Psychological Methods
2002, Vol. 7, No. 3, 323–337

1082-989X/02/$5.00 DOI: 10.1037//1082-989X.7.3.323

Risky Tests, Verisimilitude, and Path Analysis

Niels G. Waller
Vanderbilt University

Paul E. Meehl
University of Minnesota

P. E. Meehl and N. G. Waller (2002) described a novel approach for appraising the verisimilitude of path analysis models. This approach uses limited information parameter estimates when assessing model fit and a nonparametric badness-of-fit measure that relies on path diagram combinatories. R. C. MacCallum, M. W. Browne, and K. J. Preacher (2002); C. S. Reichardt (2002); and S. Mulaik (2002) have provided comments on this work. In this article the authors respond to the commentators and reemphasize the importance of subjecting causal models to severe (risky) tests.

Meehl and Waller (2002) outlined a new approach to strong appraisal of verisimilitude in recursive path analysis models. Several of our methodological prescriptions are seemingly controversial. For example, we advocated the use of limited information parameter estimates when assessing model fit. We also eschewed parametric goodness-of-fit tests in favor of nonparametric tests that are based on path diagram combinatorics. Not surprisingly, these and other features of our approach have generated questions and concerns. Before addressing these concerns, we want to thank the editor, the reviewers, and the commentators for participating in a stimulating exchange that has sharpened our thinking on the many issues that are covered in these pages.

Comments on MacCallum, Browne, and Preacher's "Comments on the Meehl–Waller (2002) Procedure for Appraisal of Path Analysis Models"

MacCallum, Browne, and Preacher (2002) suggested that three aspects of our work merit closer examination and further development: (a) the precision of VERIPATH parameter estimates, (b) the fact that risky tests may often be testing only a subset of model parameters rather than the full model of interest, and (c) the potential for different results to be obtained from analysis of equivalent models due simply to differences in the comparison sets of corrupted models. We address each of these concerns in turn.

The Precision of VERIPATH Parameter Estimates

Fifty years before the phrase *path analysis* was introduced into the vernacular of the social sciences, Sewall Wright (1921) demonstrated that (recursive) linear causal models can be represented as a system of inhomogeneous simultaneous equations. When the model is just identified and correctly specified, the model parameters—that is, the path coefficients—can be estimated through algebraic manipulation of the equations. When the model is overidentified, however, the researcher is presented with an embarrassment of riches. By choosing different subsets of equations, multiple estimates of the same parameter are sometimes derivable. In later publications, Wright (1960) resolved this dilemma by using multiple regression to estimate (unique) path coefficients.

In our original article we showed that standardized path coefficients could often be estimated with fewer than the total number of available, nonredundant correlations. We also noted that these parameter estimates might not be unique. Several commentators were bothered by this point. For instance, MacCallum et al. (2002) noted that "one could obtain parameter estimates using a different splitting of the elements of **R** into R1 and R2" (p. 302). They also suggested that when multiple solutions exist, all solutions are "equally valid" (p. 302). Rather than debate the meaning of *equally valid,* let us simply agree that the properties of VERIPATH parameter estimates merit fur-

Correspondence concerning this article should be addressed to Niels G. Waller, Department of Psychology and Human Development, Peabody College, Vanderbilt University, Box 512, Nashville, Tennessee 37203, or to Paul E. Meehl, Department of Psychology, University of Minnesota, Elliott Hall, 75 East River Road, Minneapolis, Minnesota 55455-0344. E-mail: niels.waller@vanderbilt.edu or pemeehl@umn.edu

ther consideration. One property that warrants closer scrutiny is the similarity of our estimates to so-called full information, maximum-likelihood (ML) estimates. In our original article we noted that these estimates are oftentimes very similar.

To examine this property more closely, we analyzed several models using VERIPATH and LISREL 8.5 (Jöreskog & Sörbom, 2001). For the sake of brevity, we focus on the parameter estimates and avoid discussing the substantive implications of these models. Accordingly, in the following paragraphs we designate all exogenous (independent) and endogenous (dependent) variables as simply x_i and y_i variables, respectively.

Earlier, we suggested that VERIPATH coefficients are "very similar to more efficient estimates [e.g., least squares] . . . even though the VERIPATH estimates are often calculated using 30% to 50% fewer pieces of information (i.e., correlations)" (Meehl & Waller, 2002, p. 394). We now provide further support for that claim.

Figure 1 displays a path model that was originally studied by Alwin (1988). This is a fully recursive model with three exogenous (x) and four endogenous (y) variables. Using data from Alwin's publication (specifically, for the 25 to 34-year-olds), we calculated ML estimates of the parameters (reported in Figure 1) and then used the estimates to generate a model-implied correlation matrix.

The reproduced correlations are entailed by the path model and coefficients in Figure 1. Thus, if we analyze the correlations with VERIPATH, the resulting parameter estimates should match the data generating parameters perfectly. If they do not, then our estimation method is seriously flawed. After all, the model–data fit is exact in this contrived example, and thus any set of correlations in R1 that yields a proper solution (e.g., nonimaginary numbers, predicted correlations that lie between −1 and +1) should also yield the true solution (i.e., the data generating parameters). That was exactly what we found. ML procedures would have yielded identical results.

Of course, the previous finding was to be expected because it follows directly from the logic of covariance algebra. Nevertheless, it was important to establish because it provides a baseline for comparing VERIPATH and ML estimates in more realistic examples in which model–data fit is not exact. Two sources of model–data misfit are sampling error and model misspecification. Both sources are considered in the following examples.

Sampling error was studied using the correlation matrix of the previous example. That matrix was considered a population matrix. Performing a Cholesky decomposition of the parent matrix and then applying the resulting weights to vectors of random normal deviates simulated sample data matrices from this population. Correlation matrices were generated for

