correct specification	0.072	0.020	0.010	0.000	0.000	0.000	0.000	0.620	0.224	0.038	0.000	0.000	0.000	0.000
Smallest misspecification	0.188	0.107	0.062	0.060	0.034	0.042	0.036	0.791	0.468	0.229	0.006	0.000	0.000	0.000
Moderate misspecification	0.513	0.539	0.607	0.926	0.998	1.000	1.000	0.953	0.910	0.831	0.629	0.298	0.040	0.022
Severest misspecification	0.668	0.766	0.857	0.992	1.000	1.000	1.000	0.975	0.961	0.930	0.942	0.884	0.862	0.816

Note: RMSEA = root mean error of approximation; CI = confidence interval.

When the CI is used jointly with the RMSEA point estimate, there is again no justification to use 0.05 as the universal cutoff point. The test of the lower RMSEA CI ≤ 0.05 performs inconsistently across different model specifications and sample sizes. For properly specified models, the test generally performs well. For example, only 4 percent of the properly specified models are incorrectly rejected at $n = 50$ when the lower bound of the CI is larger than 0.05. However, this approach fails to reject enough misspecified models, where there is substantial bias in the parameter estimates. For example, the test rejects no models for Model 2 with greatest misspecification at $n = 1{,}000$. The power of the test is above 90 percent only for Model 3 with moderate and severest misspecification at $n \leq 200$. Thus, our findings cast serious doubt on the utility of the test of close fit (lower RMSEA CI ≤ 0.05) offered by several major SEM packages. Though the CI is an attractive feature of the RMSEA, its use with a fixed cutoff value for the lower or upper bound does not work well in our simulations.

Our follow-up analyses evaluating the performance of the RMSEA varying the cut points indicated that to achieve a certain level of power or Type I error rate, the choice of cutoff point depends on model specifications, degrees of freedom, and sample size. When the cutoff value is larger than the population RMSEA, the test has very low power and would thus likely indicate that a misspecified model had achieved proper fit. A large critical value (larger than the population value) thus translates into a less stringent test. For example, at $n = 1{,}000$, for Model 1 with moderate misspecification (population RMSEA $= 0.040$), using 0.05 as the cutoff only rejects 6 percent of the models while the test rejects 85 percent of the models when the cutoff is 0.03.

Obviously, in practice, researchers never know the population RMSEA. Thus, any effort to identify universal cutoff points for the RMSEA is not supported and should not be pursued as a single way of assessing model fit. First, we argue that it is not optimal to strive for single-test accept/reject decisions, particularly because the nature of this test is very different from the conventional hypothesis test such as t test, for which the relationship between the critical value α and the power of the test is known. Hence, it is important to use other goodness-of-fit measures to inform global model fit and to attend to diagnostics for the sources of model misfit (Bentler 2007; Bollen and Long 1993; Hayduk et al. 2007; Tanaka 1993). Second, it is difficult to justify a cutoff of 0.05 or any other cutoff value in that the relationship between this value and the degree of misspecification depends on the structure and size of the model in complex ways that are further confounded by sample size effects. We also do not attempt to come up with an

alternative set of cutoff values since these would suffer from analogous disadvantages. Indeed, there was some recent heated debate on whether the RMSEA or other approximate fit tests should be abandoned at all, given that the appropriateness of current thresholds for several fit indices was cast in doubt by recent studies (see Barrett [2007] and responses to the articles by Bentler [2007], Goffin [2007], Hayduk et al. [2007], Markland [2007], McIntosh [2007], Mulaik [2007], and Steiger [2007]).

Some of these authors as well as SEMNET Listserv discussions question whether any fit indices besides the chi-square test statistic are ever needed to evaluate a SEM. Though we are sympathetic to the idea that researchers should investigate the sources of a significant chi-square test, we believe that the RMSEA and other fit indices have utility when used in conjunction with the chi-square test statistic. These indices can supplement the chi-square in assessing the adequacy of a model in matching the data. However, sole reliance on a single fit index seems imprudent and we would recommend that multiple indices be examined (Bentler 2007; Bollen and Long 1993; Tanaka 1993). Ultimately, a researcher must combine these statistical measures with human judgment when reaching a decision about model fit.

Notes

1. We want to note that the misspecifications that we included in the current analysis are certainly not inclusive of all styles of misspecified models. See chapter 5 in Hayduk (1996).

2. By power to detect, we refer to the power to reject a false model, using the chi-square test. The power to detect misspecifications increases with sample size, model complexity, and extent of misspecification. For example, Model 1, at $N = 50$, has a power of 0.065 to detect the smallest specification. At more severe misspecifications, or more complex models, power estimates reach 1.0. See Paxton et al. (2001) for detailed power estimates on each model specification.

