STRUCTURAL EQUATION MODELING, 7(2), 149–162

Point Estimation, Hypothesis Testing, and Interval Estimation Using the RMSEA: Some Comments and a Reply to Hayduk and Glaser

James H. Steiger
Department of Psychology
University of British Columbia

Hayduk and Glaser (2000) asserted that the most commonly used point estimate of the Root Mean Square Error of Approximation index of fit (Steiger & Lind, 1980) has two significant problems: (a) The frequently cited target value of .05 is not a stable target, but a "sample size adjustment"; and (b) the truncated point estimate $\hat{R}_t = \max(\hat{R}, 0)$ effectively throws away a substantial part of the sampling distribution of the test statistic with "proper models," rendering it useless a substantial portion of the time. In this article, I demonstrate that both issues discussed by Hayduk and Glaser are actually not problems at all. The first "problem" derives from a false premise by Hayduk and Glaser that Steiger (1995) specifically warned about in an earlier publication. The second so-called problem results from the point estimate satisfying a fundamental property of a good estimator and can be shown to have virtually no negative implications for statistical practice.

Steiger and Lind (1980) introduced the notion of noncentrality interval estimation as a method for assessing the fit of a structural equation model. The fundamental rationale behind the method is that the goal of testing whether a model is perfect should be replaced by the dual goals of ascertaining (a) how good model fit is in the population and (b) how precisely we have determined it from the sample data.

Steiger and Lind (1980) pointed out that the comparison of nested models is complicated by the fact that, for any data set and for any population, the more complex model will always fit at least as well and usually better. Consequently, some compensation for model complexity is desirable in a measure of fit.

Requests for reprints should be sent to James H. Steiger, Department of Psychology, University of British Columbia, Vancouver, B.C., Canada V6T 1Z4. E-mail: steiger@unixg.ubc.ca

The RMSEA (root mean square error of approximation) fit index draws on the fact that the discrepancy functions commonly employed in χ^2 tests in structural equation modeling are either equal to or closely approximated by a weighted sum of squared model discrepancies. Consequently, under fairly general circumstances, the *population discrepancy function* F^*, defined as the discrepancy function one would obtain if the estimation technique were applied to the population covariance matrix, is a not too unreasonable measure of population fit. Although this measure is far from problem-free, it has some real advantages, most significant of which are that it can be subjected to statistical tests and estimated with a confidence interval.

To compensate for model complexity, and to return the index of fit to the original metric of the covariance matrix, Steiger and Lind (1980) defined the RMSEA, for models with ν degrees of freedom, as

$$R = \sqrt{\frac{F^*}{\nu}} \tag{1}$$

Steiger and Lind (1980) concentrated on interval estimation of R, and Steiger (1989) implemented the interval estimation technique in the computer program EZPATH. Steiger also showed that confidence interval estimates and hypothesis tests could be produced for population counterparts of the GFI and AGFI indexes of Jöreskog and Sörbom (1984).

In their commentary, Hayduk and Glaser (2000) expressed concern about a commonly used point estimate for the RMSEA. Throughout this article, I will refer to the number of observations as N, and for simplicity of notation, define $n = N - 1$. Define F as the maximum likelihood (or GLS or IRGLS) discrepancy function (calculated on a sample of size N), $U = nF$ the χ^2 goodness of fit statistic. The *truncated estimator* $\hat{R}_t$ is defined as

$$\hat{R}_t = \sqrt{\max\left(\frac{U/\nu - 1}{n}, 0\right)} \tag{2}$$

The quantity under the radical is a point estimate of the squared RMSEA. It is commonly bounded at zero to satisfy an optimality criterion (defined and discussed later) I call the *primacy principle*. If one uses the square root of this quantity, as in Equation 2, truncating negative values at zero has the added benefit of eliminating imaginary values of $\hat{R}_t$.

Point estimates are frequently used directly in hypothesis testing, but not in the case of the RMSEA. Because of the truncation at zero in equation 2, the χ^2 statistic itself, rather than some function of the point estimate, is used directly for hypothesis testing and the calculation of associated probability levels under a null hypothe-

sis. MacCallum, Browne, and Sugawara (1996) discussed hypothesis tests, power analysis, and sample size estimation based on the use of the RMSEA.

