Sociological Methods & Research
2015, Vol. 44(3) 486-507
© The Author(s) 2014
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0049124114543236
smr.sagepub.com
SAGE

The Performance of RMSEA in Models With Small Degrees of Freedom

David A. Kenny[1], Burcu Kaniskan[2], and D. Betsy McCoach[1]

Abstract

Given that the root mean square error of approximation (RMSEA) is currently one of the most popular measures of goodness-of-model fit within structural equation modeling (SEM), it is important to know how well the RMSEA performs in models with small degrees of freedom (*df*). Unfortunately, most previous work on the RMSEA and its confidence interval has focused on models with a large *df*. Building on the work of Chen et al. to examine the impact of small *df* on the RMSEA, we conducted a theoretical analysis and a Monte Carlo simulation using correctly specified models with varying *df* and sample size. The results of our investigation indicate that when the cutoff values are used to assess the fit of the properly specified models with small *df* and small sample size, the RMSEA too often falsely indicates a poor fitting model. We recommend not computing the RMSEA for small *df* models, especially those with small sample sizes, but rather estimating parameters that were not originally specified in the model.

Keywords

structural equation modeling, model fit, RMSEA, degrees of freedom

[1] University of Connecticut, Storrs, CT, USA
[2] National Conference of Bar Examiners (NCBE) Madison, WI, USA

Corresponding Author:
David A. Kenny, University of Connecticut, Storrs, CT 06269, USA.
Email: david.kenny@uconn.edu

We begin with an example. Jane Miller runs a path model using structural equation modeling (SEM) with eight paths, one of which she thinks is zero. She has a sample size of 150. Jane has learned that a good fitting model should have a root mean square error of approximation (RMSEA) less than 0.10 (but see later discussion). However, when she runs her model fixing that one path to zero and leaving the other seven free, she obtains a $\chi^2(1) = 2.70$ ($p = .100$) and finds that her RMSEA is 0.107, above the cutoff of 0.10. She then estimates a second model that sets all eight of her paths in the model to zero. For this model, she obtains $\chi^2(8) = 17.0$ ($p = .031$). However, because her RMSEA is now 0.084 and below the cutoff of 0.10, she thinks that this second model is a better model than the original model. She draws this conclusion even though the χ^2 difference test of $\chi^2(7) = 14.3$ ($p = .046$) indicates that the latter model is the worse model. Using a cutoff value like 0.10 for the point estimate of the RMSEA by itself can be very misleading, especially when model *df* are small, a problem illustrated by what happened to Jane Miller.

One of the key challenges in SEM is the assessment of model fit. Early in the history of SEM, χ^2 test was the only way to assess model fit; however, researchers soon recognized that this statistic possessed several limitations, the most important being that it was heavily dependent on sample size. To remedy these limitations, a parade of methodologists, beginning with Bentler and Bonett (1980), developed a plethora of model fit indices for SEM and provided various suggestions and recommendations for evaluating model fit.

There are now a myriad of measures of fit, and probably the most popular is RMSEA (Steiger and Lind 1980). In a recent article, Jackson, Gillaspy, and Purc-Stephenson (2009) reviewed 194 confirmatory factor analysis studies published in American Psychological Association journals from 1998 to 2006 and found a strong increase in the number of times that RMSEA was reported. Only 37 percent of the studies reported RMSEA values from 1998 through 2000, but this number jumped to 83 percent from 2004 through 2006. McDonald and Ho (2002) conducted a similar review of several journals from 1995 to 1997, and they found RMSEA was the second most popular model fit index, after the Comparative Fit Index.

Originally introduced by Steiger and Lind (1980) and popularized by Browne and Cudeck (1993), the RMSEA is defined in the population as

$$\varepsilon = \sqrt{\frac{\lambda}{df(N-1)}}, \tag{1}$$

where λ is the noncentrality parameter of the noncentral χ^2 distribution, *df* is the model's degrees of freedom, and *N* is the sample size. In the sample, λ is estimated by $\chi^2 - df$ or zero if χ^2 is less than *df*.

Browne and Cudeck (1993) suggested population parameter values of RMSEA of about 0.05 or less are indicative of close fit of the model and values of about 0.08 or less indicate reasonable error of approximation. Moreover, they stated that they "would not want to employ a model with a RMSEA greater than 0.1" (p. 144). Hu and Bentler (1999) recommended that adequately fitting models should have RMSEA values below 0.06. MacCallum, Browne, and Sugawara (1996) used 0.01, 0.05, and 0.08 to indicate excellent, good, and mediocre fit, respectively. However, these criteria were intended as population values of the RMSEA and not as cutoffs to empirically establish good- and bad-fitting models. Further, any suggested cutoff values are by their nature arbitrary. It is not our intention to endorse the use of cutoff values or to determine optimal cutoff values.

Very different approaches have been proposed for how the RMSEA should be used (Browne and Cudeck 1993; MacCallum et al. 1996) to assess model fit. These include the "test of exact fit" and the "test of close fit" (Browne and Cudeck 1993; MacCallum et al. 1996). The test of exact fit specifies and tests a null hypothesis that the RMSEA equals zero. In this approach, if the null hypothesis is not rejected, the model is assumed to fit well. However, Browne and Cudeck (1993) argued that the exact fit approach appears to penalize the models with large sample size because it would very likely lead to a rejection of the null hypothesis that the RMSEA equals zero. Hence, they introduced the test of close fit approach and supported their claim with variety of empirical examples. In the test of close fit, the null hypothesis is that the RMSEA is less than or equal to .05. Browne and Cudeck outlined the explanation of this alternative approach:

> (T)he choice of 0.05 . . . as an upper limit for a close-fitting model is one based on a substantial amount of experience with estimates of RMSEA. It is, however, no less subjective than the choice of 5% as a significance level. What can be said about the null hypothesis of close fit . . . is that it is far less unrealistic than the null hypothesis of exact fit. (pp. 146-47)

Currently, several SEM programs support the calculation of the *p* value associated with the test of close fit. Despite the popular use of these cutoff values, many methodologists caution applied researchers about the use of cutoff values and argue against the notion of using a universal cutoff point as the sole means of assessing model fit (Chen et al. 2008; West, Taylor, and Wu 2012).

