Psychological Methods
2002, Vol. 7, No. 3, 301–306

1082-989X/02/$5.00 DOI: 10.1037//1082-989X.7.3.301

Comments on the Meehl–Waller (2002) Procedure for Appraisal of Path Analysis Models

Robert C. MacCallum, Michael W. Browne, and Kristopher J. Preacher
The Ohio State University

P. E. Meehl and N. G. Waller (2002) proposed an innovative method for assessing path analysis models wherein they subjected a given model, along with a set of alternatives, to risky tests using selected elements of a sample correlation matrix. Although the authors find much common ground with the perspective underlying the Meehl–Waller approach, they suggest that there are aspects of the proposed procedure that require close examination and further development. These include the selection of only one subset of correlations to estimate parameters when multiple solutions are generally available, the fact that the risky tests may test only a subset of parameters rather than the full model of interest, and the potential for different results to be obtained from analysis of equivalent models.

Meehl and Waller (2002) proposed an innovative procedure for assessing verisimilitude, or truth-likeness, of path analysis models by subjecting them to a creative form of risky test. Model parameters are estimated using only a subset of the elements of the sample correlation matrix, and the resulting parameter estimates are then tested by determining how well they account for the other, unused, elements of the correlation matrix. This procedure is conducted for the original model as well as for a set of similar alternative models, and the original model is then compared with the alternatives with respect to results of the risky tests. Support for verisimilitude of the original model is enhanced to the degree that it outperforms the alternative models.

We find much to commend in the Meehl–Waller (2002) approach, especially the general perspective and principles that form its foundation. We are in full agreement with their fundamental view about the implausibility of exact fit for any model and the utility of the concept of verisimilitude. This view underlies some of our own work on representing and evaluating models (e.g., Browne & Cudeck, 1993; MacCallum, in press; MacCallum & Tucker, 1991). Some of the most eminent contributors to our field have emphasized that we must recognize the inherent imperfection in our models (e.g., Box, 1979; Thurstone, 1930; Tukey, 1961). We also commend Meehl and Waller for their emphasis on the need to subject models to risky tests and to evaluate alternative models. Assessment of the fit of a single model to one set of data does not provide adequate evaluation of the model, and results may be strongly influenced by capitalization on chance. Models must be evaluated with respect to their ability to account for data other than that used to estimate model parameters, and a model of interest must be compared with alternatives rather than be examined in isolation. Otherwise the investigator learns nothing about the cross-validity of a model or about its standing relative to alternative models.

Although we fully agree with these principles that guide the Meehl–Waller (2002) approach, we wish to offer some comments about some aspects of their proposed procedure. We believe that some features of their approach merit close attention and perhaps further development and that an increased awareness of these issues will help the user to make appropriate interpretations of findings.

Robert C. MacCallum, Michael W. Browne, and Kristopher J. Preacher, Department of Psychology, The Ohio State University.

Correspondence concerning this article should be addressed to Robert C. MacCallum. Through September 30, 2002, he will be at the Department of Psychology, The Ohio State University, 1885 Neil Avenue, Columbus, Ohio 43210-1222. E-mail: maccallum.1@osu.edu. After January 1, 2003, he will be at the Department of Psychology, University of North Carolina, Chapel Hill, North Carolina 27599-3270.

Estimation and Risky Tests

For purposes of subsequent discussion, it is useful to provide a brief review of the Meehl–Waller (2002) procedure. Given a model of interest, their procedure involves construction of a set of equations relating the elements of the correlation matrix, **R**, of the measured variables to the model parameters. Let Q represent the system of equations; note that there will be one equation for each distinct element of **R**. For Model C in Figure 1 of the Meehl–Waller article, the set of equations Q is represented by Equations 6–10 in their article. Conventional parameter estimation would involve estimating the parameters of this model (*a, b, c,* and *d*) such that the resulting solution reconstructed the elements of **R** as well as possible, according to some discrepancy function such as maximum likelihood or ordinary least squares. That is, the full set of Equations 6–10 would be used to estimate the parameters. Meehl and Waller correctly recognized that the parameters could be estimated using only a subset of the equations. The set Q of equations can be split into Subsets Q1 and Q2, where equations in Q1 are used to estimate parameters. For instance, in their example, the four parameters can be estimated using only Equations 6, 7, 9, and 10 (Set Q1) and not 8 (Set Q2). Equation 8 could then be used to carry out a type of risky test to determine how well the parameter estimates obtained from the other equations will reconstruct the correlation represented by Equation 8. The closeness of that reconstruction is measured using a root-mean-square residual (*RMSr*) statistic. Equivalently, this procedure amounts to splitting the elements of **R** into two subsets, call them R1 and R2, where R1 represents a minimally sufficient subset to obtain parameter estimates. Resulting parameter estimates are then used to reconstruct the correlation coefficients in R2. For Model C in their Figure 1, the resulting *RMSr* is found to be 0.10, which is obtained by substituting their estimate of parameter *b* into Equation 8 and computing the difference between the reconstructed and observed values of $r_{y_1x_2}$.