Hayduk and Glaser (2000) expressed two primary concerns about the RMSEA:

1. They examined Equation 2 and discussed the view that the truncation at zero throws away information. They leave the reader with the vague impression that this somehow creates practical problems and that traditional probability values calculated for the RMSEA need to be corrected in some way. Hayduk expanded on this notion in a series of postings to the Internet discussion group SEMNET. For example, on March 29, 1999, he stated,

> We (via MAX) have changed the sampling distribution in a problematic way. In the true-model case, we have thrown away the bottom/left half of the distribution, so 5% OF WHAT IS LEFT OF THE DISTRIBUTION does not correspond to the top 5% of the overall sampling distribution. One has to use about 10% of the remaining top/right half of the distribution to cover what corresponds to the usual top 5% of the FULL sampling distribution.

2. They imagined that the RMSEA incorporates a hidden sample size correction that makes it "an elastic tape measure that four-steppers, and others, can stretch to let them accept models they would like to accept."

Like Hayduk and Glaser (2000), I have my own serious criticisms of both the RMSEA itself and the way that it is employed in some situations. For example, I am only partly comfortable with the very serious incorporation of the index into a Neyman–Pearson hypothesis-testing framework, and even less comfortable with rigid adherence to fixed target values. My original intention was that the RMSEA be employed in a more relaxed, heuristic fashion, both as an improvement on and a release from hypothesis testing.

However, my criticisms are quite different from those of Hayduk and Glaser (2000). Unfortunately, as I show in the following sections, their line of argument has some serious flaws.

WHY USE A TRUNCATED POINT ESTIMATE?

The quantity to be estimated, R, the RMSEA, is defined on a statistical population. Any point estimate of this quantity should satisfy certain optimality criteria, at least to a reasonable approximation.

Hayduk and Glaser (2000) failed to distinguish, in their discussion, between the particular point estimate, $\hat{R}$, which they found fault with, and the population quantity, R, which the point estimate is designed to measure. They do not mention that (a) the test statistic used for hypothesis tests and probability level calculation does not employ the truncated estimate, (b) the truncated estimate satisfies an obvious

optimality criterion, and (c) hypothesis tests need not employ the best point estimate for a parameter. Because these basic principles have eluded them, some review is in order.

Elementary textbooks on behavioral statistics discuss qualities of a good estimator, usually concentrating on unbiasedness, consistency, and efficiency. These three principles follow from more general, common-sense notions, which might be summarized as follows: (a) Seldom, in fact very seldom, will a statistic be exactly equal to the parameter it is estimating, if the parameter space is continuous; (b) consequently, for any observation on a statistic, there will be a sampling error, ε, defined as the difference between the statistic and the parameter it is estimating; (c) usually we prefer that errors not be systematically positive or negative by more than a trivial amount; and (d) we prefer a statistic that makes small errors over one that makes large errors. There are occasionally difficult tradeoffs among these principles. Principle d gives rise to the notion of efficiency, that is, that all other things being equal we prefer an estimator with low sampling variability. However, efficiency is a long-run behavior of a statistic, and, in practice, we often have only one chance to estimate a parameter. That is, we have the data in front of us, and we have to make the best effort to estimate the parameter. Suppose I were to tell you that I have two statistics that are candidates for estimates of a parameter and that they have committed errors of estimate with the same algebraic sign, but that one has committed an absolute error that is guaranteed to be less than or equal to the other. Most individuals would immediately choose the statistic with the possibly smaller error. This suggests a simple principle in statistical estimation, which I define in the following.

The Primacy Principle of Statistical Estimation

Given a statistical population P, a parameter R of that population, and a sample X of N independent vectors of observations from P, let two statistics, $\hat{R}_1$ and $\hat{R}_2$, both be functions of X used to estimate parameter R. $\hat{R}_1$ has sampling error $\varepsilon_1 = \hat{R}_1 - R$, and $\hat{R}_2$ has sampling error $\varepsilon_2 = \hat{R}_2 - R$. Let $\text{sgn}(y)$ be the sign function, with values $-1, 0, +1$ for negative, zero, and positive values of y, respectively. Suppose that, for any X sampled from P, ε_1 and ε_2 never have opposite signs (in the sense that $\text{sgn}(\varepsilon_1)\,\text{sgn}(\varepsilon_2) \geq 0$). Suppose further that, for any X, $|\varepsilon_1| \leq |\varepsilon_2|$. Then we say that $\hat{R}_1$ *has primacy over* $\hat{R}_2$.