To help assess the sampling error in the RMSEA, a confidence interval (CI) around the sample point estimate of the RMSEA can be computed. The CI provides important information about the RMSEA. First, the width of the CI demonstrates the degree of uncertainty in the estimate in the RMSEA. Furthermore, the lower value provides an optimistic estimate for the value for the RMSEA, and the upper limit provides a pessimistic value.

Several groups of investigators have investigated the performance of the RMSEA using simulations (Breivik and Olsson 2001; Chen et al. 2008; Curran et al. 2003; Curran et al. 2002; Curran, West, and Finch 1996; Hu and Bentler 1999; Kenny and McCoach 2003; Nevitt and Hancock 2000). Almost all of this simulation work for the RMSEA and its CI examined models with moderate to large *df*. In perhaps the most influential study on measures of fit in general and the RMSEA in particular, Hu and Bentler (1999) investigated the adequacy of RMSEA for models of 84 and 89 *df*. Additionally, a series of articles by Chen et al. (2008), Curran et al. (2002), and Curran et al. (2003)) all used the same three models with varying specifications, and their properly specified models had 22, 50, and 85 *df*.

Moreover, Curran et al. (1996) and Nevitt and Hancock (2000) used the same model that had 18 *df*. Nevitt and Olson concluded that the test of not-close fit for the RMSEA may not perform well for small-to-moderate-sized samples. Breivik and Olsson (2001), who examined the performance of the RMSEA for models with 24, 120, 288, and 528 *df*, found that the RMSEA tended to show better fit for larger models. In one of the few studies to use small *df*, Kenny and McCoach (2003) considered models with 2, 9, 35, 54, 77, 170, and 275 *df*. The concluded that the RMSEA tends to improve as more variables are included in the models.

In Chen et al. (2008), sample size clearly influenced the performance of the RMSEA. For example, in their correctly specified model with 22 *df*, the point estimate of the RMSEA exceeded .05 over 38 percent of the time when the sample size was 50, 26 percent of the time when the sample size was 100, 7.8 percent of the time when the sample size was 200, and only .2 percent of the time when the sample size was 400. However, based on table 1 of Chen et al. (2008), there is some evidence that rejection rates may increase as *df* decline, at least for small *N* studies. For example, when *N* is 50 and we use 0.10 as the cutoff, the rejection rates for properly specified models were .112 with 22 *df*, .090 with 50 *df*, and .032 with 85 *df*. However, once sample sizes are 100 or more, there are very few rejections when using a cutoff of .10 for the RMSEA. Even so, evidence of this pattern persists. For example, with correctly specified models and an RMSEA cutoff of .05, the rejection rates were .078 for the 22 *df* model, .024 for the 50 *df* model, and .008 for the

Table 1. Probability RMSEA > *c* (Smaller *N* Values).

		RMSEA > 0.10		RMSEA > 0.08		RMSEA > 0.05	
N	*df*	Theoretical	Simulation	Theoretical	Simulation	Theoretical	Simulation
50	1	.222	.238	.252	.264	.289	.293
50	2	.225	.216	.269	.268	.325	.326
50	3	.215	.233	.268	.275	.338	.354
50	5	.189	.203	.255	.281	.346	.372
50	10	.136	.179	.216	.274	.340	.408
50	20	.073	.121	.157	.226	.317	.411
50	50	.014	.041	.068	.164	.256	.443
100	1	.158	.161	.201	.204	.264	.261
100	2	.137	.146	.195	.205	.287	.305
100	3	.113	.128	.179	.188	.291	.308
100	5	.077	.091	.147	.172	.284	.305
100	10	.030	.037	.090	.110	.255	.300
100	20	.005	.009	.037	.056	.203	.227
100	50	.000	.000	.003	.005	.112	.161
200	1	.084	.070	.132	.117	.221	.200
200	2	.050	.056	.103	.104	.224	.234
200	3	.030	.039	.078	.091	.213	.211
200	5	.011	.012	.045	.046	.187	.187
200	10	.001	.002	.012	.017	.133	.152
200	20	.000	.000	.001	.001	.071	.091
200	50	.000	.000	.000	.000	.013	.021

85 *df* model. We might wonder if trend become even stronger with very low *df* models and how large the sample sizes would need to be to prevent too many false rejections in cases where the *df* are very small.

Thus, most of what we know about the behavior of the RMSEA refers to models with relatively large *df*. We might wonder about how RMSEA performs in models with small *df*. This lack of attention to small *df* models has not gone unnoticed:

> One set of models that might be of particular interest that we did not study here are models with a very small *df*. For example, a three–time point linear latent growth model is characterized by a single *df*, and we have found in our own applied work that resulting RMSEA values can be quite large for a model that otherwise appears to fit the data well. Further examination of the RMSEA under conditions such as these could be quite interesting. (Curran et al. 2003, p. 248)

MacCallum et al. (1996) have noted tests of hypotheses for good and bad model fit for models with small *df* (i.e., 2) have low power and require thousands of observations to have reasonable power. As shown in their table 4, for $df = 2$, the minimum *N* is 3,488 for the test of close fit, that is, the null hypothesis is that the RMSEA = 0.05, and the alternative is 0.08, and minimum *N* is 2382 for the test of not close fit, that is, the null hypothesis is that the RMSEA = 0.05 and alternative is 0.01. Additionally, both Breivik and Olsson (2001) and Kenny and McCoach (2003) noted that the average value of the RMSEA improved (i.e., declined) for larger *df* models. It might then be the case that for models with very small *df*, one or two, what may be good fitting models might produce unacceptably large RMSEA values. Obviously, it is dangerous to extrapolate; for this reason, in this article, we examine more carefully the performance of the RMSEA in low *df* models. Simply stated, our concern is that the RMSEA may show poor fit (e.g., greater than 0.10) for a given *df* and *N* in correctly specified models; we hypothesize that this probability might be unacceptably high for models with very small *df*.