Meehl and Waller (2002) then generated a set of alternative models using their delete 1–add 1 (D1-A1) rule and carried out the same procedure for each such alternative. They then compared the *RMSr* values for the original model to the distribution of *RMSr* values for the alternative models and assessed where the original model fell in that distribution. See their Table 2 for an example. The verisimilitude of Model C in their example is supported because it yields smaller *RMSr* values than do the alternatives.

Some Issues of Concern

Obtaining Only a Single RMSr Value for Each Model

It is creative of Meehl and Waller (2002) to have recognized that model parameters can be estimated using only a subset R1 of the elements of **R** and that resulting parameter estimates can be subjected to a risky test using the elements in Subset R2. A key aspect of their approach involves choosing the subset R1. Note that elements for R1 cannot be chosen randomly from **R**, because the corresponding set of equations Q1 must constitute a sufficient set of equations to estimate the parameters. Defining a general rule for designating an appropriate R1 and corresponding Q1 is not a simple problem, and we commend Meehl and Waller for developing a simple rule that accomplishes this feat. According to their rule, for each path present in the model of interest, R1 will include the simple correlation for the pair of variables connected by that path.

However, we emphasize that there generally will be other ways to select R1 for assessment of any given model. Let us return to the Meehl–Waller example using Model C of their Figure 1, and their Equations 6–10. The Meehl–Waller procedure conducts estimation using Equations 6–7–9–10 followed by a risky test using Equation 8. However, there is an obvious alternative. Note that Equations 7 and 8 each involve only the *b* parameter. Thus, it would be possible to estimate using Equations 6–8–9–10 and conduct the risky test using Equation 7. This means that one could obtain parameter estimates using a different splitting of the elements of **R** into R1 and R2 than that used by Meehl and Waller. When this is done, the resulting estimate of *b* is 0.77, rather than the estimate of 0.43 obtained by Meehl and Waller. Completing the process, the corresponding new *RMSr* is found to be 0.34, rather than the value of 0.10 obtained by Meehl and Waller. Both solutions are equally valid. Clearly, different designations of the R1 and R2 subsets may result in very different values of *RMSr* for a given model. For Model C, it happens that only two distinct designations of R1 and R2 are possible, producing the two values of *RMSr* just mentioned, 0.10 and 0.34. However, for other models, there may well be a larger number of valid *RMSr* values, depending on how many ways the minimally sufficient subset R1 can be defined.

This same phenomenon holds as well for each of the alternative models generated by the D1-A1 procedure. For each of the corrupted models D–G in Figure 1 of Meehl and Waller (2002), those authors provided a single *RMSr* value in their Table 2. However, it is in fact possible to obtain multiple values of *RMSr* for each of these corrupted models. We carried out this exercise. For each of Models D–G we specified the set of equations representing the correlation structure, analogous to Equations 6–10 for Model C. Because there are five equations (Set Q) and four parameters for each model, it was necessary to choose four of the five equations (Subset Q1) for estimation and use the remaining equation (Subset Q2) for the risky test. Therefore, for each model, we tested every possible subset of four of the five equations to determine whether those equations were sufficient for estimation of the four parameters. For those acceptable cases, we obtained parameter estimates and then, using the remaining equation, computed the reconstructed correlation and the corresponding *RMSr* value.

This exercise thus involved 30 attempted solutions for *RMSr* (5 attempts for each of the six models) and resulted in a variety of outcomes that are reported in our Table 1. For 6 of these attempts, the system of equations was not identified, thus yielding an infinite number of solutions for the parameter estimates. For the remaining 24 of the attempts, 2 attempts for Model D yielded 2 distinct solutions each for the parameter estimates, along with corresponding *RMSr* values. In addition, 2 attempts for Model G and 2 for Model H yielded complex solutions for the parameter estimates (i.e., solutions including an imaginary component) and corresponding complex *RMSr* values. Also, a total of 10 of the attempts yielded reconstructed correlations outside of the bounds of 1.0 and −1.0. This left a total of 12 solutions that yielded reconstructed correlations that were in bounds, along with corresponding *RMSr* values; this number includes the 6 *RMSr* values presented by Meehl and Waller. Thus, in addition to the 6 valid *RMSr* values of Meehl and Waller (2002), we obtained 6 additional valid values, as seen in our Table 1.

