Psychological Methods
2002, Vol. 7, No. 3, 283–300

1082-989X/02/$5.00 DOI: 10.1037//1082-989X.7.3.283

The Path Analysis Controversy: A New Statistical Approach to Strong Appraisal of Verisimilitude

Paul E. Meehl
University of Minnesota

Niels G. Waller
Vanderbilt University

A new approach for using path analysis to appraise the verisimilitude of theories is described. Rather than trying to test a model's truth (correctness), this method corroborates a class of path diagrams by determining how well they predict intra-data relations in comparison with other diagrams. The observed correlation matrix is partitioned into disjoint sets. One set is used to estimate the model parameters, and a nonoverlapping set is used to assess the model's verisimilitude. Computer code was written to generate competing models and to test the conjectured model's superiority (relative to the generated set) using diagram combinatorics and is available on the Web (http://www.vanderbilt.edu/quantmetheval/downloads.htm).

Path analysis, whatever its ultimate fate, will present historians and philosophers of science with a puzzle. Ever since Sewall Wright's classic (1921) article and the ensuing exchanges with his critic, Niles (1922, 1923; Wright, 1923), there has been sharp disagreement about the method's value (Breckler, 1990; Cliff, 1983; Karlin, Cameron, & Chakraborty, 1983; McKim & Turner, 1997). Some consider it one of the best ways to make causal inferences from correlational data; others view it as not only worthless but misleading and not to be used at all (e.g., Freedman, 1987, 1997; Rogosa, 1987). Today, after three generations of discussion by statisticians, psychologists, sociologists, economists, political scientists, geneticists, and philosophers of science, the disagreement still seems about as great and irresoluble as in the Wright–Niles exchange (for historical reviews of path analysis and structural equation models, see Aigner, Hsiao, Kapteyn, & Wansbeek, 1984; Austin & Wolfe, 1991; Bentler, 1980; Bielby & Hauser, 1977; Epstein, 1987; Shipley, 2000).

We are grateful to Judea Pearl, William Rozeboom, Keith Widaman, Leslie Yonce, and Steve West for helpful comments on an earlier version of this article.

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

Wherever one may stand as to the merits of path analysis—or to the family of techniques known as covariance structure models (CSM)—it must be admitted that the persistence of this disagreement requires explanation. After all, Wright's (1921) original equations are merely basic algebra (for the mathematics of path analysis, see Bollen, 1989; Duncan, 1975); thus it can hardly be a dispute about the mathematics, about theorems, about the formalism as such. Nor can it be a dispute about the facts, which consist of observed correlation (or covariance) coefficients that can be made as stable and robust as we wish by increasing sample size.[1] The explicit goal of the procedure is the inference from a set of correlations to a set of causal paths with their inferred quantitative weights. We propose that the persisting disagreement is neither mathematical nor factual, but methodological or, if you like, "philosophical" in nature. Although the rules of the formalism are not problematic, the empirical scientist still has a certain amount of freedom, because one can move through the formalism via different epistemic paths. Developing and defending this thesis is the purpose of this article, along with a novel statistical approach that is suggested thereby.[2]

[1] That some social scientists use path analysis on samples so small as to permit disputes about the estimates is a reflection of a widespread bad habit of our profession, not a valid objection to the method.

[2] Some readers may question our focus on path analysis rather than latent variable models, the latter being known as

The core of our position is implicit in the first two paragraphs of Wright's 1921 article. First he said that when the causal relations are known, path analysis enables one to infer the causal weights. That causal structure is of course among the questionable "assumptions" challenged by Freedman (1987, 1997) and others (e.g., Clogg & Haritou, 1997). Wright was dealing with biology, in which the causal relations are sometimes known on other grounds (e.g., experimental manipulation, strongly corroborated theory) or—as in his guinea pig example (Wright, 1921)—are easily limited to a small number of possibilities. In the second paragraph, however, he said that when the causal relations are not known, the method provides an empirical test of hypotheses regarding them. This second sort of application leads to an important thought in the framework of contemporary philosophy of science. If the structure of a theory permits facts to refute it, then, at least in principle, facts can corroborate the theory by failing to refute it, provided that the factual prediction is otherwise (somehow) risky. For the corroboration to be strong, we have to have "Popperian risk" (Popper, 1959/1977, 1962), "severe test" (Mayo, 1991, 1996), or what philosopher Wesley Salmon called a *highly improbable coincidence* (Salmon, 1984). We show how that can be achieved in the path analytic context.