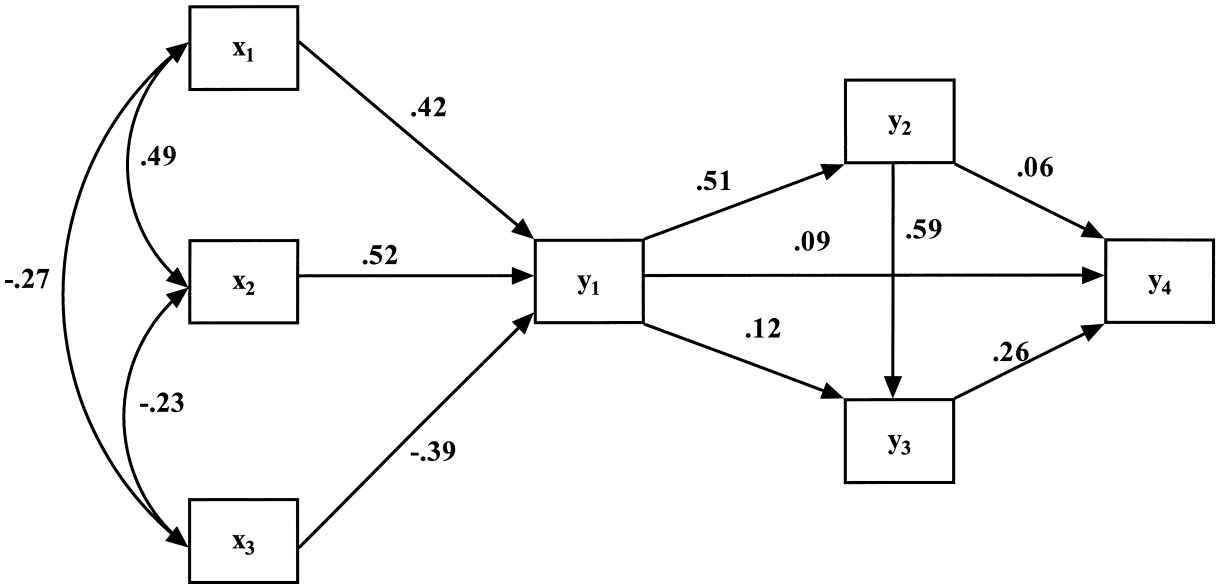


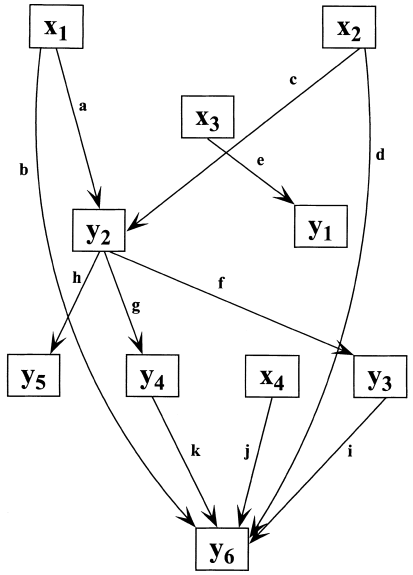
Figure 1. Model proposed by Alwin (1988) with maximum-likelihood parameter estimates added. The model includes three exogenous (x) and four endogenous (y) variables.

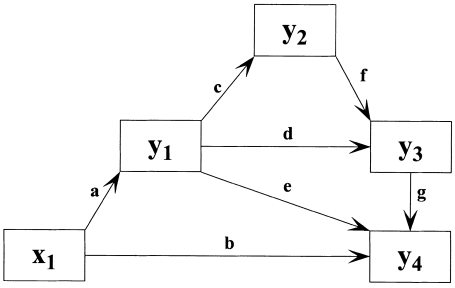
sample sizes of 200 to 600. Each matrix was analyzed with VERIPATH and LISREL 8.5 using the model displayed in Figure 1.

The findings from these analyses are instructive. For all sample sizes—representing different degrees of sampling error—the VERIPATH and ML parameter estimates were identical even though the VERIPATH estimates were based on only 12 of the 21 available nonredundant correlations. One could argue that these findings are noteworthy but of limited value to applied researchers who are unlikely to encounter "true" causal models at any time during their research careers. We would agree with that contention. Although these views may sound pessimistic, we believe that they are accurate.

We are not alone in holding a skeptical view of the veracity of causal models (see, e.g., Browne & Cudeck, 1993; MacCallum & Tucker, 1991). For instance, Cudeck (1991) proclaimed, "A 'correctly specified model' is, always has been, and always will be a fiction. . . . All that can be hoped is that a model captures some reasonable approximation to the truth" (p. 261); Steiger (2000) added that "perfect models. . . . occur with probability very close to zero in practice" (p. 159); Jöreskog (1993) admitted that "the use of chi-square as a central χ^2-statistic is based on the assumption that the model holds exactly in the population . . . this may be an unreasonable assumption in most empirical research" (p. 309); and McDonald and Ho (2002) opined that "it has long been recognized that all SEMs [structural equation models] are simplified approximations to reality, not hypotheses that might possibly be true" (p. 71). These and similar comments point to the importance of investigating the performance of VERIPATH in models that are literally false—in other words, in misspecified models.

We studied the influence of model misspecification on VERIPATH parameter estimates using two models from recent publications. Admittedly, both models are based on real data where the latent structure is unknown; thus either model could be correctly specified. Nevertheless, for previously stated reasons, we feel confident that both models have some, and perhaps a large, degree of model misspecification. Both models, displayed in Figure 2, are contemporary examples of path analysis. Model 1 (Ho, Davidson, Van Dyke, & Agar-Wilson, 2000) recently appeared in the *Journal*





MODEL 1 **MODEL 2**

Figure 2. Two (probably misspecified) models: Model 1 from Ho, Davidson, Van Dyke, and Agar-Wilson (2000) and Model 2 from Chan, Schmitt, DeShon, Clause, and Delbridge (1997).

of Health Psychology, whereas Model 2 (Chan, Schmitt, DeShon, Clause, & Delbridge, 1997) was published in the *Journal of Applied Psychology.*

The two models were analyzed with VERIPATH and LISREL 8.5 using data from the original articles. For comparative purposes, we calculated ML, instrumental variables (IV), and two-stage least squares (TS) estimates of the path coefficients. Like VERIPATH, the latter two methods provide limited information estimates. All estimates are listed in Table 1.

A striking feature of Table 1 is the similarity of the VERIPATH and ML parameter estimates. Although the correlation matrix for Model 1 contains 45 distinct elements, VERIPATH uses only 17 correlations (38%) to estimate the parameters. In Model 2, 7 of 10 correlations are used during parameter estimation. Nevertheless, for both models, the VERIPATH and ML estimates are remarkably close: 12 of the 18 path coefficients are identical, 1 differs by no more than |.02|, and none differ by more than |.06|.

We have focused on the quality of VERIPATH parameter estimates to assuage concerns that they are somehow arbitrary. VERIPATH chooses one solution among many, but in our experience it always chooses a highly accurate solution (if accuracy is gauged by closeness to full information solutions). We realize that this line of reasoning—that is, generalizing from past successes to probable future success—is fraught with logical difficulties. We are able to prove that if an endogenous variable is influenced only by exogenous variables, then all paths leading to the endogenous variable will have identical VERIPATH and ML parameter estimates. We have not been able to derive the sampling distributions for paths linking two endogenous variables. Users of VERIPATH should remember that if the precision of VERIPATH estimates is ever in doubt, one can always compare them to full information estimates (because the true model is unknown, we cannot determine their absolute accuracy).

Table 1
Full and Limited Information Parameter Estimates for Path Models 1 and 2

Parameter	ML	VP	IV	TS
		Model 1		
a	.30	.30	.36	.30
b	−.36	−.31	−.36	−.36
c	.39	.39	.44	.39
d	−.27	−.21	−.27	−.27
e	.66	.66	.67	.66
f	.33	.33	.03	.04
g	.41	.41	.34	.19
h	.34	.34	.19	.22
i	−.20	−.26	−.20	−.20
j	.21	.16	.21	.21
k	−.35	−.31	−.35	−.35
		Model 2		
a	−.41	−.41	−.41	−.41
b	−.08	−.09	−.08	−.08
c	.29	.29	.34	.34
d	.26	.26	.26	.26
e	.77	.77	.77	.77
f	.37	.37	.37	.37
g	.10	.10	.10	.10

Note. ML = maximum likelihood; VP = VERIPATH; IV = instrumental variables; TS = two-stage least squares.

Our preferred badness-of-fit measure, the root-mean-square residual (*RMSr*), is a function of the observed and expected correlations (but only those correlations not involved in parameter estimation). Obviously, if VERIPATH parameter estimates are highly biased, then our fit index will also be biased. MacCallum et al. (2002) noted that if multiple solution sets are possible, then multiple *RMSr* values are also possible. Furthermore, they wondered why any particular *RMSr* is more valid than the next. We have already shown that VERIPATH parameter estimates are typically very close to full information estimates. In Figure 3 we offer additional support for the sensitivity of the *RMSr* as a measure of model verisimilitude.