For any data set you will ever have, the statistic that has primacy over a competitor will always be as close to the true parameter as the competitor, and on some occasions it will be closer. Except in rather unusual circumstances, this would seem to be a powerful advantage.

The primacy principle arises quite naturally in connection with bounded parameter spaces, when (a) the nonpreferred statistic is unbiased, or close to it,

and (b) the actual parameter value is close to the boundary of the parameter space. Take, for example, multiple regression, when we are trying to estimate the squared multiple correlation, R_p^2. In this case, the parameter space (the set of all parameter values) ranges from 0 to 1. Unfortunately, when the sample size is small relative to the number of predictors, the "raw" sample squared multiple correlation R^2 is a rather positively biased estimate of the corresponding population quantity. Consequently, an "adjusted" estimator is reported frequently. If there are p predictor variables, and the sample size is N, the adjusted formula is

$$R_a^2 = \frac{(N-1)R^2 - p}{N-p-1} \tag{3}$$

Although it is seldom computed, a unique unbiased estimate of R_p^2 is also available (Olkin & Pratt, 1958) and is (for $n > p \geq 2$) given by

$$R_u^2 = 1 - \frac{N-3}{N-p-1}(1-R^2)F(1,1,(N-p+1)/2,1-R^2) \tag{4}$$

(F in the above refers to a hypergeometric function, not the F distribution.) Frequently, when sample size is small and the relationship between variables is weak, both R_a^2 and R_u^2 can be negative. In this case a "preferred statistic" satisfying the primacy principle may be constructed by setting all negative values to zero. For the unbiased estimator,

$$R_{u+}^2 = \max(R_u^2, 0) \tag{5}$$

Because the true parameter is known to be between 0 and 1, it is a certainty that moving the negative estimate to zero reduces sampling error. It is one of those golden moments in statistics when one can be absolutely certain one is improving the quality of an estimate! Of course, the preferred estimate is no longer unbiased, but in this case, the tradeoff seems acceptable. Indeed, as Kendall and Stuart (1979) pointed out, it is impossible, when a parameter space is between 0 and 1, to construct an unbiased estimator that always takes on values between 0 and 1. To eliminate what they call "the absurdity of negative estimates," they suggest the same modification that we employ.

Point estimation of the RMSEA has much in common with point estimation of the squared multiple correlation. Because the RMSEA is by definition nonnegative, an estimate that is unbiased (or close to it) will of necessity take on negative values for some values of the population parameter. For such an estimate, eliminating those values, by setting them equal to zero, is guaranteed to reduce error of estimation for those values.

Hence it seems that the objection of Hayduk and Glaser (2000) is ill founded, if the RMSEA is used as originally intended, that is, to estimate fit in a single model. It is the square root of an estimate that satisfies the statistical primacy principle, relative to an untruncated competitor.

THE NONEXISTENT "SAMPLE SIZE ADJUSTMENT"

Hayduk and Glaser (2000) made some incorrect assumptions about the relationship between point estimation of the RMSEA (and the proper statistical goals associated with such estimation) and inference about a population value of the RMSEA.

For example, Hayduk and Glaser (2000) computed and plotted (their Figure 2) χ^2/ν values corresponding to sample values of the RMSEA point estimate. Their graph demonstrated that, if one chooses a fixed "target value" of the point estimate $\hat{R}_t$ (say .05), one finds that the χ^2/ν value corresponding to a target value changes as a function of sample size. Because these changing values of χ^2/ν correspond to changing probability levels, Hayduk and Glaser claimed to have discovered that a fixed value of the RMSEA represents a "sample size adjustment" to a criterion for good fit and that this represents a defect that must be corrected.

There are several problems with this line of reasoning. One is the assumption that a probability level of a statistic has a static interpretation in terms of model fit. This misconception has been debunked so thoroughly and so often by modern writers on hypothesis testing and effect size estimation that its appearance is surprising. Steiger and Fouladi (1997) examined this notion in connection with a two-way analysis of variance (ANOVA) and demonstrated its falsity. Within the same ANOVA, different tests having the same probability level can have different implications about effect size.

In a similar way, χ^2 statistics based on different sample sizes, having the same probability levels, will have different implications about the population RMSEA. Yet Hayduk and Glaser (2000) appeared to assume, as the cornerstone of their argument, that constant ratios of χ^2/ν have the same statistical meaning. In other words, Hayduk and Glaser have turned the truth on its ear and are now shaking it excitedly. Needless to say, the truth is suffering.