Why worry at all about small *df* models? They are much more common in Confirmatory factor analyses (CFAs) and SEM than might be thought. Most path models (models with observed variables as causes) or path models with single indicators that use the Williams and Hazer (1986) method of fixing the error variances have relatively small *df*. For instance, the well-known Fishbein and Ajzen (1975) model of the theory of reasoned action has just two *df*. Two others examples are Segrin et al. (2005) who tested a path model that had just one *df* and Frone, Russell, and Cooper (1994) who tested a nonrecursive model with just four *df*.

Certainly, CFA models tend to have large *df*. A model with 12 indicators and three factors, with each indicator loading on only one factor has 51 *df*. However, a single-factor model with few indicators, which are sometimes first estimated in the development a larger CFA, can have few *df*. For example, a model with four indicators and one factor has only two *df*.

Behavior genetics models typically have relatively small *df*. For instance, many of the classic behavior genetics models discussed by Heath et al. (1989) have just one or two *df*.

Finally, a very common SEM with a small number of *df* is a latent growth curve model (LGM). The standard three-wave linear growth model (Curran 2000) has just 1 *df* (see above-mentioned quote from Curran et al. 2003). Models with more waves have more *df*, but if allowances are made for nonlinearity and correlated errors, the number of *df* can still be very small, for example, Llabre et al. (2004) with 1 *df* and McAuley et al. (2003) with 6 *df* and Kaugars et al. (2008) with 5 *df*.

Hence, low *df* models do occur. Moreover, by understanding the behavior of the RMSEA under "extreme" circumstances, we might better understand how it works in general. Our approach is as follows: First, we develop a theoretical model that predicts the rejection rate for RMSEA, using various cutoffs. We then conduct a simulation model to verify the theoretical model.

A Theoretical Approach

We can somewhat understand the distribution of the RMSEA by considering the distribution of the RMSEA squared under the null hypothesis, without setting it to zero when $df \geq \chi^2$, that is, the distribution of $(\chi^2 - df)/[df(N - 1)]$ (see also Rigdon 1996). Based on the distribution of χ^2, it has a population mean of zero and a variance of $2/[df(N - 1)^2]$ with a positive skew of $\sqrt{8}/[\sqrt{df}(N - 1)df]$. We can see that as both *df* and *N* increase, both the variance and the positive skew decline nonlinearly, and the magnitude of the decrease for a 1 *df* change is greatest at the smallest *df*. Although we might expect a somewhat weaker relationship between *df* and *N* with the variance and skew of the RMSEA than with the squared RMSEA without truncation, these results are suggestive. Finding more extreme values of the RMSEA is much more likely when either *df* or *N* are small. We would then expect that the RMSEA would be highly variable with many "large" values when both *df* and *N* are small.

We can approximate the probability that the RMSEA exceeds a certain cutoff for correctly specified models. We assume the computed χ^2 value is exactly, not approximately, distributed as χ^2. We need to know the model *df*, the sample size or *N*, and the RMSEA cutoff or *c*. Then using the χ^2 distribution with *df*, we determine the probability that χ^2 will be greater than the implied value of χ^2 at the cutoff. The resulting formula is

$$P\left(\chi^2_{df} > df\left[(N - 1)c^2 + 1\right]\right). \tag{2}$$

We note that $df[(N - 1)c^2 + 1]$ is the value the χ^2 must equal for the estimated RMSEA to equal *c*. Such a computation is straightforward, and we provide a spreadsheet to accomplish this (http://davidakenny.net/papers/RMSEA_small_df/compute_ps.xls).

Tables 1 and 2 present these theoretical probabilities (simulation probabilities are discussed later) for various samples sizes (from 50 to 1,000), various model *df* (1 to 50), and RMSEA cutoffs (0.10, 0.08, and 0.05). Note that the rows of these tables refer to only one condition (*df* and *N*) and that the model rejection rates for this condition are listed for three different

Table 2. Probability RMSEA > *c* (Larger *N* Values).

		RMSEA > 0.10		RMSEA > 0.08		RMSEA > 0.05	
N	*df*	Theoretical	Simulation	Theoretical	Simulation	Theoretical	Simulation
400	1	.025	.025	.059	.054	.158	.154
400	2	.007	.006	.029	.027	.136	.141
400	3	.002	.002	.014	.013	.112	.127
400	5	.000	.000	.003	.003	.076	.069
400	10	.000	.000	.000	.000	.029	.039
400	20	.000	.000	.000	.000	.005	.004
400	50	.000	.000	.000	.000	.000	.000
600	1	.008	.006	.028	.028	.114	.108
600	2	.001	.000	.008	.009	.082	.089
600	3	.000	.000	.002	.004	.058	.055
600	5	.000	.000	.000	.000	.029	.025
600	10	.000	.000	.000	.000	.005	.012
600	20	.000	.000	.000	.000	.000	.001
600	50	.000	.000	.000	.000	.000	.000
1,000	1	.001	.002	.007	.007	.061	.057
1,000	2	.000	.000	.001	.000	.030	.032
1,000	3	.000	.000	.000	.000	.015	.017
1,000	5	.000	.000	.000	.000	.004	.003
1,000	10	.000	.000	.000	.000	.000	.000
1,000	20	.000	.000	.000	.000	.000	.000
1,000	50	.000	.000	.000	.000	.000	.000

cutoff values. It is important to keep in mind what is contained in these tables. They are the probabilities of finding that RMSEA is greater than some cutoff value (e.g., 0.10) and not the probability that χ^2 exceeds some critical value (e.g., $p < .05$).

We first note that there is a relatively large percentage of poor fitting cases for models with small *df*, especially if *N* is small and *c* is low. Consider what happens when the *df* equal 1 and the sample size is very small. We see in Table 1 that when $N = 100$ and $df = 1$, the proportion of RMSEA values that exceed 0.10 is .264.

However, even when the sample size is 400 or more, there are still large rejection rates for very low *df*. In particular, when $N = 400$, $df = 1$, and we set the RMSEA cutoff at 0.05, the probability has an unacceptably high value of .158. Even when the sample size is as large as 1,000, the probability of obtaining an RMSEA above .05 is .061 when the model has 1 *df*. Thus, we find, as we hypothesized, too many indications of poor model fit, especially when *df* and *N* are small.