A useful feature of VERIPATH is its ability to create a stacked LISREL file for all alternative models that are generated by the delete 1–add 1 rule. This feature makes it easy to compare VERIPATH output with analogous information from LISREL. Figure 3 displays data from the 199 alternative models that were generated in our reanalysis of data on custodial fathers (Rettig, Leichtentritt, & Stanton, 1999). Each data point represents a pairing of one *RMSr* value from VERIPATH with the corresponding root-mean-square error of the approximation (RMSEA) from LISREL. A glance at the figure indicates that the two indices are clearly measuring something similar. They correlate .93, indicating that both measures provide similar rankings of model fit. We view this as an important finding because the two measures are founded on highly different theoretical assumptions. The RMSEA is based on an ML noncentral chi-square/sampling-theory perspective, whereas VERIPATH uses admittedly inefficient parameter estimates and makes only weak assumptions about the joint density

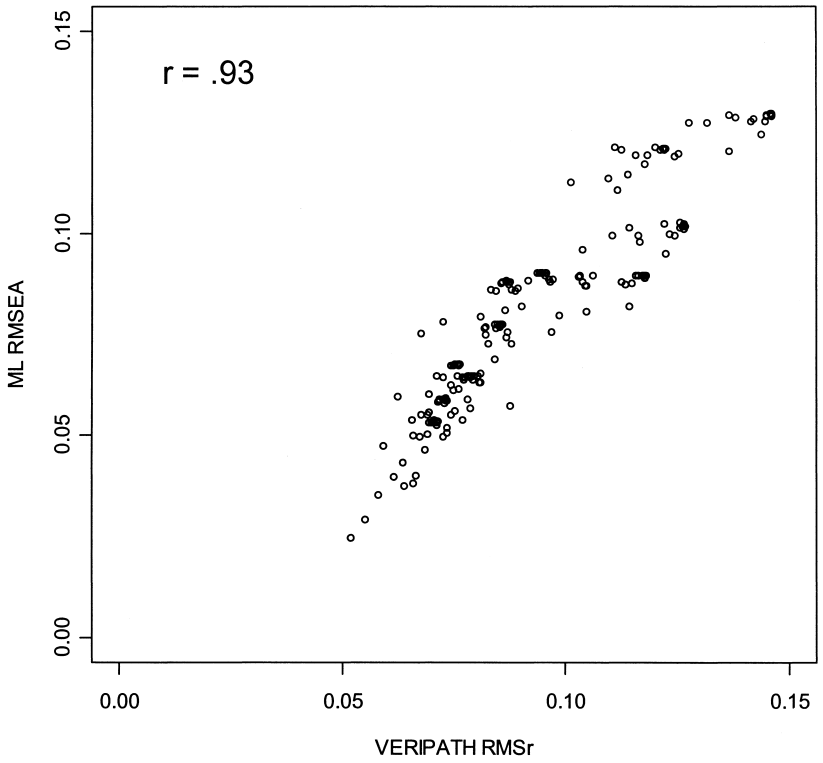


Figure 3. Correlation between maximum-likelihood root-mean-square error of the approximation (ML RMSEA) and root-mean-square residual (*RMSr*) for 199 models generated for data from Rettig, Leichtenritt, and Stanton (1999).

of the observed variables. Rather than consider the RMSEA as a gold standard, we view this finding as providing mutual support for the validity of both fit measures.

Risky Tests May Often Be Testing Only a Subset of Model Parameters

A second concern of MacCallum et al. (2002) is that our proposed risky tests may be based on only a subset of model parameters. For instance, the *RMSr* of Example 1 (Meehl & Waller, 2002) was based solely on equations involving coefficient *b*. On this point our critics are correct. Example 1 was chosen to be as simple as possible so that readers could follow the mathematics of our method. This illustrative model contains a single degree of freedom, and only five alternative models are generated by the delete 1–add 1 rule. In very simple models with few alternatives, the distribution of *RMSr* values has too few numbers to generate a stable distribution. This is a situation that merits caution when interpreting a combinatoric fit index. Obviously, if the delete 1–add 1 rule generates 15 alternative models for one example and 150 alternative models for a second example, then a D* (theoretically preferred model) ranking of 85% offers different probative weight for the two cases. Numbers are not self-interpreting, and thus we must always evaluate a fit index within the context and conditions in which it was generated.

In larger models, or models with more degrees of freedom, the delete 1–add 1 rule typically generates hundreds of alternative models. Moreover, as we have seen in earlier examples, VERIPATH often uses a small fraction of the available data to estimate the model parameters. This leaves a sizable portion of the data for model testing purposes. We illustrate this point with the data in Table 2.

Table 2 lists the correlations that were used to fit Model 1 (Ho et al., 2000) in Figure 1. Notice that 28 correlations were held in abeyance to assess the model's verisimilitude. Although not shown in the table, it is noteworthy that all of the parameters were used to calculate the reproduced correlations. Moreover, most parameters were used numerous times. For instance, path coefficients *a* and *c* (reported in Table 1) were

Table 2
Correlations and Discrepancies for Model 1 in Figure 1

Variable	1	2	3	4	5	6	7	8	9	10
1	—	.296	.055	.136	.073	−.258	.085	−.015	.656	−.152
2	.271	—	.332	.406	.344	−.420	.062	.198	.145	−.062
3	.047	***r*3,2**	—	.306	.298	−.323	.083	−.071	.068	−.015
4	.126	***r*4,2**	.171	—	.726	−.371	−.094	.105	.126	.162
5	.064	***r*5,2**	.184	.586	—	−.271	.014	−.041	.122	.217
6	−.218	−.144	***r*6,3**	***r*6,4**	−.176	—	−.218	−.046	−.207	.200
7	.027	***r*7,2**	.062	−.119	−.007	***r*7,6**	—	−.623	.088	−.202
8	−.035	***r*8,2**	−.137	.025	−.109	***r*8,6**	***r*8,7**	—	.031	.116
9	***r*9,1**	.106	.055	.110	.109	−.146	***r*9,7**	***r*9,8**	—	−.124
10	−.071	−.046	−.010	.169	.223	***r*10,6**	***r*10,7**	***r*10,8**	***r*10,9**	—

Note. Observed correlations are reported above the main diagonal. Model discrepancies (residuals) are reported below the main diagonal. Boldface placeholders denote correlations used to estimate model parameters.

included in 25 of the 28 equations that generated the reproduced correlations. No parameter was used in less than 4 equations. This type of information is easily derived from VERIPATH, and thus researchers can determine the extent to which path coefficients are used in the derivation of *RMSr* values. VERIPATH provides a riskier test when all or most path coefficients are used to reproduce the correlations in R2.

Some readers may fear that the correlations in R2 are highly constrained by the values in R1. If that were true, then the entire rationale of our method would be in jeopardy. This fear can be more clearly expressed as follows.

1. A correlation matrix **R** is partitioned into two nonoverlapping sets of correlations: R1 and R2.

2. The elements in R1 are used to estimate path coefficients. These coefficients perfectly reproduce the correlations in R1 because the coefficients were calculated from a just-identified series of equations.