One hears the objection "Correlation does not prove causality." If *prove* means *deduce,* of course it cannot in any empirical domain—courts of law, business, common life, or sciences. However, causal inference can be strongly corroborated—*proved,* in the usual sense of the term—by correlation. A nonartifactual replicable correlation between x and y shows that some sort of causal influence is at work. All science proceeds on that basis. Either x causes y; or y causes x; or some factor z causes both; or factors u and v, which cause x and y, respectively, are in turn caused by z; or. . . . The familiar cautionary adage should be amended to read something like "Correlation evidences causation but does not immediately reveal what causes what, and how much." Unraveling the causal influences may require complicated statistical inferences. One who did not believe such inferences are rational and feasible would not engage in path analysis in the first place. Wright (1921, 1923) invented the procedure solely for the purpose of making causal inferences; path analysis would be a needless and controversial complication if one had only a descriptive or predictive aim.

The epistemology is simple, straightforward, and similar to inferences in other empirical sciences: A substantive theory T implies factual relations $\{F_i\}$; if verified, $\{F_i\}$ corroborates T. If T contains causal statements $\{C\}$ playing an essential role in the derivation $\mathrm{T} \rightarrow \{F_i\}$, then $\{C\}$ is corroborated (along with whatever noncausal statements T contains). It is that simple, qualitatively speaking. A causal component of T has nothing special, suspicious, or dangerous that should lead one to cavil at it as such; it is merely somewhat difficult to unscramble those relations via the statistics. If for some reason (e.g., background knowledge) we assign a very low probability (everyone, not only Bayesians, takes account of the context in which statistics are computed) to a causal component of T, then the question is whether an amended T without that component is capable of deriving the (successful) risky consequences $\{F_i\}$. If not, $\{C\}$ seems essential, so we face the usual scientific judgment call on partially conflicting evidence.

In psychology, all of our causal theories are incomplete, and in the "soft fields" (e.g., psychopathology, personality theory, social psychology), they are always partially false. In the context of causal modeling, Browne and Cudeck (1993) expressed a similar view:

> In the social sciences it is implausible that any model we use is anything more than an approximation to reality. Since a null hypothesis that a model fits exactly in some population is known *a priori* to be false, it seems pointless even to try to test whether it is true. (p. 137)

Other authors have also stressed the view that models are approximations and are therefore incomplete (Cudeck & Henley, 1991; Linhart & Zucchini, 1986). Although a substantive causal theory may be too weak

covariance structure models, LISREL models, or structural equations with latent variables. There are three reasons for focusing on path analysis. First, the mathematics is less complicated and thus easier to discuss. Second, although we have not tested this claim, we see no reason why our approach could not be used on the structural portion of a latent variable model once the measurement model was fixed (Anderson & Gerbing, 1988). Third, although latent variable models are popular in the social sciences, path analysis continues to be used regularly to test causal structures. For example, in their recent review of causal modeling practices, MacCallum and Austin (2000) noted that 25% of the studies they reviewed used path analysis with no latent variables.

to generate precise numerical predictions, especially in the behavioral sciences, where marked individual differences (e.g., between organisms) play an important role, the theory may be structurally strong enough to deduce that observations should be related in certain ways. This gives us an alternative way to obtain risky predictions, namely, we can try to derive one portion of the data from another portion. The best we can hope for in those fields is to work with an idealized formulation of the theory that is capable of deriving a topological structure; use that structure to derive quantitative relations between some variables (observed) and others (predicted and to be observed); compute an index that compares the predicted values with the observed; and then—what is currently not done, or is done with *ad hoc*ery—compare the predictive performance of our theory with that of its competitor theories. Arguing from successful prediction of a subset of the observations from another subset uses inductive logic (rather than strict deduction); hence the degree of corroboration given by a failure to falsify the hypothesis depends on the predictive risk we take. Modifying this general principle of inductive logic for the idealized (literally false) case, we say that a weak theory's less erroneous prediction ("closer to the facts") corroborates it to the extent that the selection of the prediction-mediating causal diagram was antecedently improbable absent the theory. We can quantify this riskiness for path diagrams by using combinatorics.