What is particularly ironic is that, in the documentation (Steiger, 1995) to my program SEPATH, I (a) derived the formula that generates the results in Figure 2 of Hayduk and Glaser (2000), and (b) carefully explained the fallacy of assuming that a χ^2/ν ratio has a constant meaning.

In what follows, I review the actual statistical facts and reveal exactly where Hayduk and Glaser (2000) went wrong. I then demonstrate that their logic, transferred to another domain, would imply that some familiar measures of effect size in ANOVA (such as the sum of squared standardized effects) are invalid.

Recall the philosophy behind estimation of the RMSEA. It is that the population badness of fit is seldom zero, and our task is to ascertain how large it is and how precisely we have determined it. Yet early in their argument, Hayduk and Glaser (2000) discuss probability levels, derived under the assumption that population badness of fit is zero (i.e., that fit is perfect).

Assume, for the time being, that an RMSEA of a given value (say .05) in the population has a fixed interpretation (later we will question this assumption, but for reasons completely different from those of Hayduk and Glaser [2000]). Our job as statisticians is to estimate the population RMSEA.

Steiger, Shapiro, and Browne (1985) demonstrated that, under reasonable assumptions, the χ^2 statistic $U = nF$ has a distribution that is approximately $\chi^2_{\nu,\lambda}$, where λ, the noncentrality parameter, is given by

$$\lambda = nF^* \tag{6}$$

The expected value of a $\chi^2_{\nu,\lambda}$ variate is

$$E(U) = \nu + \lambda \tag{7}$$

Hence, a bias-corrected (based on the asymptotic result) point estimate of F^* is

$$\hat{F}^*_{bc} = (U - \nu)/n \tag{8}$$

Whenever U is less than ν, the unbiased estimate will be negative. However, the primacy principle dictates that, on any occasion where that occurs, we can always obtain a better estimate (i.e., one that is closer to the parameter value than the unbiased estimate) by simply changing the value to zero. Hence the revised statistic,

$$\hat{F}^*_a = \max\{(U - \nu)/n, 0\} \tag{9}$$

There are numerous trivial consequences of Equations 6 and 7. For example, we note that

$$E(U) = \nu + nF^* \tag{10}$$

For constant degrees of freedom, this tells us that the mean of the χ^2 goodness-of-fit statistic will increase as a function of n, for a given value of F^*. Because R, the RMSEA, is, for fixed ν, a simple transformation of F^*, its mean will also increase as a function of n. Another way of viewing this is that, as n increases, it will require a larger χ^2 statistic to imply an equivalent R.

One obvious implication of this, stressed by Steiger (1995), is that the quantity χ^2/ν is not a stable indicator of the size of population badness-of-fit function F^*. Because Hayduk and Glaser appear to be unaware of my earlier discussion, I reproduce it here. It may be shown easily that a bound on the point estimate of R implies a corresponding bound on the ratio of the χ^2 statistic to its degrees of freedom. Specifically, suppose, for example, you have decided that, for your purposes, the point estimate of the RMSEA index should be less than some value c, that is,

$$\hat{R} < c \tag{11}$$

Steiger (1995) showed (his Equation 91) that this inequality implies

$$\frac{U}{\nu} < 1 + nc^2 \tag{12}$$

So, for example, the rule that the point estimate of the RMSEA should be less than .05 translates into the equivalent rule (Equation 92, Steiger [1995]) that

$$\frac{U}{\nu} < 1 + \frac{n}{400} \tag{13}$$

This agrees with the calculations of Hayduk and Glaser (2000).

However, the preceding result is not evidence that the RMSEA has a hidden "sample size adjustment." Rather, it is evidence that the noncentrality parameter is a function of n! It appears to be a sample size adjustment, if one adopts the mistaken belief that χ^2/ν ratios have a stable statistical implication. They do not. I specifically warned about this in Steiger (1995):

> Rules of thumb that cite a single value for a critical ratio of χ^2/ν ignore the point that the chi-square statistic has an expected value that is a function of degrees of freedom, population badness of fit, and N. Hence, for a fixed level of population badness of fit, the expected value of the chi-square statistic will increase as sample size increases. The rule of Equation 91 compensates for this, and hence it may be useful as a quick and easy criterion for assessing fit. To avoid misinterpretation, we should emphasize at this point that our primary emphasis is on a confidence-interval-based approach, rather than one based on point estimates. The confidence interval approach incorporates information about precision of estimate into the assessment of population badness of fit. Simple rules of thumb (such as that of Equation 91) based on point estimates ignore these finer statistical considerations. (p. 3670)