We wish to emphasize several points regarding results of this exercise. First, when carrying out the process of identifying a minimally sufficient subset of equations to obtain parameter estimates and then conducting the Meehl–Waller (2002) risky test using the remaining equations, if one considers all possible such solutions, it appears that quite a variety of solutions can occur. Some of these solutions can be reasonably deemed to be invalid because of reconstructed correlations being out of bounds or other reasons. However, even granting that, it is most important to recognize that completely valid solutions may exist other than those arising via the Meehl–Waller method of selecting which equations or correlations to use for estimation and which to use for the risky tests. The critical point here is that the Meehl–

Table 1
Original and Alternative Root-Mean-Square Residual (RMSr) Values for Example 1 of Meehl and Waller (2002)

	RMSr		
Model	From Meehl and Waller	From other valid solutions to the systems of equations	From invalid solutions to the systems of equations
C	0.10	0.34	
D	0.32	0.47, 0.49,[a] 0.65, 0.80[b]	5.45,[a] 3.08[b]
E	0.55	1.14	0.80, 1.39, 4.97
F	0.36		1.20
G	0.59		1.51, 2.78, $0.78 \pm 0.72i$, $1.28 \pm 1.02i$
H	0.55		4.90, 2.20, $0.02 \pm 1.84i$, $0.43 \pm 0.73i$

Note. Valid solutions: Reconstructed r is ≤ 1 and ≥ -1. Invalid solutions: Reconstructed r is >1 or <-1, or complex. $i = \sqrt{-1}$.

[a] Two distinct solutions to a single set of equations; one yields a valid solution and the other an invalid solution.

[b] Two distinct solutions to a single set of equations; one yields a valid solution and the other an invalid solution.

Waller procedure yields only a single value for *RMSr* for each model, when in fact multiple valid solutions may well exist, possibly in large numbers for more complex models. Thus, there exists a source of variability in results that is not taken into account in the Meehl–Waller procedure with respect to specification of the R1 and R2 subsets of correlation coefficients. As a result, the evaluation of the standing of a given model relative to a set of alternatives may be affected. For instance, Meehl and Waller evaluated Model C as surviving its risky test better than did Model D because their *RMSr* for C is smaller than that for D (0.10 v. 0.32). However, another valid solution for *RMSr* for C is 0.34, which changes the perspective of the performance of C as compared with D. Of course, other values for D exist as well.

An interesting feature of the results shown in our Table 1 is that, of the alternative values of *RMSr* obtained for each model, the value produced by the Meehl–Waller (2002) approach is always the smallest. It may be the case that the Meehl–Waller procedure for choosing R1 and R2 will necessarily produce the smallest *RMSr,* as compared with other choices of R1 and R2. This would be a difficult, if not impossible, property to prove, but we have not seen counterexamples. If it is true, it is not clear whether this would be a desirable property in that it would in some sense bias the result for each model by yielding only the best measure of performance rather than a representative value. That is, the Meehl–Waller procedure may be evaluating models by comparing the smallest possible *RMSr* for each model and ignoring the rest of the distributions of *RMSr* values. Regardless, we believe that the most desirable approach would be to obtain all valid solutions rather than just one.

The issue just discussed leaves the user of the Meehl–Waller (2002) procedure with a problem. A full resolution of this problem would apparently require an exhaustive treatment of the system of equations, Q, representing each model so as to seek all possible permissible designations of Q1 and Q2 and to obtain parameter estimates and *RMSr* for each such splitting. Such an exhaustive search probably becomes impractical when models and numbers of parameters become large. The number of systems of equations to solve would become extremely large, and the process of obtaining those solutions would likely yield a mix of valid and invalid solutions as seen in the example just presented. At this point, we simply see this issue as an unresolved problem, but we are not comfortable with a procedure that bases model evaluation on only a single *RMSr* value when an unknown number of alternative valid values exist.