The Simulation Study

These theoretical values are disturbing, but we might wonder what would happen with actual data. Therefore, we conducted a Monte Carlo simulation study in order to validate our theoretical probabilities with empirical ones. The purpose of this study is to examine RMSEA in models with small *df* in terms of its (1) performance when cutoff values are used (descriptive comparison with fixed cutoffs), (2) estimation accuracy (bias, 1 [*SD*], and root mean squared error), (3) the CI (the 90 percent CI width and coverages), and (4) power. Each of these research questions has been examined by a separate study in the past using large *df* models (see Chen et al. 2008; Curran et al. 2003); MacCallum et al. 1996; Kim 2005; Rigdon 1996). We varied *df* and *N*, and then we examined the model fit using selected cutoff values. We made sure that we included low *df* models. Moreover, we investigated the coverage of the CI as well as *SD* of the range of the CI of the estimated RMSEA.

Method

In the literature, as for the cutoff value of RMSEA, MacCallum et al. (1996) suggested 0.01 (excellent fit), 0.05 (good fit), and 0.08 (mediocre fit); however, Hu and Bentler (1999) have recommended 0.06 as the cutoff for good fit. Furthermore, SEM packages such as AMOS and LISREL provide the probability value at the cutoff value of RMSEA at .05, we decided to examine cutoff values of 0.05 (good fit), 0.08 (mediocre fit), and 0.10 (poor fit).

We wanted to include some very small sample sizes in the simulation. We chose the smallest sample size as 50. Although this might be considered too small a sample size for latent variable SEM, some researchers may have sample sizes of this magnitude when they conduct path analysis with a small number of variables. Additionally, we included sample sizes of 100 200, 400, 600, and 1,000. For model *df*, we wanted some very small values, and so we chose 1, 2, 3, 5, 10, 20, and 50. Thus, the study employed Monte Carlo simulation with factorial design of 6 (*N*) × 7 (*df*). Each condition was replicated 1,000 times.

Simulation Model

To vary *df*, the study had seven different models. All models were correctly specified growth models, and we present a five-wave growth model in

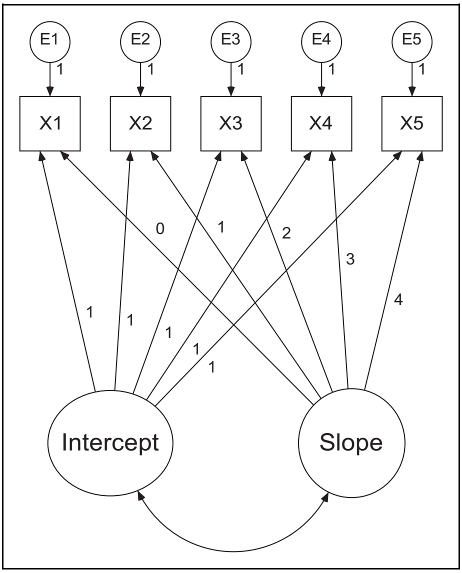

Figure 1. Simulation model for 10 *df.*

Figure 1. The model population values for all of the models were as follows: Intercept loadings were all fixed to one whereas the slope loadings were fixed to zero for wave 1 and increased in one-unit increments thereafter. The population mean of the intercept factor was 0.5 and the variance was set at 1.0: The population mean of the slope factor was 1.0 and its variance was 0.2. The covariance between slope and intercept was 0.1, and all error variances were set to 0.5. The models were as follows and are designated by their *df*:*df* = 1: 3-wave growth model, *df* = 2: 3-wave growth model, with equal error variances and the loading for the slope factor at wave 3 free, *df* = 3: 3-wave growth model, with equal error variances, *df* = 5: 4-wave growth model, *df* = 10: 5-wave growth model (see Figure 1), *df* = 20: 7-wave growth model, with loadings on the slope factor for the last three times free, and *df* = 50: 10-wave growth model.

Data generation and estimation. We used the simulation feature of MPlus Version 5 (Muthén and Muthén 2007) to generate the data, and maximum likelihood estimation was used to estimate the parameters. Results from the 42,000 trials were saved.

Distribution. The raw data were generated from a multivariate normal distribution.

Replications. There were total of 42 experimental conditions, and in each condition there were 1,000 replications. We did have some improper solutions, 2.3 percent overall, and consistent with the findings of the previous studies (e.g., Chen et al. 2008), these were concentrated for the small N and small df models. The condition that had the highest percentage of improper solutions was $N = 50$ and $df = 1$ condition where we had a 38.5 percent nonadmissible solution (usually a negative error, slope, or intercept variance). We compared mean RMSEA between the sample with improper solutions and sample without an improper solution in the model with highest percentage of improper solutions, and there was not a statistically significant difference between these two samples. Every model that we ran converged.

Results

Rejection Rates

In this study, we calculated the percentages of model rejection rates defined as the number of times that the model's RMSEA exceeded a given cutoff value. Based on the earlier presented theoretical work, we did expect higher rejection rates for small df and small N cases.

We checked to see how the simulation values compared to the theoretical values. We judged that the correspondence was quite good, the correlation being .990 for the 0.10 cutoff, .981 for 0.08, and .974 for 0.05. The reader can compare the values in both Tables 1 and 2. As prior theory and simulation work would suggest, our theoretical values were, on average, a bit too small, especially for small N and lower cutoffs.

Because there is a strong agreement between the theoretical and simulated probabilities, we can be confident that our results do not depend on the specific SEMs that we simulated. That is, we believe that our findings would have been the same had we not studied LGM models but had instead studied CFA or path analysis or any SEM.

As we found with the theoretical values, models with small df had high rejection rates unless the model had very large sample sizes. The results of the test of null hypothesis are provided in Table 1 for the smaller sample sizes. Overall, when the df were small, the rejection rates were high. For instance, for $df = 1$ and $N = 200$ and a cutoff of 0.05, we have a 20.0 percent rejection rate, but for 50 df model, the rejection rate shrinks to 2.1 percent. We see a parallel, but weaker, pattern for large sample sizes in Table 2.