The notion that a χ^2/ν ratio has a stable statistical implication is a remnant of an earlier time, when the relationships between the likelihood ratio statistic, population

badness of fit, and the noncentral chi-square distribution were not yet fully understood. This notion is misguided in the present context, because the noncentrality parameter affects the mean and variance of the distribution of the test statistic and is itself a function of sample size. Equally erroneous is the notion that, across values of n, the probability level of an observed χ^2 has a constant interpretation. The only situation in which this supposition would be correct is if F^*, the population discrepancy function, is known or assumed to be zero. Of course, if that were true, it would be nonsensical to be estimating F^*.

The fallacy in the analysis of Hayduk and Glaser (2000) is grasped more clearly by constructing a very precise analogy between estimation of badness of fit in structural equation modeling and estimation of effect size in ANOVA.

Consider a one-way fixed effects ANOVA, with J groups and N observations per group. A reasonable measure of effect size is the Root Mean Square Standardized Effect, defined by Steiger and Fouladi (1997) as

$$\Psi = \sqrt{\frac{1}{J-1}\sum_{j=1}^{J}\left(\frac{\alpha_j}{\sigma}\right)^2} \tag{14}$$

(Note that the "averaging" is by the number of degrees of freedom, rather than the number of groups, to compensate for the one restriction imposed on the effects.) The standard F statistic for the test of nil effect in ANOVA has a distribution that is $F_{\nu 1,\nu 2,\lambda}$, i.e., noncentral F with ν_1 and ν_2 degrees of freedom, and noncentrality parameter λ. Degrees of freedom are $\nu_1 = J-1$ and $\nu_2 = J(N-1)$. There is a direct relationship between Ψ and λ, given by

$$\Psi = \sqrt{\frac{\lambda}{(J-1)N}} \tag{15}$$

Conversely, of course,

$$\lambda = (J-1)N\Psi^2 = \nu_1 N\Psi^2 \tag{16}$$

An unbiased estimator of λ is obtained from the observed F statistic as

$$\hat{\lambda} = \nu_1\left(\frac{\nu_2}{\nu_2 - 2}F - 1\right) \tag{17}$$

Consequently, an unbiased estimator of Ψ^2 is

$$\widehat{\Psi^2} = \frac{1}{N}\left(\frac{\nu_2}{\nu_2 - 2}F - 1\right) \tag{18}$$

Suppose you have a "target value" in mind for $\widehat{\Psi^2}$, representing a "minimally trivial" average squared standardized effect. If this target is c^2, there is a one-to-one correspondence between a target c^2 and the F statistic, just as there is with the RMSEA in structural modeling. Specifically, we have

$$F = \frac{v_2}{v_2 - 2}\left(N\widehat{\Psi^2} + 1\right) \tag{19}$$

Note that the value of F corresponding to a target value of $\widehat{\Psi^2}$ varies as a function of N. Again, this is because the noncentrality parameter is a function of N. Suppose, for simplicity, we confine ourselves to a four-group ANOVA. In that case, we have

$$F = \frac{\left(1 + N\widehat{\Psi^2}\right)(4N - 6)}{4N - 4} \tag{20}$$

Figure 1 shows the relationship between target values of F and sample size N for selected values of $\widehat{\Psi^2}$. Note the similarity between this figure and Figure 2 in Hayduk and Glaser (2000).