In the social sciences, the path diagram motivated by a substantive causal theory is almost always literally incorrect, with or without the path coefficients written in. As noted by Browne and Cudeck (1993), statistical testing of the diagram's correctness, taken literally, is usually pointless because we know that the theory is imperfect and hence the diagram, taken literally, is incorrect; thus, when we do a purportedly exact statistical test, all we are actually doing is studying the statistical power function.

Furthermore, as Freedman (1987) and others have forcibly argued, even if the topological structure were known, or even if (almost never the case) the theory were strong enough to deduce the numerical values of the path coefficients, the statistical search or test procedure typically relies on at least seven statistical auxiliaries that are doubtful and usually false in the behavioral and life sciences. For example, limiting our attention to the form of path models used most commonly in the literature, that is, recursive path models without constraints (thereby allowing us to work with correlations rather than covariances; see Cudeck, 1989), we assume the following:

1. linear causal relations,
2. no reciprocal feedback (reciprocal feedback is implied for variables A and B whenever $A \rightarrow B$ and $B \rightarrow A$),
3. no causal loops (causal loops are implied for variables A, B, and C if $A \rightarrow B \rightarrow C \rightarrow A$),
4. uncorrelated disturbances,
5. manifest variables are direct measures of causal factors rather than proxies (e.g., socioeconomic status is a proxy for the rich complex of a home environment and social reinforcement regime),
6. model self-containment (self-containment is achieved whenever all relevant causes of an effect are included in the model; failure to achieve self-containment is known as the *omitted variables problem*), and
7. manifest variables are perfectly reliable.[3]

The average probability of any one of these auxiliaries' being literally correct in the social sciences is surely $p < .50$ (we suggest that only the first of these assumptions has an even chance), so if their truth frequency in a class of experiments were independent and if one performed 100 path analyses with recursive models in the course of a research career, in less than one study would they all be correct. When the weakness of the theories is combined with the improbability of the auxiliaries, it seems as if Freedman (1987, 1997) wins the argument hands down if we restrict our attention to the assessment of exact fit.

We think there is a solution to the problematic use of path analysis in the life sciences. Our philosophy of science adopts a different stance with regard to what

[3] In nonrecursive models (Berry, 1984), the second and fourth assumptions are relaxed. Kenny and Judd (1984) and others (Li et al., 1998) have demonstrated how to include nonlinear and interactive effects (the first assumption) in structural equation models. Latent variable models were developed to avoid the seventh assumption. An insightful discussion of the assumptions underlying causal analysis can be found in chapters 2 and 3 of James, Mulaik, and Brett (1982).

is being tested and how a severe test can be provided despite the admitted vagueness and incompleteness of a theory, the resulting imperfection of the motivated path diagram(s), and the presumed falsity of the conjoined auxiliaries. The crucial distinction between our approach and the conventional one lies in our reformulation of (a) the character of the appraisal task, (b) the selection of an alternative epistemic path via the formalism, and (c) a severe nonparametric test based on path diagram combinatorics.

Subjecting Path Models to Severe Tests

Let D denote a diagram implied by a path analysis model. Although any particular diagram D is incorrect—in the sense that D represents a model known a priori to be false—a substantive theory T motivates a preferred diagram D*, and it entails a class of tolerated diagrams, $T \rightarrow \{D_T\}$. Tolerated diagrams are those that the causal theory T permits. T favors D* but allows specified path substitutions that generate $\{D_T\}$. If members of nontolerated set $\{-D_T\}$ predict better, T is strongly discorroborated. Our aim is to compare the success of our conjectured D*, and a narrow class of diagrams $\{D_T\}$ tolerated by T, with an important subclass of diagrams $\{-D_T\}$ that T forbids. In doing this, we are indirectly comparing the adequacy of T with conceivable competing theories whether known or not (although not with all competing theories).