3. The path coefficients are then used to estimate the values in R2.

4. The correlations in R2 must be reproduced accurately (although not perfectly) because the elements in R1 and R2 are highly dependent, being selected from the same correlation matrix.

Fortunately, the reasoning behind this concern is virtually never sound for the types of correlation matrices that are observed in the social sciences. We can explore this issue in greater depth by returning to Example 1 in Meehl and Waller (2002). We also describe why analytic support for this view is difficult to come by except in unusual situations.

Recall that Example 1 was a simple model with only four variables. The topological structure of the model was portrayed in Figure 1C of Meehl and Waller (2002). We argued that r_{y_1,x_2} could be used for model appraisal because it was not used to estimate the model parameters. The observed and predicted correlations for r_{y_1,x_2} were .23 and .13, respectively. In our original discussion of this example, we did not show that the reproduced correlation could have been much farther from .23 without violating the essential properties of the correlation matrix. We do so now with the help of symbolic algebra. More specifically, we can determine the range of possible values for r_{y_1,x_2} by solving a quadratic equation that involves the matrix determinant. VERIPATH automatically solves this equation for us and shows that the correlation matrix will have a positive determinant whenever $.109 \leq r_{y_1,x_2} \leq .989$. In plain English, r_{y_1,x_2} was not highly constrained by the elements in R1. Stated otherwise, there was plenty of room for the observed and reproduced correlation to differ.

Unfortunately, the aforementioned method does not work when R2 contains more than a single correlation. What we need under this condition is a means of determining simultaneous bounds for all correlations in R2 given the values in R1. Although the literature contains some relevant work on this problem (Glass & Collins, 1970; Hubert, 1972; Olkin, 1981), analytic solutions are apparently only available when R1 and R2 are proper submatrices of **R** (the literature considers range restrictions in submatrices of partitioned matrices only). This will almost never be the case with VERIPATH. Nevertheless, we can answer our

question using simulation methods. A possible simulation could be designed as follows: (a) Partition **R** into R1 and R2; (b) replace the entries in R2 by random numbers between −1.00 and +1.00, and call this matrix **K**; (c) compute the determinant of **K**, and if the determinant is greater than 0.00, then **K** satisfies the conditions of a proper correlation matrix; (d) compute an *RMSr* for **K** by using the path coefficients derived from R1; and (e) save the *RMSr* and repeat. Following this design, we simulated 1 million plausible *RMSr* values for the first model in Figure 1 of the current article. The largest *RMSr* was .80. Clearly, in this example, the correlations in R2 were not highly constrained by those in R1.

Equivalent Models

The final issue raised by MacCallum et al. (2002) concerns the behavior of the delete 1–add 1 rule when faced with equivalent models. The authors defined *equivalent models* as "models that are parameterized differently but that fit any given data equally well" (p. 305). This definition is problematic because it entails the logical impossibility of discovering a fit index that can distinguish equivalent models. If the models are distinguishable, they cannot be equivalent. We suspect that the authors meant something different, namely, that equivalent models are models that are covariance equivalent (models are covariance equivalent when they imply equivalent reproduced covariance matrices; see Pearl, 2000, p. 145). This definition leaves open the possibility of finding a discriminating index that uses information in addition to the model residuals. We are not claiming to have found such an index. What we are claiming is that our verisimilitude index is a potential candidate because of the manner in which it is calculated (i.e., by considering the class of alternative models generated by the delete 1–add 1 rule). At this point we consider the question unresolved and hope that future simulation studies will shed further light on the issue. However, we may get an inkling of what those simulations will reveal by considering the equivalent models in MacCallum et al.

MacCallum et al. (2002) noted that EQ1 and EQ2 (a) are equivalent (according to our interpretation of their meaning), (b) produce identical *RMSr* values (.44), and (c) have different sets of alternative models as generated by the delete 1–add 1 rule. Parenthetically, we note that the correct number of alternative models is 17 (not 12 as reported) and the *RMSr* for EQ2 is .53 (not .50).

To clarify any confusion concerning the application of the delete 1–add 1 rule, we consider the alternative models for EQ1 and EQ2 in some detail. Path diagrams for these models are portrayed in Figures 4 and 5. The *RMSr* for each model is displayed above the diagrams. A comparison of these figures quickly supports the observation that the alternative models are not pairwise equivalent. MacCallum et al. (2002) found this observation troubling. We were neither surprised nor discouraged by this result. To understand our position requires that we further explicate the role of the delete 1–add 1 rule in generating alternative causal structures.

We have argued, as have many others, that it is useful to assess a model's performance by comparing it to relevant alternative models. Practicing what one preaches in this case, however, presents an obvious difficulty. Namely, there are a vast number of alternative models for any path diagram (e.g., models could include interaction terms, nonlinear relations, or various parameter constraints). In earlier examples we demonstrated that, even when degrees of freedom are held constant, one could often generate hundreds of alternative models. The delete 1–add 1 rule was designed to limit this plethora of riches and to focus consideration on a manageable set of structurally close models. Models EQ1 and EQ2 are covariance equivalent, but according to our definition they are not structurally close. This statement requires further elaboration.

Notice in Model EQ1 that y_2 is an endogenous variable that is influenced by multiple conjectured causes (x_1 and x_2). Accordingly, the endogenous status of y_2 is an important component of the model. Moreover, it is a conjecture that should not be modified in the alternative models; otherwise they would not be structurally close. The delete 1–add 1 rule respects the endogenous status of y_2. Failing to keep this feature of the model would have produced a class of models with a significantly different causal structure than that entailed by EQ1. In EQ2, on the other hand, the endogenous status of y_2 is more tenuous. Deleting a single path from the model can turn y_2 into an exogenous variable. This actually occurs four times in the models that are displayed in Figure 5.

Before leaving this section, we wish to comment on some additional features of Figures 4 and 5. MacCallum et al. (2002) noted that their equivalent models have *RMSr* values of .44. They neglected to mention that these values indicate that neither model does an adequate job of reproducing the observed correla-

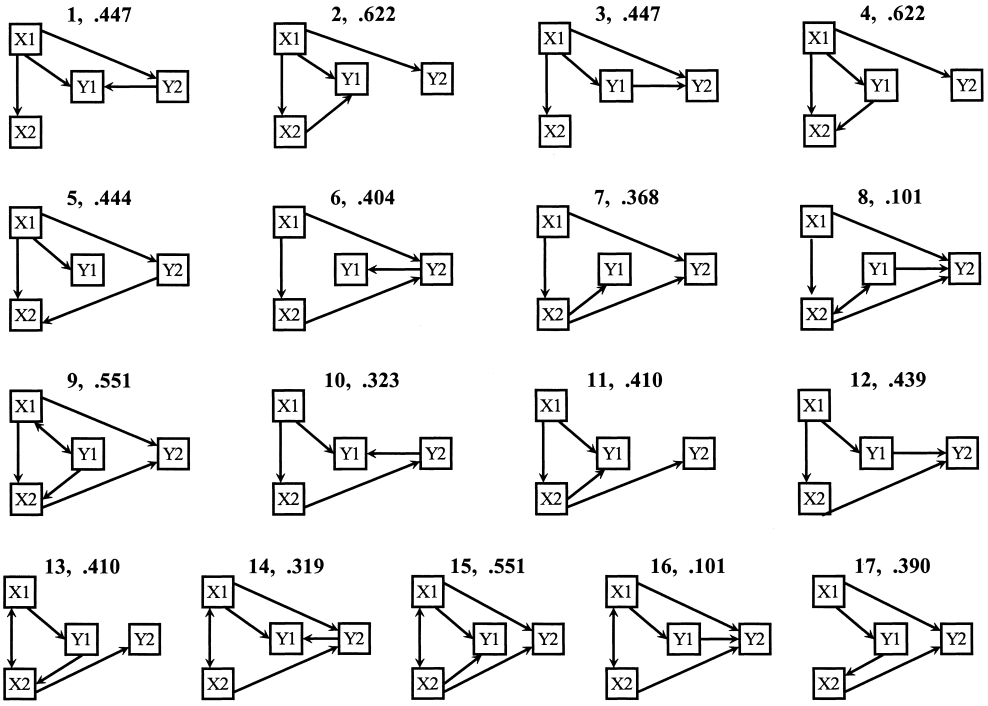


Figure 4. Alternative models for EQ1.