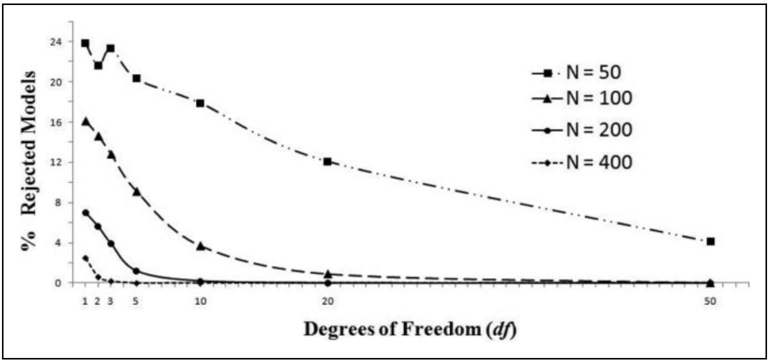

Figure 2. Model rejection rates in percentages for RMSEA $\leq$ 0.10 in correctly specified models for selected sample sizes (*N*) and degrees of freedom (*df*).

Because the overall rejection rate is much lower, the pattern is less clear. The strong association of *df* and *N* leading to fewer rejections is clearly seen in Figure 2 which graphs some of the data in Tables 1 and 2.

Replicating and extending the results of Chen et al. (2008), we have also found that when both the *df* and *N* were large, we virtually never found that the RMSEA exceeded the cutoff, even the lowest cutoff of 0.05. Further, when the sample size was very small (i.e., $N = 50$), the RMSEA exceeded the cutoff of .05 29–44 percent of the time, regardless of the number of *df*. Thus, the RMSEA is elevated with very small sample sizes, and the RMSEA was problematic regardless of the *df*. However, extending the findings in Chen et al. (2008), we found that for very small *df*, the RMSEA was substantially elevated for correctly specified models, even in moderate and large samples. For instance, when $N = 400$, the RMSEA exceeded .05 15.4 percent of the time in 1 *df* models, 14 percent of the time in 2 *df* models, and 12.7 percent of the time in 3 *df* models. Because these are correctly specified models, we would hope that the RMSEA would never exceed 0.05. In contrast, the RMSEA exceeded the cutoff only .04 percent of the time in 20 *df* models and never exceeded the cutoff of .05 in 50 *df* models. This difference is striking. In sum, replicating the work of Chen et al., at very small sample sizes (at or below 100), RMSEA cutoffs reject substantial proportions of correctly specified models. Extending the work of Chen et al., we find that small *df* studies far too often exceed cutoff values for the RMSEA. Sample size does appear to somewhat moderate the effect of low *df* on model rejection rates. Models with very small *df* and very large sample sizes do have lower model rejection rates than models with very small *df* and moderate sample sizes. However, models with very low *df* do have elevated RMSEA, even with sample sizes as large as 1,000. For instance, in our simulation, with sample sizes of 1,000, the observed RMSEA exceeded a cutoff of .05 5.7 percent of the time for our correctly specified 1 *df* model, 3.2 percent for our correctly specified 2 *df* model, and 1.7 percent of the time for our correctly specified 3 *df* model. At the heart of the problem is that such studies have more variability in χ^2 which in turn leads to greater variability of the RMSEA, a topic that we now discuss.

Estimation Accuracy of RMSEA

Because the simulation model is correctly specified, the population value of the RMSEA equals zero, and so the sample mean represents its bias. As seen in Table 3, we see the expected result that this bias declines as the sample size increases. However, we also see that the bias is much larger for smaller *df* models. In fact, for $df = 1$ and $N = 50$, the bias exceeds 0.05! Additionally, the *SD*s of the RMSEA parallel the bias. Again, small *df* models have much larger *SD*s than large *df* models.

The bias and variance can be combined into a single number, the root mean square error (RMSE) which is defined as the square root of the bias squared plus the variance of the estimate. As can be seen in Table 3, in the models with small *df*, RMSE decreases as the sample size increases. Within each of the same sample sizes, as the model *df* increases, the estimation accuracy increases. The bias values are negligibly small for the larger sample sizes.

CI of the RMSEA

We also examined the CI of the RMSEA. First, we computed the average width of the CI. We see in Table 3, especially that when both *N* and *df* are small, the CI is widest. For instance, when *N* is 50 and *df* are 1, the width is over 16 times wider than when *N* is 1,000 and *df* are 100. Quite clearly, there is a great deal of uncertainty about the RMSEA in the case of small *N* and small *df*. Next, we examined the coverages of the CI or the percentage of time zero was in the CI or coverage. We note that coverage should be .95 for a 90 percent CI because only the lower value is examined. We replicated the findings of Curran et al. (2003) that the CI was too narrow for small samples. However, we can extend their result: In our study, for small *df* models, the coverages are quite good, even for small *N* models. For instance, we find 94.1 percent coverage for small *df* models (1, 2, and 3) with an *N* of 50. However, as for 50 *df* and an *N* of 50, the coverage drops to 86.8 percent. A careful

Table 3. Descriptive Statistics and Power.