It should be stressed at this point that our limited aim is to corroborate the topological structure of a theory—as portrayed in a path diagram—rather than specific components of that structure. Accordingly, our approach does not allow one to test hypotheses of the form path $a \geq$ path b (unless $b = 0$) or to test various linear or nonlinear constraints on parameter estimates.

The preferred diagram D* relates the observed correlations $\{r_j\}$ to a conjectured set of causal arrows with path coefficients $\{c_i\}$, thus $D^* \rightarrow \text{Rel}[\{c_i\}, \{r_j\}]$, where $\text{Rel}[\{c_i\}, \{r_j\}]$ denotes the overall complex relations between the *r*s and the *c*s. This total configural relationship between the observed correlations and the conjectured causal paths can be parsed into relations between various subsets, each defining two composite relations Rel_1 and Rel_2 that have the *c*s in common but are disjoint with respect to the *r*s. Thus, having chosen a subset of the observed correlations, we can write a relation $\text{Rel}_1[\{c_i\}, \{r_j\}]$ and another relation $\text{Rel}_2[\{c_i\}, \{r_l\}]$ using the other correlations. If there are k path coefficients in the diagram for which we wish to solve, we require k polynomial equations, which means choosing k of the observed *r*s, plus the *r*s among the exogenous variables, as the source of our predictions. Not all subsets of correlations will work. However, we have found that the k polynomials can always be solved by choosing *r*s using the following rules: (a) Choose all *r*s among the exogenous variables, and (b) if there is a directed path, c_{ij}, between variables i and j, choose r_{ij}.

Solving these k polynomials for the k path coefficients that are the unknowns in our diagram, we plug these *c*s into the polynomials for generating the unused *r*s. We then compute an overall index of error that quantifies the internal consistency of the model-to-data and data-to-data relations. Specifically, using the m correlations that were not used to solve the k polynomials (and k path coefficients), we compute a root-mean-square residual (*RMSr*) correlation:

$$RMSr = \sqrt{\frac{1}{m}\sum_{i \neq j}(r_{ij} - \hat{r}_{ij})^2}. \quad (1)$$

Equation 1 has several attractive properties. First, recent work by Hu and Bentler (1998, 1999) demonstrates that the standardized *RMSr* is sensitive to model misspecification and relatively more sensitive to underparameterized misspecification than other popular fit indices.[4] Second, the interpretation of Equation 1 is straightforward because it quantifies model–data differences in the metric of the data (i.e., in the metric of the correlations). We prefer not to call Equation 1 a "goodness-of-fit" index but rather a "badness-of-fit" index, emphasizing that what we are doing is looking for the least bad job of internal data-to-data prediction achievable by the diagram. We know that T is imperfect (and probably false), and hence we know that D* is false, as are all of the diagrams $\{D_T\}$ that T tolerates.

We now get a "corrupted" D by substituting a causal arrow that was left blank in D*, striking an arrow in D*, and repeat the above procedure for that corrupted D; we do this repeatedly, getting a badness of fit for each corrupted D. We call this method for generating corrupted diagrams the *delete 1–add 1* rule. Obviously, one could devise an unlimited number of rules for generating so-called corrupted dia-

[4] Hu and Bentler studied a root-mean-square index that was defined slightly differently from our own index. The critical differences are noted in the text.

grams. We favor the delete 1–add 1 rule because it generates structurally close diagrams with the same number of directed paths.

Our first, most risky prediction is that the preferred D* does better than any of the corrupted Ds. However, we also permit ourselves a less risky prediction, given that a subset of alternative arrows is tolerated (although not preferred) by T, so it is possible that one of the corrupted diagrams will do better than D*. For example, if D* has seven causal arrows, and T tolerates two others in a slightly corrupted D, then $2 \times 7 = 14$ such substitutions are tolerated. This still leaves us with a strong prediction that the best diagram, although it may not be D* as we rationally hope, is to be found among this 14. This subset of $14 + 1 = 15$ tolerated diagrams is often a small proportion of the possible diagrams in highly overidentified models.