tions. Further evidence that the models are badly misspecified can be gleaned from the alternative path diagrams. Notice that many *RMSr* values in these diagrams are also close to .44. A strength of VERIPATH is that it generates a distribution of fit indices rather than a single measure of global fit. Users may be tempted to cast a line in this distribution and fish for a well-fitting model. Do not do it! Such behavior would be counter to the spirit of VERIPATH. We designed VERIPATH so that researchers could submit theoretically preferred models to risky tests. VERIPATH is not a model exploration technique. Lastly, we feel compelled to mention that in terms of absolute and relative fit, our original model fares much better than either EQ1 or EQ2. Readers should remember that although all three models are covariance equivalent, only Model C is structurally close to the model that actually generated the data.

Comments on Reichardt's "The Priority of Just-Identified, Recursive Models"

Reichardt (2002) crafted a commentary that is rich in ideas and stimulating conjectures. He surveyed our work from a shared methodological vantage point and he offered fresh proposals for model assessment. Although we comment briefly on several features of his alternative methods, this forum is not the appropriate outlet for a thorough exegesis of these fascinating ideas. Rather, in the following paragraphs we limit our comments to those aspects of Reichardt's remarks that bear directly on the novel aspects of VERIPATH.

Norm- Versus Criterion-Referenced Model Appraisal

We are concerned that Reichardt (2002) painted a narrow portrait of our ideas. Although our article contains many prescriptions for model assessment that depart from standard practice in the path analysis literature, Reichardt emphasized only one of them: the delete 1–add 1 rule. Other elements of our proposal are given short shrift. For instance, Reichardt stated, "My purpose is not to oppose Meehl and Waller's delete one–add one (D1-A1) method but to juxtapose it with an alternative" (p. 307). He claimed, "Meehl and Waller (2002) proposed an innovative method for theory appraisal that. . . . is only norm-referenced" (p. 307); "the D1-A1 method is norm referenced, as opposed to criterion referenced, because it assesses the negligibility of paths by comparing the fits of alternative models" (p. 309); and "the D1-A1 method is solely norm-referenced because it appraises verisimilitude based only on the relative performance of

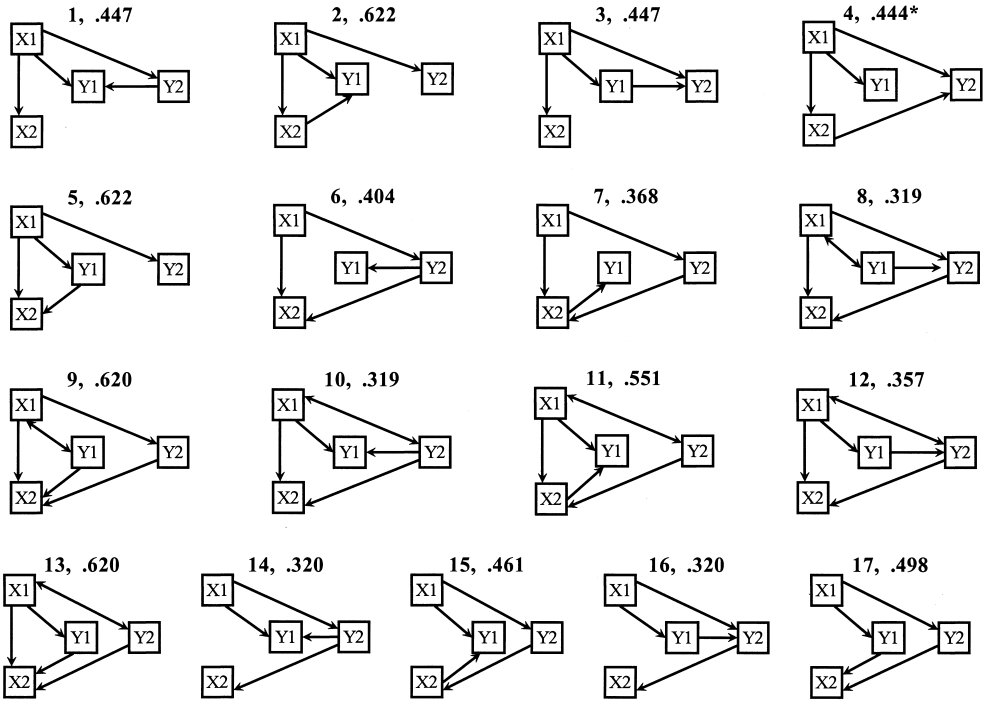


Figure 5. Alternative models for EQ2. The asterisk model is identical to Model C, Figure 1, in Meehl and Waller (2002).

alternative models" (p. 310). Furthermore, while describing his own work, he suggested that researchers may "calculate a goodness-of-fit (GOF) score, which could be Meehl and Waller's (2002) RMSr measure *or any other GOF measure* [italics added]" (p. 309). These and other comments indicate to us that important aspects of our coordinated approach to model assessment have been underappreciated.

Reichardt's (2002) commentary leaves one with the false impression that we emphasized norm-referenced assessment to the exclusion of criterion-referenced assessment. That impression is misleading as testified by comments in our original article. For instance, in our description of Example 1 we noted that "if the *RMSr* was large, we would reject the model no matter how many degrees of freedom were available" (Meehl & Waller, 2002, p. 289). In other words, we clearly encouraged researchers to consider the absolute magnitude of $RMSr_{D^*}$ (the *RMSr* of the theoretically preferred model). If $RMSr_{D^*}$ is large—the game stops. The model fails to account for the data and thus fails to clear the first hurdle of our multihurdle course. If $RMSr_{D^*}$ is smallish, the model is competitive and advances to our norm-referenced test, which is provided by the delete 1–add 1 rule. Recall that we used this rule to generate a combinatorically defined verisimilitude index that indicates whether D* accounts for the data better than a large class of topologically close models.

Notice in this description that we begin model appraisal with a criterion-referenced test before proceeding to a norm-referenced test. In both tests the *RMSr* plays a central role. Unfortunately, the logical justification for this role also seems underappreciated in Reichardt's (2002) commentary. For instance, he suggested that researchers could use "Meehl and Waller's (2002) *RMSr* measure or any other GOF measure" (p. 309). We would not use the *RMSr* in this context because it is always zero or it is based on a single correlation when models are designed according to Reichardt's proposals.