Model Comparison Using RMSr Values

Meehl and Waller (2002) proposed comparing *RMSr* values produced for different models as a basis for assessing verisimilitude in terms of relative capacity of each model to survive a risky test. A close examination of the nature of the test to which each model is subjected reveals that users must be extremely cautious about interpretation of these model comparisons. Consider the first example presented by Meehl and Waller. Model C in their Figure 1 is represented by the correlation structure given in their Equations 6–10. For this model, the *RMSr* value of 0.10 produced by Meehl and Waller is obtained by using Equations 6–7–9–10 to estimate parameters and Equation 8 to conduct the risky test. Note that the only parameter in Equation 8 is *b.* Thus, the outcome of this risky test depends only on the estimate of *b* and not at all on the estimates of the other parameters. Likewise, the alternative *RMSr* value of 0.34 for Model C in our Table 1 uses Equation 7 to conduct the risky test and again depends only on the estimate of *b.* Therefore, the risky tests for Model C do not reflect on performance of the full model but rather just reflect the capacity of the estimate of parameter *b* to account for information other than that used to make the estimate. The estimates of the other three parameters are irrelevant. Furthermore, because of the structure of the equations representing Model C, it is impossible to conduct any other risky tests of parameters other than *b.*

In a similar manner, we can determine that risky tests of other models in the set of alternative models may depend on only subsets of the model parameters. For instance, it can be shown that the *RMSr* values for Model F in Figure 1 of Meehl and Waller (2002) depend only on the estimate of parameter *w* in that model. Again, because of the structure of this model, it is impossible to conduct risky tests involving any of the other parameters. Thus, a comparison of *RMSr* for Models C and F does not reflect anything about the full models but rather reflects only the results of risky tests of different parameters of these two models. Meehl and Waller's finding that Model C produces a smaller *RMSr* than Model F does not reflect better verisimilitude of the entire Model C versus the entire Model F but rather indicates only that the estimate of parameter *b* in Model C survives a risky test better than does the estimate of parameter *w* in Model F. For some models, such as Model D, it can be shown that the *RMSr* values depend on estimates of all four of the model parameters. Regardless, it is difficult to justify use of these *RMSr* values for model comparison when such values for different models often depend on different subsets of model parameters.

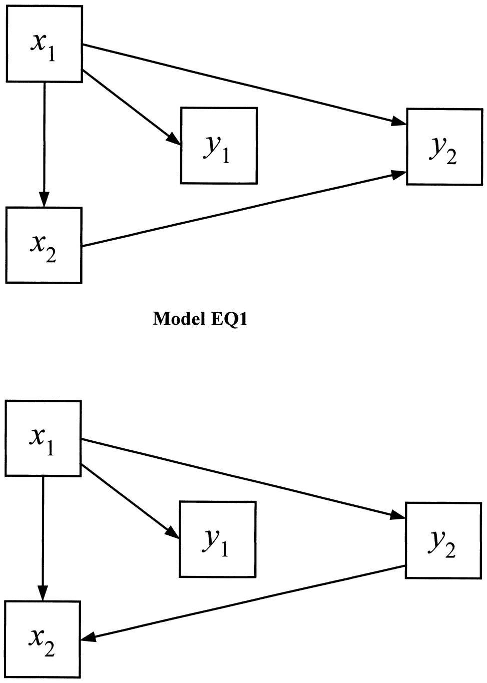
The same phenomenon applies to the multiple *RMSr* values that can be obtained for a single model as presented in our Table 1. For any single model, it can be shown that the various *RMSr* values may not represent outcomes of risky tests of the same parameter(s). For instance, for Model E, the Meehl–Waller (2002) *RMSr* value of 0.55 represents the result of a test involving all four parameters. Other *RMSr* values for Model E in our Table 1 represent results of tests involving only two parameters each. The values 1.14 and 0.80 arise from risky tests using equations that involve only *b* and *w,* and the values 1.39 and 4.97 result from equations involving only *c* and *d.*

These observations must call into question the Meehl–Waller (2002) approach of assessing the standing of the original model relative to the set of alternatives by determining the rank of the *RMSr* for the original model in the distribution of *RMSr* values representing all of the alternatives. In the sense just explained, these various *RMSr* values do not all reflect outcomes of risky tests of entire models, but rather different parts of the models. Interpretation of the percentile rank of a given model, as presented by Meehl and Waller, should not be taken as reflecting an assessment of relative verisimilitude of the models being considered.