df	*N*	*M* RMSEA	*SD* RMSEA	RMSE[a]	*M* 90 Percent CI Width	Coverage[b]	Power[c]
1	50	0.052	0.089	0.103	0.316	0.939	0.069
1	100	0.034	0.060	0.069	0.221	0.951	0.085
1	200	0.022	0.041	0.047	0.155	0.954	0.109
1	400	0.017	0.03	0.035	0.11	0.951	0.148
1	600	0.014	0.024	0.028	0.091	0.951	0.181
1	1,000	0.011	0.019	0.022	0.069	0.955	0.243
2	50	0.047	0.074	0.087	0.241	0.945	0.075
2	100	0.035	0.052	0.063	0.172	0.946	0.097
2	200	0.024	0.037	0.044	0.124	0.944	0.134
2	400	0.016	0.026	0.031	0.086	0.944	0.199
2	600	0.014	0.022	0.026	0.071	0.938	0.259
2	1,000	0.011	0.016	0.020	0.054	0.943	0.368
3	50	0.046	0.067	0.082	0.211	0.939	0.081
3	100	0.033	0.047	0.058	0.148	0.948	0.108
3	200	0.022	0.034	0.040	0.105	0.942	0.157
3	400	0.016	0.023	0.028	0.074	0.959	0.247
3	600	0.013	0.019	0.023	0.061	0.949	0.330
3	1,000	0.010	0.015	0.018	0.046	0.952	0.476
5	50	0.044	0.059	0.073	0.175	0.941	0.089
5	100	0.031	0.043	0.053	0.124	0.945	0.127
5	200	0.020	0.029	0.036	0.085	0.947	0.199
5	400	0.015	0.020	0.025	0.061	0.962	0.335
5	600	0.012	0.016	0.02	0.050	0.951	0.456
5	1,000	0.009	0.013	0.016	0.038	0.953	0.650
10	50	0.044	0.051	0.067	0.146	0.930	0.107
10	100	0.029	0.036	0.046	0.101	0.932	0.169
10	200	0.019	0.025	0.031	0.070	0.939	0.294
10	400	0.013	0.017	0.022	0.049	0.943	0.520
10	600	0.011	0.014	0.018	0.040	0.944	0.691
10	1,000	0.008	0.011	0.013	0.031	0.947	0.886
20	50	0.040	0.044	0.059	0.117	0.911	0.136
20	100	0.025	0.030	0.038	0.080	0.930	0.241
20	200	0.017	0.021	0.027	0.056	0.935	0.454
20	400	0.011	0.014	0.018	0.039	0.951	0.766
20	600	0.009	0.011	0.015	0.032	0.951	0.913
20	1,000	0.007	0.009	0.012	0.025	0.944	0.991
50	50	0.042	0.035	0.055	0.096	0.868	0.210
50	100	0.023	0.023	0.033	0.064	0.931	0.424
50	200	0.014	0.016	0.022	0.043	0.929	0.769

(continued)

Table 3. (continued)

df	*N*	*M* RMSEA	*SD* RMSEA	RMSE[a]	*M* 90 Percent CI Width	Coverage[b]	Power[c]
50	400	0.009	0.011	0.014	0.029	0.950	0.981
50	600	0.007	0.009	0.012	0.024	0.944	0.999
50	1,000	0.006	0.007	0.009	0.019	0.958	1.000

Note: CI = confidential interval.

[a] The root mean square error or the square root of the bias squared plus the variance in estimate of the RMSEA.

[b] The proportion of time zero is in the CI, the expectation being .95.

[c] The power to reject the null hypothesis that the RMSEA = 0.05 when the actual RMSEA = 0.08.

examination of the results in Curran et al. (2003) reveals the very same pattern. Coverage tends to be poor for large *df*, small *N* studies. This study, however, shows that this poor coverage for small *N* studies is hardly apparent for small *df* studies.

Overall, the incorporation of CI of RMSEA in assessing model fit reinforced the results obtained from the point estimate of RMSEA. In other words, RMSEA was more precisely estimated for models with small *df* as the sample size increased. Also, the model rejection rates for the test of close fit using point estimate of RMSEA for models with small degrees had seriously high rejection rates, unless the models had extremely large sample size, regardless of the cutoff values chosen.

Our results are consistent with the findings of Chen et al. (2008) such that the model rejection rates were remarkably lower when the lower band of 90 percent CI used as opposed to use of point estimate of RMSEA. As can be seen in Table 3, both models with low *df* and large *df* the rejection rates were less than 5 percent in all models regardless of the sample size. However, when the upper 90 percent CI was used, RMSEA behaved worse unlike the use of lower band.

Power of RMSEA

Although we did not simulate models with specification error, we can employ a method for power analyses developed by MacCallum et al. (1996). We used their procedure to compute the power which is probability of rejecting the hypothesis of the test of close fit (the null value of the RMSEA = 0.05) when the true model is reasonable fit (alternative value of the RMSEA = 0.08). These results are presented in the last column of Table 3.

For models with 1 *df* sample size and an *N* of 50, the power to reject a model of close fit is only .069, whereas the power for a model with 50 *df* and an *N* of 200 is .769. Even with a sample size of 1,000, for 1 *df* models the power is only .243. Thus, we see that there is very little power in testing the RMSEA in models with small *df*, especially with small sample sizes. Note that with 50 *df*, acceptable levels of power are achieved with reasonable sample sizes (e.g., 200). These power analyses illustrate that models with low *df* have low power to reject a model of close fit. Therefore, even though the RMSEA is elevated in correctly specified models, using RMSEA cutoffs in small *df* models may not provide adequate power to reject moderately misspecified models.

What Is to Be Done?

Clearly, what Jane Miller experienced is not unique. For models with small *df*, the RMSEA can exceed cutoffs very often, even when the model is correctly specified. The major purpose of this article is to document that finding, and we have done so both using statistical theory and simulation results. Following the previous advice of others, we discourage the practice of examining the point estimate of the RMSEA and comparing it with some arbitrary cutoff point. For instance, Millsap (2007) stated that ‘‘research has challenged the generality of some of the commonly used thresholds for approximate fit indices such as the RMSEA’’ (p. 875). This point has been made repeatedly in the literature, but it continues to be ignored not only by practitioners but also by reviewers and editors.

We might think that researchers, like Jane Miller, could examine the CI and determine whether the desired value of the RMSEA (e.g., 0.00 or 0.05) is within that interval. However, for Jane Miller and her single degree of freedom model, the 90 percent CI limits are 0.0 and 0.269. This CI suggests that Jane Miller's model is likely somewhere between perfect and extremely horrible! Clearly, any RMSEA value with a CI this wide is of no value. Another alternative would be to test whether the sample RMSEA is statistically significantly larger than some desired value of the RMSEA (i.e., a test of close fit). However, for small *df* models, especially those with not very large sample sizes, the width of the CI would be large and the power of the test of close fit would be low (see Table 3).

If cutoffs, the CI, and tests of Close Fit are of no help, what then is the solution to the Jane Miller problem? We offer three suggestions. First, we may not need a fit index at all. One of the main purposes of a fit index is to get around the problem that one has a reasonable model, even with a statistically significant χ^2 test. However, if χ^2 is not statistically significant, we fail to reject the null hypothesis that the covariance matrix equals the model implied covariance matrix. In such a case, we know that the model relatively closely reproduces the data. Of course, a nonsignificant χ^2 does not prove that the model is correct.