Conjectures and Refutations: The Logical Priority of Falsification

More important, Reichardt (2002) and other critics may have failed to appreciate how the *RMSr* embodies a philosophical stance that favors falsification (refutation) over verification. We are persuaded by Popper's thesis (1959, 1962, 1983) that efforts to falsify theories by subjecting them to risky tests are the most efficient means of gauging a theory's mettle. We

also believe that such efforts should play a larger role in the appraisal of causal models, and our article was written in that spirit. From this perspective, the *RMSr* can be viewed as a measure of the "theoretical consistency" of a correlation matrix. Other indices could serve this function, but any index must share a common feature with the *RMSr* to fit within the spirit of our approach. Namely, the index must be based on data that were not used to estimate the model parameters. The following comments should clarify this point.

A path model entails a system of equations that relate correlations in **R** to model parameters, **P**. If the path model is recursive and overidentified, **R** can be partitioned into nonoverlapping sets of correlations, R1 and R2. These sets play logically distinct roles in our method of theory corroboration. Path coefficients are estimated from the correlations in R1. These correlations are reproduced perfectly because the path coefficients are derived from a just-identified set of simultaneous equations. The risky test comes into play when we use these coefficients to produce point estimates of the correlations in R2. (We have previously shown that the correlations in R2 are not highly constrained by the values in R1 unless they have coefficients of determination near 1.00, a situation that rarely obtains in the behavioral sciences.) We judge the accuracy of these point predictions by computing a root-mean-square error of discrepancy using the correlations—and only those correlations—that were not used to estimate the model parameters. This process of parameter estimation, data prediction, and model assessment can be summarized as a logical argument that includes a hypothesis, *H;* a deductively inferred consequence, *C;* and one or more obtained events E_i:

H: If D* has verisimilitude and (auxiliary assumption) path coefficients, **P**, are accurately estimated from R1, then

C: **P** entails a set of point predictions for the elements in R2.

*E*1: Prediction is deemed accurate by small $RMSr_{D^*}$

or

*E*2: Prediction is deemed inaccurate.

Notice that if $RMSr_{D^*}$ is large, then by the *modus tollens* argument, either D* does not have verisimilitude or the auxiliary assumption is untenable. Choosing between these alternatives is accomplished by comparing the estimates in **P** with full information parameter estimates (e.g., ML estimates). If the two sets of estimates are close and $RMSr_{D^*}$ is large, D* is no longer considered a plausible approximation of the causal structure.

Naturally, passing the above test does not imply that D* is true. The literal falsity of D* is acknowledged before the study is conducted. All we can say at this point, and that is saying a lot, is that D* can accurately account for data points that were not used to estimate the model parameters. Using Mulaik's (2002) terminology, this aspect of our approach achieves *objectivity* (see Mulaik's example of line fitting). However, we wish to say more. Specifically, we want to know whether D* accounts for the correlations in R2 better than a large class of structurally close models. If D* truly includes the important causal paths of the system under study, then our ability to accurately predict those correlations should decrease whenever D* is modified, say, by deleting a model-consistent and adding a model-inconsistent causal path.

Again, by a *modus tollens* argument, if D* does not reproduce the correlations in R2 better than a large class of theory-inconsistent models, then the evidentiary weight for D* is weak. However, if D* passes this second risky test, then it has proved its mettle for a second time and we conclude that it has verisimilitude. Lest we be misunderstood, we view this assessment as open to modification as we continue testing the causal implications of our model under novel conditions.

The Metric Problem: Standardized Versus Unstandardized Variables

Before leaving this section we briefly describe why VERIPATH uses standardized path coefficients. Reichardt (2002) correctly noted that standardized coefficients tend to vary across populations more than unstandardized coefficients. Unfortunately, this observation neglects the fact that there are many unstandardized coefficients for a given path and that, from one point of view, the metric of a coefficient is arbitrary. Consider his example, which, parenthetically, is highly unusual because it is deterministic with no residual variances. A group of youngsters measure their heights prior to getting dressed. The children put on their clothes, step into their shoes, and measure their heights a second time. This simple model includes three variables: (a) height before getting dressed (HB), (b) shoe height (SH), and (c) height

after getting dressed (HA). Reichardt presumed that all variables are measured in centimeters.

To breathe life into this example, we simulated height data for 20 fourth graders. Using a figure we located on the Internet, we assumed that the average fourth grader is 53 in. (134.62 cm) tall ($SD = 2$ in., or 5.08 cm). We also assumed that shoe soles have an average height of 1 in. (2.54 cm; $SD = 0.2$ in., or 0.51 cm) and that taller children have slightly larger shoes than shorter children. Thus the correlation between HB and SH was set to .50 in our model. Using these parameters, we simulated data for 20 fourth-grade boys. These data are reported in Table 3. Figure 6 displays the path model for this example with alternative scalings for the path coefficients.

Recall that all of Reichardt's (2002) variables denote a common attribute (height) that is measured on a common scale (centimeters). This is another aspect of the model that differentiates it from traditional path models in the social sciences. Figure 6A portrays Reichardt's model and shows that it is deterministic because HA is an additive function of HB and SH. When the dependent variable is a unit-weighted sum of the independent variables, and when all variables are measured on a common metric, the path coefficients are indeed 1.00 as noted by Reichardt.

Table 3
Sample Data to Illustrate the Effects of Standardization on Path Coefficients

Height before dressed (cm)	Shoe height (cm)	Shoe height (ft)	Height after dressed (cm)
127.13	2.20	0.072	129.33
141.79	2.95	0.096	144.74
144.35	2.12	0.069	146.47
136.20	2.86	0.093	139.06
143.89	2.80	0.091	146.69
137.31	3.48	0.113	140.79
121.38	2.38	0.077	123.76
145.74	3.61	0.117	149.35
135.73	1.56	0.051	137.29
129.96	1.72	0.056	131.68
132.87	2.63	0.085	135.50
132.66	2.61	0.085	135.27
124.92	2.05	0.067	126.97
128.24	2.53	0.082	130.77
131.87	2.58	0.084	134.45
130.69	2.59	0.084	133.28
129.39	2.09	0.068	131.48
144.91	2.57	0.084	147.48
139.16	2.58	0.084	141.74
136.63	2.91	0.095	139.54

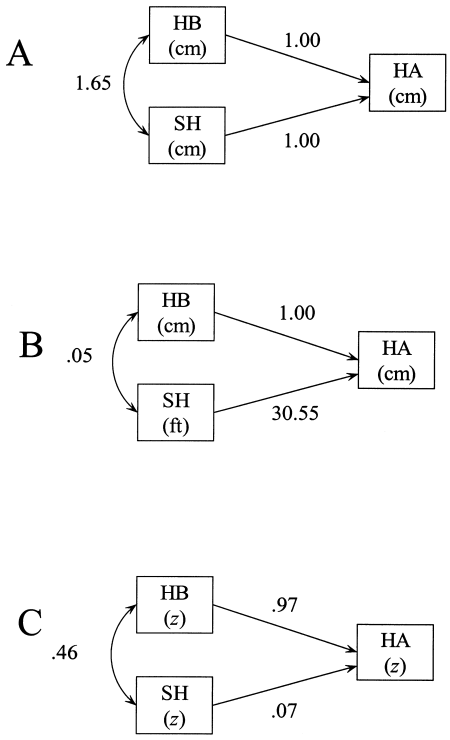


Figure 6. Three alternative scalings for path coefficients. A: Reichardt's (2002) model. B: The same model with shoe height (SH) reported in feet rather than in centimeters. C: The model with standardized coefficients. HB = height before getting dressed; HA = height after getting dressed.