3. Some key findings include: improper solutions are more common in small samples than in large ones (ranging from 0 percent at $n = 1{,}000$ for a perfectly specified model to 16 percent of the replications at $n = 50$); no simple positive relation between the degree of misspecification and improper solutions; no practical difference in chi-square test statistics between samples with proper solutions and those with improper solutions; higher bias in parameter estimates in samples with improper solutions than those with proper solutions; and similar mean asymptotic standard errors between samples with proper and improper solutions.

4. Only 23 conditions contained fewer than 500 replications, with the median number of replications as high as 492 and a minimum of 463.

5. In this simulation project, we consider three types of misspecification and regard rejection of these models as correct decisions. However, the root mean square error of approximation (RMSEA) is designed as a measure of approximate fit and some degree of misfit may

well be tolerated. Therefore, it is arguable whether it is "correct" to reject some of the models with minor degrees of misspecification. We believe that the degrees of misspecification that we identify are severe enough for them to be considered misspecified models, but we acknowledge that it is a subjective call.

References

Barrett, P. 2007. "Structural Equation Modeling: Adjusting Model Fit." *Personality and Individual Differences* 42:815-24.

Bentler, P. M. 1990. "Comparative Fit Indexes in Structural Models." *Psychological Bulletin* 107:238-246.

Bentler, P. M. 1995. *EQS: Structural equations program manual, Version 5.0.* Los Angeles: BMDP Statistical Software.

Bentler, P. 2007. "On Tests and Indices for Evaluating Structural Models. *Personality and Individual Differences* 42:825-29.

Bollen, K. A. and J. S. Long. 1993. "Introduction." Pp. 1-9 in *Testing Structural Equation Models*, edited by K. Bollen and J. Long. Newbury Park, CA: Sage.

Browne, M. W., and Cudeck, R. 1993. "Alternative Ways of Assessing Model Fit." Pp. 136-162 in *Testing Structural Equation Models*, edited by K. Bollen and J. Long. Newbury Park, CA: Sage.

Chen, F., K. A. Bollen, P. J. Curran, P. Paxton, & J. Kirby. 2001. "Improper Solutions in Structural Equation Modeling: Causes, Consequences and Strategies." *Sociological Methods and Research* 29: 468-508.

Curran, P. J., K. A. Bollen, P. Paxton, J. Kirby, and F. Chen. 2002. "The Noncentral Chi-Square Distribution in Structural Equation Modeling: Use or Abuse?" *Multivariate Behavioral Research* 37:1-36.

Curran, P. J., K. A. Bollen, F. Chen, P. Paxton, and J. Kirby. 2003. "Finite Sampling Properties of the Point Estimates and Confidence Intervals of the RMSEA." *Sociological Methods and Research* 32:208-52.

Fan, X., B. Thompson, and L. Wang. 1999. "Effects of Sample Size, Estimation Methods, and Model Specification on Structural Equation Modeling Fit Indexes." *Structural Equation Modeling* 6:56-83.

Goffin, R. D. 1993. "A Comparison of Two New Indices for the Assessment of Fit of Structural Equation Models." *Multivariate Behavioral Research* 28:205-14.

Goffin, R. D. 2007. "Assessing the Adequacy of Structural Equation Models: Golden Rules and Editorial Policies." *Personality and Individual Differences* 42:831-39.

Hancock, G. R. and M. J. Freeman. 2001. "Power and Sample Size for the Root Mean Square Error of Approximation Test of Not Close Fit in Structural Equation Modeling." *Educational and Psychological Measurement* 61:741-58.

Hayduk, L., G. Cummings, K. Boadu, H. Pazderka-Robinson, and S. Boulianne. 2007. "Testing! Testing! One, Two, Three-Testing the Theory in Structure Equation Models." *Personality and Individual Differences* 42:841-50.

Hayduk, L. A. and D. N. Glaser. 2000. "Jiving the Four-Step, Waltzing Around Factor Analysis, and Other Serious Fun." *Structural Equation Modeling* 7:1-35.

Hu, L. and P. M. Bentler. 1998. "Fit Indices in Covariance Structure Modeling: Sensitivity to Underparameterized Model Misspecification." *Psychological Methods* 3:424-53.

Hu, L. and P. M. Bentler. 1999. "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives." *Structural Equation Modeling* 6:1-55.

Kenny, D. A. and D. McCoach. 2003. "Effect of the Number of Variables on Measures of Fit in Structural Equation Modeling. *Structural Equation Modeling* 10:333-51.

MacCallum, R. C., M. W. Browne, and H. M. Sugawara. 1996. "Power Analysis and Determination of Sample Size for Covariance Structure Modeling." *Psychological Methods* 1:130-49.

Marsh, H. W., K. Hau, and Z. Wen. 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* 11:320-41.

Markland, D. 2007. "The Golden Rule Is That There Are No Golden Rules: A Commentary on Paul Barrett's Recommendations for Reporting Model Fit in Structural Equation Modeling." *Personality and Individual Differences* 42:851-58.