Second, in some cases, it might be possible to redesign the study, so that it has more *df* by adding more indicators in a latent variable model or adding more waves in a growth curve study. Thus, it may be possible to avoid having a low *df* model. For instance, changing a path model to a latent variable model by using items as indicators instead of using variables as observed variables would increase the number of *df* in the measurement model, and hence the overall number of *df* for the hybrid model. However, such a strategy may not always be feasible or desirable. For instance, adding another wave of data in a longitudinal study may be too costly. Additionally, adding new waves or more indicators might change the character of the model in important and subtle ways. For instance, estimating latent variables allows for the possibility of (unmodeled) covariances among the errors of the items, either within or across factors. This type of misspecification, which is not detectable when estimating path models using observed variables, may also result in worse fit. Of course, in some of these situations, there may be no way around a low *df* model, which leads us to our third and final suggestion.

Third, we suggest not computing the RMSEA for very low *df* models that do not have a large sample size.[1] What then replaces the RMSEA for the small *df* model? If the model has small *df*, that means that there are a few extra parameters that could have been estimated but were not. We suggest, when possible, determining what these parameters are and estimating them. Following suggestions of McDonald (2010) and West et al. (2012), we urge focusing on the sensitivity of parameters that were not included in the model. For instance, in a three-wave growth curve model, the extra parameter is most plausibly the slope loading for the second wave (assuming the loading for wave 1 is fixed to zero). Instead of fixing the second loading to 1 and the third loading to 2 as is conventionally done, the time 2 loading could be freed. Then, the researcher could test how far the new free loading is from one. As another example, for most path analyses, there are usually missing paths (i.e., paths fixed to zero). The most plausible missing paths could be reintroduced and tested. For instance, for the Fishbein and Ajzen (1975) model, there are two paths fixed to zero, from Attitudes and Social Norms to Behavior. These two paths can be tested and it can be determined if they are needed or not. In feedback models, the *df* indicate that are extra instrumental variables. For a single-factor model with four indicators, there are several alternatives about what two parameters to add, for example, two correlated errors or a second factor with a correlated error. For any very low *df* model, it should be possible to think of plausible parameters that might be added that would turn the model into a saturated model. Such a strategy is sensible for very small *df* models, but it is not viable for moderate to large *df* models.

For low *df* models, it makes more sense to estimate a model in which the constrained parameter is free and report its parameter value and CI. We would suggest that Jane Miller adopt this last strategy. She has one missing path in her study, and she should report its CI and note that whether zero is in that interval. For low *df* studies with small *N*, the power to reject the null hypothesis that the parameter value is 0 would be low; however, in such scenarios, the CI would be quite wide, indicating that the parameter is estimated with a great deal of uncertainty and imprecision. In such cases, the lack of statistical significance for the parameter, coupled with the large CIs for that and other parameter estimates, should call into question the ability of the model to give us a clear picture about the underlying mechanisms of interest.

Conclusion

Using the RMSEA to assess the model fit in models with small *df* is problematic and potentially misleading unless the sample size is very large. We urge researchers, reviewers, and editors not to dismiss models with large RMSEA values with small *df* without examining other information. In fact, we think that it advisable for researchers to completely avoid computing the RMSEA when model *df* are small. In such cases, poor fit can be diagnosed by specifying additional models that include deleted parameters and determining if those additional parameters are needed in the model.

Our study does have limitations. All of the specified models here in this study met the assumption of multivariate normality. We do hypothesize that when data fail to conform to this assumption, the behavior of RMSEA might possibly be worse than what we found. That is, nonnormality tends to further inflate χ^2 (Curran et al. 1996) which would exacerbate the problem that we have described. Also our simulation did not include any specification error.

As MacCallum (1990) and others have suggested, it is important to realize that no one model fit index can be considered to be the litmus test of a good fitting model. Researchers should examine the fit of their models in conjunction with the residual covariance matrices and interpret this information using an underlying theoretical framework to support their results. Reliance on any one number to determine if one has a good model is an almost certain recipe for disaster.

Acknowledgment

We thank Thomas Ledermann and Randi Garcia who provided helpful comments on earlier version of this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. Not only is the RMSEA a less than useful measure of fit for small *df* models, but we believe, but do not know, that other measures of fit would also show high variability for those models. When we examined other measures of fit from our simulation, the results for those measured parallel those for the RMSEA. We suspect that all fit indices perform poorly with low *df* models because fit indices depend in large part on the χ^2 statistic and that value is much more variable with low *df* than with high *df*. Due to the space limitations, we are unable to report these details from our simulation. For more information, please refer to http://davidakenny.net/papers/RMSEA_small_df/details.pdf.

References

Bentler, Peter M. and Douglas G. Bonett. 1980. "Significance Tests and Goodness of Fit in the Analysis of Covariance Structures." *Psychological Bulletin* 88:588-606.

Breivik, Einar and Ulf H. Olsson. 2001. "Adding Variables to Improve Fit: The Effect of Model Size on Fit Assessment in LISREL." Pp. 169-94 in *Structural Equation Modeling: Present and Future*, edited by Robert Cudeck, S. H. C. du Toit, and Dag Sörbom. Lincolnwood, IL: Scientific Software International.

Browne, Michael W. and Robert Cudeck. 1993. "Alternative Ways of Assessing Model Fit." Pp. 136-62 in *Testing Structural Equation Models*, edited by Kenneth A. Bollen and J. Scott Long. Newbury Park, CA: Sage.

Chen, Feinian, Patrick J. Curran, Kenneth A. Bollen, James B. Kirby, and Pamela Paxton. 2008. "An Empirical Evaluation of the Use of Fixed Cutoff Points in

RMSEA Test Statistic in Structural Equation Models." *Sociological Methods and Research* 36:462-94.

Curran, Patrick J. 2000. "A Latent Curve Framework for Studying Developmental Trajectories of Adolescent Substance Use." Pp. 1-42 in *Multivariate Applications in Substance Use Research*, edited by Jennifer S. Rose, Laurie Chassin, Clark C. Presson, and Steven J. Sherman. Hillsdale, NJ: Erlbaum.