Equivalent Models

It is interesting to consider the outcome of the Meehl–Waller (2002) procedure when applied to equivalent initial models. For present purposes, let us define equivalent models as models that are parameterized differently but that fit any given data equally well. Some rules for generating equivalent models were described by Lee and Hershberger (1990), and the routine occurrence of equivalent models in substantive research was examined by MacCallum, Wegener, Uchino, and Fabrigar (1993). Ignoring for the moment the issues discussed above, suppose that the Meehl–Waller procedure were applied, in exactly the manner they proposed, to two different initial models that were equivalent in the sense just defined. The *RMSr* values for those two models would be the same, we expect. However, differences are likely to arise with respect to the nature of the sets of alternative models generated by the D1-A1 procedure. This could occur in at least two ways. First, if the two equivalent models differ with respect to which variables are exogenous, then the resulting sets of alternative models may differ in number. This would occur because the Meehl–Waller approach treats correlational paths between exogenous variables as inviolate, whereas directional paths involving endogenous variables can be deleted during the D1-A1 procedure. Thus, if equivalent Models M1 and M2 differed in that M1 contained correlations between a number of exogenous variables whereas M2 replaced some of those correlations with directional paths, then the set of alternative models generated by the D1-A1 procedure applied to M1 may be smaller than the set generated for M2. Second, if the two equivalent models differ by reversal of a path between endogenous variables, application of the D1-A1 procedure to these models may produce sets of alternative models that are not equivalent to each other. If Models M3 and M4 differ in this way, then the models in the alternative set produced for M3 will not be equivalent to those produced for M4.

We offer a small-scale illustration of this last point by modifying the first example presented by Meehl and Waller (2002). Ignoring their models, but using



Model EQ1

Model EQ2

Figure 1. Two equivalent models fit to the correlation matrix in Table 1 of Meehl and Waller (2002).

the correlation matrix in their Table 1, we constructed two equivalent models, designated Models EQ1 and EQ2 and shown in our Figure 1. Note that these models differ with respect to the reversal of a single path. Application of the Meehl–Waller procedure using each of these models in turn as the original model yields the same *RMSr* value of 0.44 for EQ1 and EQ2, along with results for 12 alternative models in each case. However, the two distributions of *RMSr* values for the alternative models are not the same, because the 12 modified models in each set are not pairwise equivalent. As a result, the percentile ranks of EQ1 and EQ2 are found to be different (0.33 vs. 0.50), implying better relative verisimilitude of EQ1 even though the two models are equivalent. We have carried out other small-scale demonstrations showing differences in the set of alternative models and corresponding distributions of *RMSr* values given equivalent initial models. These findings suggest that this phenomenon may be a significant matter for further study. They also raise at least some concern in that the Meehl–Waller procedure as proposed can indicate differential performance for equivalent initial models, simply as a function of the sets of alternatives generated by their D1-A1 procedure.

Summary

As noted at the outset, we find much common ground with the perspective underlying the Meehl–Waller (2002) approach. However, we are not completely comfortable with a number of aspects of their proposed procedure. These include the provision of only a single *RMSr* value for each model when multiple such values generally exist, the fact that *RMSr* values for different models may reflect the outcome of risky tests for different parts of the models rather than the full models, and the potential for different results for equivalent models simply due to differences in the sets of alternative models generated for comparison. Although we commend Meehl and Waller for their sound guiding principles and we admire the creativity inherent in their proposed procedure, we do believe that the matters we have discussed raise some cause for concern about methods and interpretation of results. Users of the procedure should be aware of these issues, and those of us who study and develop methods could perhaps contribute to refining the innovative approach proposed by Meehl and Waller.

References

Box, G. E. P. (1979). Some problems of statistics and everyday life. *Journal of the American Statistical Association, 74,* 1–4.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Lee, S., & Hershberger, S. (1990). A simple rule for generating equivalent models in covariance structure modeling. *Multivariate Behavioral Research, 25,* 313–334.

MacCallum, R. C. (in press). Working with imperfect models. *Multivariate Behavioral Research.*

MacCallum, R. C., & Tucker, L. R (1991). Representing sources of error in the common factor model: Implications for theory and practice. *Psychological Bulletin, 109,* 502–511.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin, 114,* 185–199.

Meehl, P. E., & Waller, N. G. (2002). The path analysis controversy: A new statistical approach to strong appraisal of verisimilitude. *Psychological Methods, 7,* 283–300.

Thurstone, L. L. (1930). The learning function. *Journal of General Psychology, 3,* 469–478.

Tukey, J. W. (1961). Discussion, emphasizing the connection between analysis of variance and spectrum analysis. *Technometrics, 3,* 191–219.

Received February 21, 2002
Revision received March 29, 2002
Accepted March 29, 2002 ■