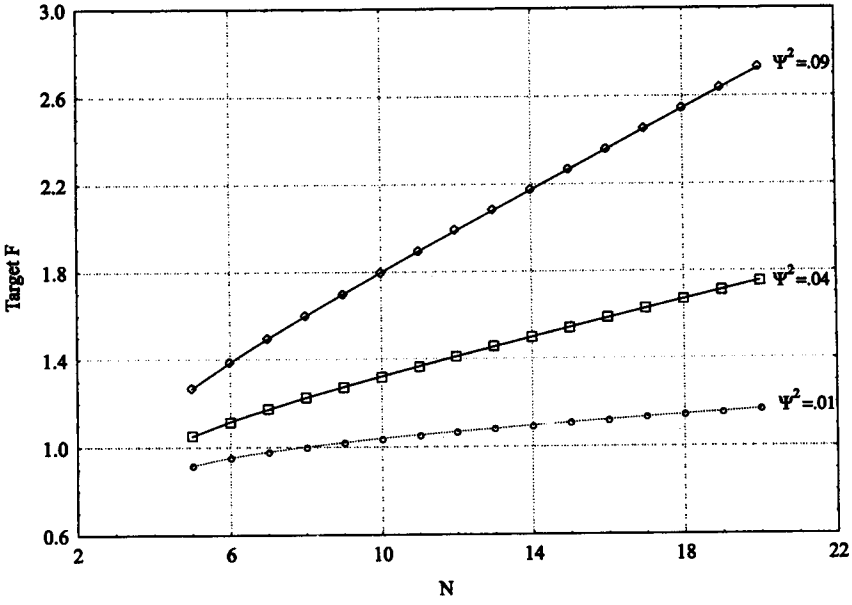

FIGURE 1 F values corresponding to target values of Ψ^2 for a one-way ANOVA with four groups.

Would one, after viewing Figure 1, conclude that the variance of standardized effects (or, alternatively, the variance of population means, in standard deviation units) is an ill-founded measure of effect size in ANOVA and incorporates a previously unnoticed "sample size adjustment"? One hopes not, but one might reach such a questionable conclusion following the logic of Hayduk and Glaser.

IMAGINARY PENALTIES OF TRUNCATION

Hayduk and Glaser (2000) expressed concern that the truncated point estimate of the RMSEA creates problems in practice and stated (incorrectly) that "Proper models would result in zero RMSEA values (via use of Max) about half the time, while nearly proper models would result in zero RMSEA values nearly half the time" (p. 28).

The statement is overly pessimistic, because it confuses "proper" with "perfect." Perfect models (i.e., those with zero RMSEA values) occur with probability very close to zero in practice.

The question arises, "How often in practice should we expect zero values of the RMSEA point estimate?" Note that this is an empirical question, and it appears that the answer is "very seldom," because only a tiny minority of the structural equation models ever published report χ^2 statistics that are less than or equal to their degrees of freedom. Using the asymptotic noncentral χ^2 approximation, one may estimate how frequently such estimates will be encountered. One simply evaluates the following probability:

$$\Pr(\chi^2_{\nu, n\nu R^2} < \nu) \tag{21}$$

For example, if the χ^2 statistic has 15 *df*, and if, using the guidelines in MacCallum et al. (1996), one chooses a sample size that will yield a power of at least .80 to reject the hypothesis that $R \leq .05$ when $R = .08$, values of the point estimate will be truncated at zero less than 1% of the time.

Consequently, the empirical evidence represented by several decades worth of published work in factor analysis and structural equation modeling suggests that Hayduk and Glaser are seriously overstating the extent of the problem caused by truncating the point estimate. Very few models fit well enough to generate point estimates of R that are zero, when sample size is adequate to establish reasonable precision.

The question then arises, "What damage is done on those occasions when the truncated point estimate is zero?" The answer would appear to be very little. The point estimate of the RMSEA is never used in statistical testing or interval estimation. The χ^2 statistic is used directly to construct the confidence interval, so no part of its sampling distribution is "thrown away" by truncating the point estimate. Because $\lambda = n\nu R^2$, it follows immediately that any hypothesis about R implies a

noncentral χ^2 distribution with a specific λ. So, for example, to test the hypothesis that $R \leq .05$, one simply compares the observed value of U to a noncentral χ^2 distribution with $\lambda = n\nu(.05)^2 = n\nu/400$. Contrary to what one might infer from the comments of Hayduk and Glaser (2000), for values of U that are less than or equal to ν, the reported probability level for the test statistic shows appropriate variation, although the point estimate remains (quite properly, according to the statistical primacy principle) at zero.

Clearly, if one is comparing two models with zero point estimates, one cannot differentiate between the models on the basis of these point estimates. This penalty of truncation would seem minimal, because this type of situation would tend to exist either (a) when population fit is actually very outstanding for both models or (b) when precision of estimate is inadequate to distinguish between models.

CONCEPTUAL FOUNDATIONS FOR THE RMSEA—SOME FURTHER COMMENTS

Although the misconceptions of Hayduk and Glaser (2000) render their commentary largely irrelevant to the use of the RMSEA in assessing the fit of structural models, I should stress that, in my view, this index, like any attempt to describe a multivariate quantity in a single summary measure, can certainly be questioned.