Curran, Patrick J., Kenneth A. Bollen, Feinian Chen, Pamela Paxton, and James B. Kirby. 2003. "Finite Sampling Properties of the Point Estimates and Confidence Intervals of the RMSEA." *Sociological Methods and Research* 32:208-52.

Curran, Patrick J., Kenneth A. Bollen, Pamela Paxton, James B. Kirby, and Feinian Chen. 2002. "The Noncentral Chi-square Distribution in Misspecified Structural Equation Models: Finite Sample Results from a Monte Carlo Simulation." *Multivariate Behavioral Research* 37:1-36.

Curran, Patrick J., Stephen G. West, and John Finch. 1996. "The Robustness of Test Statistics to Non-normality and Specification Error in Confirmatory Factor Analysis." *Psychological Methods* 1:16-29.

Fishbein, Martin and Icek Ajzen. 1975. *Belief, Attitude, Intervention, and Behavior: An Introduction to Theory and Research*. Reading, MA: Addison-Wesley.

Frone, Michael R., Marcia Russell, and M. Lynn Cooper. 1994. "Relationship between Job and Family Satisfaction: Causal or Noncausal Covariation?" *Journal of Management* 20:565-79.

Heath, Andrew C., Michael C. Neale, John K. Hewitt, Lindon J. Eaves, and David W. Fulker. 1989. "Testing Structural Equation Models for Twin Data Using LISREL." *Behavior Genetics* 19:9-36.

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

Jackson, Dennis L., J. Arthur Gillaspy, and Rebecca Purc-Stephenson. 2009. "Reporting Practices in Confirmatory Factor Analysis: An Overview and Some Recommendations." *Psychological Methods* 14:6-23.

Kaugars, Astrida S., Mary D. Klinnert, Jane L. Robinson, and Martin Ho. 2008. "Reciprocal Influences in Children's and Families' Adaptation to Early Childhood Wheezing." *Health Psychology* 27:258-67.

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

Kim, Kevin H. 2005. "The Relation Among Fit Indexes, Power, and Sample Size in Structural Equation Modeling." *Structural Equation Modeling* 12: 368-90.

Llabre, Maria M., Susan B. Spitzer, Scott D. Siegel, Patrice G. Saab, and Neil Schneiderman. 2004. “Applying Latent Growth Curve Modeling to the Investigation of Individual Differences in Cardiovascular Recovery from Stress.” *Psychosomatic Medicine* 66: 29-34.

MacCallum, Robert C. 1990. “The Need for Alternative Measures of Fit in Covariance Structure Modeling.” *Multivariate Behavioral Research* 25:157-62.

MacCallum, Robert C., Michael W. Browne, and Hazuki M. Sugawara. 1996. “Power Analysis and Determination of Sample Size for Covariance Structure Modeling.” *Psychological Methods* 1:130-149.

McAuley, Edward, Gerald J. Jerome, David Xavier Marquez, Steriani Elavsky, and Bryan Blissmer. 2003. “Exercise Self-efficacy in Older Adults: Social, Affective, and Behavioral Influences.” *Society of Behavioral Medicine* 25:1-7.

McDonald, Roderick P. 2010. “Structural Models and the Art of Approximation.” *Perspectives on Psychological Science* 5:675-86.

McDonald, Roderick P. and Moon-Ho R. Ho. 2002. “Principles and Practice in Reporting Structural Equation Analyses.” *Psychological Methods* 7:64-82.

Millsap, Roger E. 2007. “Structural Equation Modeling Made Difficult.” *Personality and Individual Differences* 42:875-881.

Muthén, Linda K. and Bengt O. Muthén. 2007. *Mplus User’s Guide. Version 5*. Los Angeles, CA: Muthén & Muthén.

Nevitt, Jonathon and Gregory 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.

Rigdon, Edward E. (1996). “CFI versus RMSEA: A Comparison of Two Fit Indexes for Structural Equation Modeling.” *Structural Equation Modeling* 3: 369-79.

Segrin, Chris, Terry A. Badger, Paula Meek, Ana M. Lopez, Elizabeth Bonham, and Amelia Sieger. 2005. “Dyadic Interdependence on Affect and Quality of Life Trajectories among Women with Breast Cancer and Their Partners.” *Journal of Social and Personal Relationships* 22:673-89.

Steiger, James H. and John 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.

West, Stephen G., Aaron B. Taylor, and Wei Wu. (2012). “Model Fit and Model Selection in Structural Equation Modeling.” Pp. 209-31 in *Handbook of Structural Equation Modeling*, edited by R. H. Hoyle. New York: Guilford.

Williams, Larry J. and John T. Hazer. 1986. “Antecedents and Consequences of Satisfaction and Commitment in Turnover Models: A Reanalysis Using Latent Variable Structural Equation Methods.” *Journal of Applied Psychology* 71: 219-31.

Author Biographies

David A. Kenny is professor emeritus in Psychology at the University of Connecticut. He is the author of six books and has written extensively in the areas of mediational analysis, interpersonal perception, and the analysis of social interaction data. He was elected as a fellow of the American Academy of Arts and Sciences and was the inaugural winner of the Society of Personality and Social Psychology's Methodological Innovation award.

Burcu Kaniskan is a research psychometrician at National Conference of Bar Examiners (NCBE) where conducts full-service psychometric activities including calibration, equating, and scaling, item and student data analysis of the Multistate Bar Examination and the Multistate Professional Responsibility Examination, psychometric research and consulting. Burcu Kaniskan earned her PhD from Measurement, Evaluation, and Assessment at the University of Connecticut. Her dissertation, which compared the prediction and projection accuracy of several growth models using a state assessment program in mathematics and reading, won the Outstanding Dissertation Award from American Educational Research Association (AERA) Division H: Research, Evaluation, and Assessment in Schools.

D. Betsy McCoach is professor and program coordinator of the measurement, evaluation, and assessment program in the Educational Psychology department at the University of Connecticut. Her methodlogical research interests include structural equation modeling, hierarchical linear modeling, longitudinal data analysis, and instrument design in the affective domain. Her substantive research interests are in gifted education, and she is the current editor of *Gifted Child Quarterly*.