McIntosh, C. N. 2007. "Rethinking Fit Assessment in Structural Equation Modeling: A Commentary and Elaboration on Barrett 2007." *Personality and Individual Differences* 42:859-67.

Mulaik, S. 2007. "There Is a Place for Approximate Fit in Structural Equation Modeling." *Personality and Individual Differences* 42:883-91.

Nasser, F. and J. Wisenbaker. 2003. "A Monte Carlo study investigating the impact of item parceling on measures of fit in confirmatory factor analysis." *Educational and Psychological Measurement* 63:729-57.

Nevitt, J. and G. R. Hancock. 2000. "Improving the root mean square error of approximation for nonnormal conditions in structural equation modeling." *Journal of Experimental Education* 68:251-68.

Olsson, U. H., T. Foss, and E. Breivik. 2004. "Two equivalent discrepancy functions for maximum likelihood estimation: Do their test statistics follow a non-central chi-square distribution under model misspecification?" *Sociological Methods & Research* 32:453-500.

Paxton, P., K. Bollen, P. Curran, J. Kirby, and F. Chen. 2001. "Monte Carlo Experiments: Design and Implementation." *Structural Equation Modeling* 8:287-312.

Steiger, J. H. 1989. *Causal modeling: a supplementary module for SYSTAT and SYGRAPH*. Evanston, IL: SYSTAT.

Steiger, J. H. 2000. "Point estimation, hypothesis testing, and interval estimation using the RMSEA: Some comments and a reply to Hayduk and Glaser." *Structural Equation Modeling* 7:149-62.

Steiger, J. H. 2007. "Understanding the limitations of global fit assessment in structural equation modeling." *Personality and Individual Differences* 42:893-98.

Steiger, J. H. and J. C. Lind. 1980. "Statistically based tests for the number of common factors." Paper presented at the annual meeting of the Psychometric Society, May, Iowa, City, IA.

Steiger, J. H., A. Shapiro, and M. W. Browne. 1985. "On the multivariate asymptotic distribution of sequential chi-square tests." *Psychometrika* 50:253-64.

Sugawara, H. M. and R. C. MacCallum. 1993. "Effect of estimation method on incremental fit indexes for covariance structure models." *Applied Psychological Measurement* 17:365-77.

Tanaka, J. S. 1993. "Multifaceted conceptions of fit in structural equation models." Pp. 10-40 in *Testing Structural Equation Models*, edited by K. Bollen and J. Long. Newbury Park, CA: Sage.

Yuan, K. 2005. "Fit indices versus test statistics." *Multivariate Behavioral Research* 40:115-48.
Yuan, K., K. Hayashi, and P. Bentler. 2007. "Normal theory likelihood ratio statistic for mean and covariance structure analysis under alternative hypotheses." *Journal of Multivariate Analysis* 9:1262-1282.

Feinian Chen is an associate professor of sociology at the North Carolina State University. Her research interests include family, demography, aging, and quantitative methods. Her recent work has appeared in *Social Forces* and *Population and Development Review*. She is currently investigating the impact of grandparent caregiving on grandparents' health.

Patrick J. Curran is a professor in the L.L. Thurstone Psychometric Laboratory in the Department of Psychology at the University of North Carolina at Chapel Hill. His primary area of research focuses on the development of new quantitative methods for the analysis of longitudinal data, particularly as applied to the study of drug and alcohol use in children and adolescents. He has recently published papers in *Psychological Methods, Multivariate Behavioral Research, Structural Equation Modeling, Developmental Psychology*, and *Development and Psychopathology*. He also coauthored *Latent Curve Models: A Structural Equation Perspective* (Wiley) with Kenneth Bollen.

Kenneth A. Bollen is the director of the Odum Institute for Research in Social Science and the Immerwahr Distinguished Professor at the University of North Carolina. In 2000 he received the *Lazarsfeld Award for Methodological Contributions in Sociology*. The *ISI* named him among the *World's Most Cited Authors* in the Social Sciences. In 2005, three of his articles were recognized as among the most cited articles in the history of the *American Sociological Review*. He has published *Latent Curve Models* (with P. Curran, 2006, Wiley), *Structural Equation Models with Latent Variables* (1989, Wiley) and over 100 papers. Bollen's primary research areas are structural equation models, population studies, and democratization.

James B. Kirby is a senior social scientist at the Agency for Healthcare Research and Quality. His research focuses on family and community influences on health, health behaviors, and access to health care in the United States, with an emphasis on racial, ethnic, and socioeconomic disparities. Dr. Kirby also conducts methodological research on structural equation models and multilevel models.

Pamela Paxton is an associate professor of sociology and political science (by courtesy) at the Ohio State University. Some of her previous substantive research on trust and social capital appears in the *American Sociological Review*, the *American Journal of Sociology*, and *Social Forces*. With Melanie Hughes, she is the coauthor of the 2007 book, *Women, Politics, and Power: A Global Perspective*. Her current research considers women's political participation over time and connections between social capital and social networks.