The RMSEA is a population measure. It is a simple transformation of the "population discrepancy function." A key problem is the extent to which the population discrepancy function adequately describes badness of fit.

Recall that the population discrepancy function, if it is from the GLS family, is a weighted sum of squared discrepancies. (The ML discrepancy function is closely approximated by an iteratively reweighted GLS discrepancy function.) The choice of the weight function is made to allow the discrepancy function to have a convenient asymptotic probability distribution.

For example, the GLS discrepancy function is of the form $e'We$, where e is a vector of discrepancies between the observed and reproduced covariance matrices and nW is the inverse of the asymptotic covariance matrix of the observed covariances. The key question is the extent to which the use of "sample directed" weighting (i.e., weighting that is used to accomplish desirable distribution of a sample quantity) creates problems in computing a measure of "population badness of fit."

To bring the problem into sharp focus, consider a simple special case: the hypothesis that the correlation ρ between two variables is ρ_0, a constant. In this case, for a given level of difference between the actual correlation ρ and the hypothesized correlation ρ, the value of the population discrepancy function becomes greater as ρ_0 departs from zero, primarily because the sampling variance of the correlation coefficient decreases as ρ approaches 1.

Some representative values are shown in Table 1. Note that as ρ_0 remains within the levels typically encountered, the statistical weighting of discrepancy

TABLE 1
Value of the Population Maximum Likelihood Discrepancy Function F^* and RMSEA R for the Hypothesis $\rho = \rho_0$

ρ_0	ρ	F^*	R
.00	.05	.0025	.0500
.25	.30	.0029	.0541
.45	.50	.0041	.0645
.65	.70	.0085	.0919
.85	.90	.0463	.2152

Note. RMSEA = root mean square error of approximation.

has little effect. However, as ρ_0 approaches 1, the impact of the weighting becomes substantial. Of course, what one considers a "proper measure" of the discrepancy between two correlations is itself open to debate and will therefore color one's perception of this phenomenon.

My space is limited here, and a full description of what I see as problems at the foundation of the RMSEA requires a different venue. However, if one considers the structure of the measure, it seems clear that the use of precise numerical "cutoff values" (like .05) should not be taken too seriously, anymore than a precise value for its analog, the RMSSE in ANOVA, or similar measures like "Cohen's *d*." In Steiger (1999), I described situations in ANOVA where different levels of experimental effect give rise to identical values for the RMSSE. So it is, in essence, impossible to define a single value of the RMSSE that conveys a given "level of experimental effect." In a rather analogous (but somewhat more serious) way, it is possible to take issue with any particular RMSEA value as a precise indicator of model fit. This does not render the measure useless, or even less than valuable, anymore than interval estimation of the root-mean-square standardized effect in ANOVA is rendered useless by the issues discussed in Steiger (1999). However, religious adherence to a particular numerical guideline not only violates the original spirit in which Steiger and Lind (1980) offered the measure, but also strikes me as a mistake.

The fundamental contributions of the RMSEA and the noncentrality interval estimation approach are that they rescue model-fitting from the destructive grasp of accept-support statistical testing and place it on a much firmer foundation, where increased sample size and experimental precision work for, rather than against, the investigator's interests. These contributions were an important step forward, and with careful critical analysis, may serve as the foundation for further improvements in structural modeling.

REFERENCES

Hayduk, L. A., & Glaser, D. N. (2000). Jiving' the four-step, waltzing around factor analysis, and other serious fun. *Structural Equation Modeling, 7,* 1–35.

Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI. Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods.* Mooresville, IN: Scientific Software International.

Kendall, M. G., & Stuart, A. (1979). *The advanced theory of statistics* (Vol. 2). New York: Macmillan.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130–149.

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics, 29,* 201–211.

Steiger, J. H. (1989). *EzPATH: A supplementary module for SYSTAT and SYGRAPH.* Evanston, IL: SYSTAT.

Steiger, J. H. (1995). Structural equation modeling (SEPATH). In *Statistica/W* (Version 5). (pp. 3539–3688). Tulsa, OK: StatSoft.

Steiger, J. H. (1999). *STATISTICA power analysis.* Tulsa, OK: StatSoft.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–257). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of factors.* Paper presented at the annual spring meeting of the Psychometric Society, Iowa City, IA.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika, 50,* 253–264.