Figure 6B portrays the same model with shoe height reported in feet rather than centimeters. Notice that the path from shoe height to final height is more than 30 times its former value. What is important to realize here is that both figures report unstandardized path coefficients. Now suppose that we broaden the model to include a second dependent variable called basketball prowess (BP). Presumably, among fourth-grade boys, height confers an advantage on the basketball court as well as the dance floor. Suppose we are interested in the path from HB → BP. What is its expected size? The answer is: We do not know and we cannot know unless we are given additional information about the metric of the basketball prowess scale. Moreover, unless we were unusually familiar with this scale and how it covaries with height measured in centimeters, we would be hard-pressed to classify the HB → BP path as being negligible or sizable. Yet we are asked to make exactly this type of decision if we are to implement one of Reichardt's (2002) model

appraisal methods. Let us be clear. We do not mean to criticize the underlying rationale of his approach, which we find very interesting. We merely wish to point out that the utility of his method may be severely limited by the lack of well-established metrics for psychology variables. Continuing with this example, Figure 6C reports standardized coefficients and correctly shows that HB is an important determinant of HA whereas the causal impact of SH is minimal (this is why children with stunted growth are given growth hormones rather than elevator shoes). The delete 1–add 1 rule does not deem a path negligible because it is smaller than an arbitrary number. Paths are deemed negligible when they can be fixed at zero without seriously compromising our ability to reproduce the correlations in R2. It is easy to construct models where the elimination of a small path would seriously degrade the model's performance.

Comments on Mulaik's "Commentary on Meehl and Waller's (2002) Path Analysis and Verisimilitude"

Mulaik (2002) emphasized the role of metaphor in scientific reasoning. He raised known weaknesses in several philosophical schools and concluded that VERIPATH is based on faulty logic. More specifically, he claimed that "the quasi-deductive argument that Meehl and Waller (2002) used to frame their method is not a sound deduction" (p. 316). We respectfully disagree. In previous sections we touched on the logic of VERIPATH. In this section we address the broader philosophical underpinnings of our work.

We adhere to no philosophical school, holding that a scientist should treat philosophical writings cafeteria style, taking what is useful in thinking about a scientific problem and bypassing the rest. For example, although we have learned lessons from the logical positivists, the chief inspiration for developing VERIPATH was Sir Karl Popper (1959, 1962, 1983), who was not a member of the Vienna Circle and who in his autobiography boasted of being the "murderer" of logical positivism. Mulaik (2002) elaborated his personal philosophy of science, parts of which we agree with and parts not, and aspects of both categories are still controversial among philosophers. Adequate examination of them requires more space than is available and would be inappropriate in this forum. Most are, we think, irrelevant to VERIPATH. Instead we prefer to state a minimal set of philosophical assumptions necessary to rationalize our VERIPATH proposals:

1. There is an objective physical world that exists independently of our minds.
2. Causal relations are a feature of that world.
3. Describing those relations symbolically (text, mathematics, schematic diagrams, pictures) is a legitimate aim of science.
4. Formulations of those relations differ in their accuracy (verisimilitude).
5. Verisimilitude is a correlate of the formulations' power to predict novel observational facts (to pass severe, risky tests).

Most philosophers and, so far as we know, all practicing scientists, act in accordance with this last assumption, despite the absence of a satisfactory metatheorem to that effect (for a proposed amelioration of this deficit, see Meehl, 1992a, 2002, in press). The importance of severe tests, although articulated most vigorously in recent years by Popper (1959, 1962, 1983), was of course not his invention; it can be found in 19th century thinkers such as Whewell (1847/1966), Jevons (1874/1958), and Peirce (1986) and is argued by contemporaries such as Mayo (1991, 1996) and Salmon (1989). The relation between normative and descriptive metatheory is a complex question currently lacking consensus. The norms of theory appraisal are rules (logic, semiotic, probability calculus, and honest protocols) and guidelines (e.g., parsimony, numerical precision, and novelty and diversity of predictions). Metatheory is in a sense an empirical discipline, whose database consists of history of science and current scientific practice. The normative stems from the descriptive for one who considers science to be, on the average and in the long run, a success. On this view, the main factual source of prescriptions is episodes in the history of science, the psychology and sociology of science being of lesser importance (Faust & Meehl, 1992, in press; Meehl, 1992a, 1992b, 2002, in press). Our stance in developing VERIPATH is normative. One who does not accept the first four philosophical assumptions listed above would presumably not be interested in path analysis in the first place; anyone who accepted the first four but rejected the relation between verisimilitude and passing a severe test would not be interested in our proposed solution.

As to induction versus deduction, the theory T should be strong enough to deduce the set of permis-

sible diagrams, a topological consequence without the causal weights. If T is too weak to provide even that little, one should probably not engage in path analysis. The inductive component is the (probabilistic, nondeductive) inference from passing a risky test to theory verisimilitude, briefly stated in the fifth metapostulate listed above. We bypass the philosophers' persisting disagreements on what licenses this transition (for discussion of these points, see Mayo, 1991, 1996; or Salmon, 1989). Rather, we assume that usual scientific practice warrants reliance on it. Whether one prefers inductivist terms such as *confirmation* (Carnap, 1936–1937, 1945, 1949; and see the discussion of degree of confirmation in Meehl, 1954/1996, pp. 34–36) or deductivist *corroboration* (Popper, 1959—merely failure to refute *modus tollens*), the common scientific feature is quite clear: If a conjecture, however arrived at, gained no rational credibility as an approximation to objective truth (verisimilitude) by generating (nomologically or probabilistically) risky numerical predictions, then nearly all empirical science would be impossible. Since the time of Galileo, empirical scientists have proceeded in this manner (see, e.g., Salmon, 1989), and it is good enough for us. One hopes the philosophers will be able to rigorize a metaproof that this is a rational policy, but science cannot be suspended in the meantime. Meanwhile, whether we label the scientific inference process as *inductive* or *deductive* is unimportant semantics. The logical empiricists realized over a half-century ago that in the so-called deductive–nomological model of explanation, the "deduction" from theoretical premises to observational conclusion is itself often deduction of a probability. The important thing is to be clear about the structure of the inference and its data source, which we think VERIPATH achieves. For example, as Meehl and Waller (2002) stated clearly, the transition from the correlations in R1 to the estimated coefficients (weights) via the topological path diagram is deductive computationally, as is the deduction of the predicted correlations in R2 from the weights. However, the procedure as a whole is inductive, starting with observed correlations subject to sampling error, with these correlations relying on probabilistic statistical inference, and concluding with an appraisal of T based on the probability of $\{D_T\}$ doing as well as it did if the motivating T had small verisimilitude.

This should clarify where our views differ from Mulaik's (2002) on important foundational issues. On some issues we agree; however on others Mulaik detected agreement where none exists. For instance, Mulaik is "pleased to say, Meehl and Waller (2002) also showed an inclination to abandon the idea of not only a final truth but exact truth" (p. 317). Once again, we respectfully submit "it just is not so." As philosophical realists, we believe in both final and exact truths (the terms are synonymous in our view). Nevertheless, as we stated previously, we do not consider path analysis models as accurate snapshots of nature. Rather, we view them as impressionistic paintings. The topological structure of the landscape is recognizable, but the details are fuzzy. The recent review by McDonald and Ho (2002, see Table 2) should convince anyone that most path models are gross representations of causal systems, at best.