We understand our weak theory T does not literally say that all null arrows in the tolerated set $\{D_T\}$ are exactly zero but only that they are (we hope) "negligible" in causal strength compared with those drawn in members of $\{D_T\}$. Our calculations treat them as zero, despite our knowing better. Thus, our strong prediction is that all of these corrupting substitutions will make things worse. If our risky prediction motivated by T turns out to be correct, T is corroborated. T, despite being fairly weak in content, is strongly corroborated if the fit index (Equation 1) associated with D* is substantially smaller than the fit indices associated with a large number of corrupted diagrams.

An Illustration of the Method With a Hypothetical Example

Let us illustrate our method for appraising path models by considering the series of models in Figure 1. Assume that Figure 1A depicts the true and unknown relations among a set of variables. This simple model with four manifest variables includes five directed paths and an observed correlation (an undirected path) between two exogenous variables. The five path coefficients and the exogenous correlation exhaust the six nonredundant pieces of information in the 4×4 correlation matrix. Thus, we have a just-identified recursive model with two exogenous variables (x_1 and x_2) and two endogenous variables (y_1 and y_2). In Figure 1B, we have substituted letters for the path coefficients for ease of communication. Note that in all diagrams, the paths from the residual factors (disturbance terms, error variables) have been omitted to avoid clutter; nevertheless, they are estimated in the procedures outlined below.

The system of equations that are implied by a path diagram can be conveniently expressed using matrix algebra. Assume that Figure 1C represents a researcher's preferred model for the causal relations among variables x_1, x_2, y_1, and y_2. The system of equations that are implied by this path diagram can be written as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ a & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} b & 0 \\ c & d \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix} \quad (2)$$

or more compactly as

$$\boldsymbol{y} = \mathbf{B}\boldsymbol{y} + \boldsymbol{\Gamma}\boldsymbol{x} + \boldsymbol{\zeta}, \quad (3)$$

where $\mathbf{B}$ is a $p \times p$ matrix of coefficients denoting directed paths between endogenous variables, $\boldsymbol{\Gamma}$ is a $p \times q$ matrix of coefficients specifying paths from exogenous to endogenous variables, $\boldsymbol{y}$ is a $p \times 1$ vector of endogenous variables, $\boldsymbol{x}$ is a $q \times 1$ vector of exogenous variables and $\boldsymbol{\zeta}$ is a $p \times 1$ vector of errors in the equations. As previously stated, we assume that the variables in $\boldsymbol{y}$ and $\boldsymbol{x}$ are standardized such that each variable has a mean of 0.00 and a standard deviation of 1.00. Let $\boldsymbol{\Phi} = E(\boldsymbol{x}\boldsymbol{x}')$, the observed correlations among the exogenous variables, and $\boldsymbol{\Psi} = E(\boldsymbol{\zeta}\boldsymbol{\zeta}')$. In the recursive models considered in this article, $\boldsymbol{\Psi}$ is a diagonal matrix of residual variances. Under the aforementioned definitions and assumptions, the fixed and free parameters ($\boldsymbol{\theta}$) in Equation 3 imply a predicted correlation matrix:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{pmatrix} (\mathbf{I}-\mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi})(\mathbf{I}-\mathbf{B})^{-1'} & (\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi} \\ \boldsymbol{\Phi}\boldsymbol{\Gamma}'(\mathbf{I}-\mathbf{B})^{-1'} & \boldsymbol{\Phi} \end{pmatrix}. \quad (4)$$

In today's world of high-speed computers, researchers rarely estimate path coefficients using the original procedures outlined by Wright (1921; i.e., solving inhomogeneous polynomials). Some researchers use multiple regression to estimate path coefficients (see Pedhazur, 1997, pp. 765–803), although the most common approach is to use maximum likelihood (ML) or generalized least squares algorithms in CSM software (e.g., Arbuckle, 1995; Bentler, 1995; Jöreskog & Sörbom, 2000). These algorithms provide estimates of path coefficients with attractive statistical properties, such as asymptotic normality and efficiency, when their underlying distributional assumptions are satisfied. Because these techniques consider all of the correlations simultaneously, however, they cannot be used to test a model's verisimilitude using the procedures outlined in this article; moreover, because path models gener-