We have never suggested that VERIPATH be used for exploratory purposes. Thus we are baffled by Mulaik's claim that "because ostensibly one would pick the best fitting model of the compared alternative models generated, [Meehl and Waller's] proposal is essentially no different from what takes place in the parameter estimation process" (Mulaik, 2002, p. 319). What is the logical rationale for this statement? If a researcher wittingly divides a probability value in half because editors prefer small probability values to large ones, we do not point an accusatory finger at Karl Pearson or Sir Ronald Fisher. Likewise, if someone uses VERIPATH for other than its intended purpose: Do not blame us! VERIPATH is not an exploratory tool, and thus it has no "similarities to what we do already (see the third and fourth step of the four-step procedure of Mulaik & Millsap, 2000)" (Mulaik, 2002, p. 320).

Summary

The strictly statistical criticisms of our procedure seem to divide into two broad classes, the first consisting of valid points, suggesting refinements that might lead to improved parameter estimates. Whether they are right or not can be ascertained by empirical and Monte Carlo investigation. Our data suggest that VERIPATH estimates are frequently identical to ML estimates. When they are not, they are typically very close or good enough. The second class says because we do not know how accurate some of the parameters are, the proposed method is not useful. Yes, we should look into such things as how inaccurate the parameter estimates are, how they might be made more accurate, and whether one index of badness of fit is better than another. However, pending those investigations, we still have the final step of the argu-

ment, that if the cumulative effect of all the error sources (leading to and including bad or less than optimal parameter estimates) were vitiating, then we would not be able to pass the multiple risky tests of our approach. Our method gets the "right answer." Not the "right numbers"—we know the numbers are wrong, and the whole point of our procedure is to bypass that incurable problem that the numbers will (most probably) always be, strictly speaking, inaccurate.

References

Alwin, D. F. (1988). Measurement and the interpretation of effects in structural equation models. In J. S. Long (Ed.), *Common problems/proper solutions: Avoiding error in quantitative research* (pp. 15–45). Newbury Park, CA: Sage.

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.

Carnap, R. (1936–1937). Testability and meaning. *Philosophy of Science, 3,* 420–471.

Carnap, R. (1945). The two concepts of probability. *Philosophy and Phenomenological Research, 5,* 513–532.

Carnap, R. (1949). Truth and confirmation. In H. Feigl & W. Sellars (Eds.), *Readings in philosophical analysis* (pp. 119–127). New York: Appleton-Century-Crofts.

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test taking motivation. *Journal of Applied Psychology, 82,* 300–310.

Cudeck, R. (1991). Comments on "Using causal models to estimate indirect effects." In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 260–263). Washington, DC: American Psychological Association.

Faust, D., & Meehl, P. E. (1992). Using scientific methods to resolve enduring questions within the history and philosophy of science: Some illustrations. *Behavior Therapy, 23,* 195–211.

Faust, D., & Meehl, P. E. (in press). Using meta-scientific studies to clarify or resolve questions in the philosophy and history of science. *Philosophy of Science.*

Glass, G. V., & Collins, J. R. (1970). Geometric proof of the restriction of possible values of r_{xy} when r_{xz} and r_{yz} are fixed. *Educational and Psychological Measurement, 30,* 37–39.

Ho, R., Davidson, G., Van Dyke, M., & Agar-Wilson, M. (2000). The impact of motor vehicle accidents on the psychological well-being of at-fault drivers and related passengers. *Journal of Health Psychology, 5,* 33–51.

Hubert, L. J. (1972). A note on the restriction of range for Pearson product–moment correlation coefficients. *Educational and Psychological Measurement, 32,* 767–770.

Jevons, W. S. (1958). *The principles of science.* New York: Dover. (Original work published 1874)

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage.

Jöreskog, K. G., & Sörbom, D. (2001). *LISREL 8.50 user's reference guide* [Computer software manual]. Chicago: Scientific Software.

MacCallum, R. C., Browne, M. W., & Preacher, K. J. (2002). Comments on the Meehl–Waller (2002) procedure for appraisal of path analysis models. *Psychological Methods, 7,* 301–306.

MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common factor model: Implications for theory and practice. *Psychological Bulletin, 109,* 501–511.

McDonald, R. P., & Ho, M. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7,* 64–82.

Mayo, D. G. (1991). Novel evidence and severe tests. *Philosophy of Science, 58,* 523–552.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge.* Chicago: University of Chicago Press.

Meehl, P. E. (1992a). Cliometric metatheory: The actuarial approach to empirical, history-based philosophy of science. *Psychological Reports, 71,* 339–467.

Meehl, P. E. (1992b). The Miracle Argument for realism: An important lesson to be learned by generalizing from Carrier's counter-examples. *Studies in History and Philosophy of Science, 23,* 267–282.

Meehl, P. E. (1996). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Northvale, NJ: Jason Aronson. (Original work published 1954)

Meehl, P. E. (2002). *Cliometric metatheory III: Peircean consensus, verisimilitude, and the asymptotic method.* Manuscript in preparation.

Meehl, P. E. (in press). Cliometric metatheory II: Criteria scientists use in theory appraisal and why it is rational to do so. *Psychological Reports.*

Meehl, P. E., & Waller, N. G. (2002). The path analysis controversy: A new statistical approach to strong ap-

praisal of verisimilitude. *Psychological Methods, 7,* 283–300.

Mulaik, S. A. (2002). Commentary on Meehl and Waller's (2002) path analysis and verisimilitude. *Psychological Methods, 7,* 316–322.

Olkin, I. (1981). Range restrictions for product–moment correlation matrices. *Psychometrika, 46,* 469–472.

Pearl, J. (2000). *Causality.* Cambridge, England: Cambridge University Press.

Peirce, C. S. (1986). *Writings of Charles S. Peirce* (C. J. W. Kloesel, Ed.). Bloomington: Indiana University Press.

Popper, K. R. (1959). *The logic of scientific discovery.* New York: Basic Books. (Original work published 1935)

Popper, K. R. (1962). *Conjectures and refutations.* New York: Basic Books.

Popper, K. R. (1983). *Postscript: Vol. 1. Realism and the aim of science.* Totowa, NJ: Rowman & Littlefield.

Reichardt, C. S. (2002). The priority of just-identified, recursive models. *Psychological Methods, 7,* 307–315.

Rettig, K. D., Leichtentritt, R. D., & Stanton, L. M. (1999). Understanding noncustodial fathers' family and life satisfaction from resource theory perspective. *Journal of Family Issues, 20,* 507–538.

Salmon, W. C. (1989). *Four decades of scientific explanation.* Minneapolis: University of Minnesota Press.

Steiger, J. H. (2000). Point estimation, hypothesis testing, and interval estimation using the RMSEA: Some comments and a reply to Hayduk and Glaser. *Structural Equation Modeling, 7,* 149–162.

Whewell, W. (1966). *The philosophy of the inductive sciences, founded upon their history.* New York: Johnson Reprint. (Original work published 1847)

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 20,* 557–585.

Wright, S. (1960). Path coefficients and path regression: Alternative or complementary concepts? *Biometrics, 16,* 189–202.

Received April 30, 2002
Revision received May 14, 2002
Accepted May 14, 2002 ■