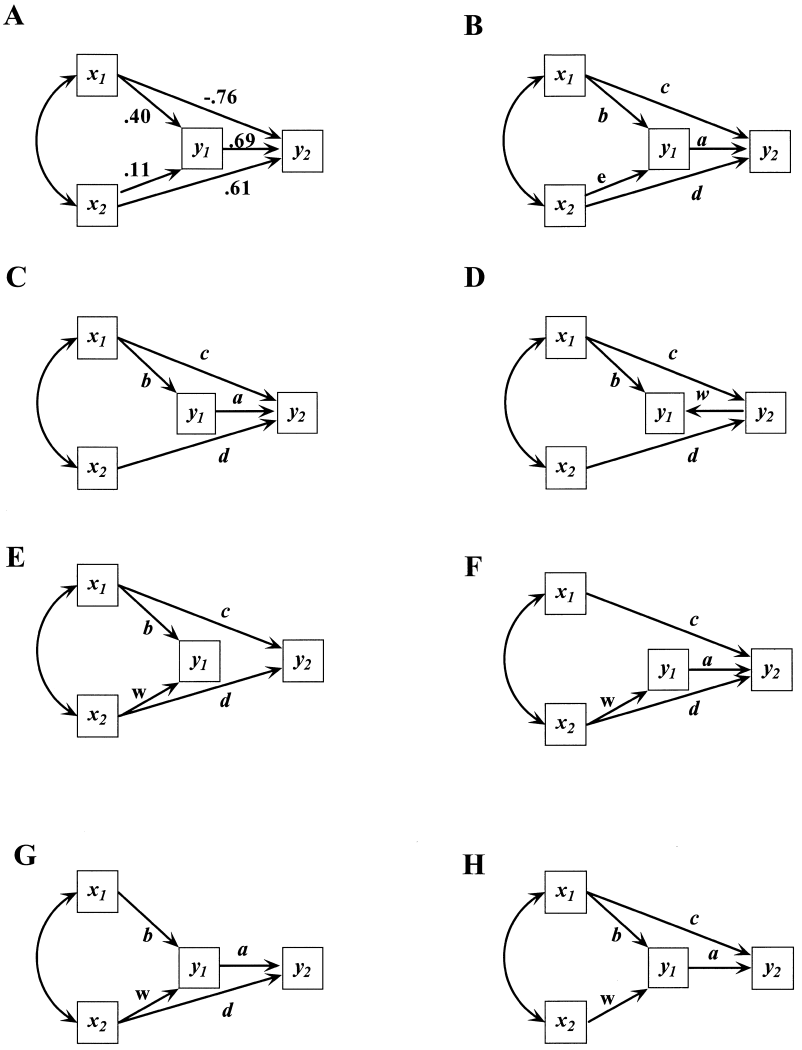

Figure 1. Example path diagrams generated by our delete 1–add 1 rule. A: The true and unknown relations among a set of variables. B: Paths with letters (*a–e*) substituted for path coefficients (for ease of communication). C: A researcher's preferred model for the causal relations among variables. D–H: Corrupted path models generated by the delete 1–add 1 rule. x_1 and x_2 are exogenous variables; y_1 and y_2 are endogenous variables; *w* represents the new path.

ally are not scale invariant (see Cudeck, 1989), full information methods should be applied to covariance matrices. For instance, ML estimates of path coefficients are calculated by minimizing a discrepancy function, such as the ML discrepancy function shown below:

$$F = \log \|\boldsymbol{\Sigma}\| + tr(\boldsymbol{S}\ \boldsymbol{\Sigma}^{-1}) - \log \|\boldsymbol{S}\| - (p + q), \quad (5)$$

where $\boldsymbol{S}$ denotes the observed correlation matrix, *tr* denotes the trace operator, $\|\boldsymbol{A}\|$ denotes the determinant of matrix $\boldsymbol{A}$ and all other terms are defined as above. Equation 5 illustrates an important property of full information methods; namely, all elements in the observed covariance or correlation matrix are taken into account when locating the ML estimates. This is an attractive feature of ML estimates. However, for model testing purposes, we argue that less efficient estimates may be preferable when appraising the verisimilitude of a model.[5] In the sections below, we dem-

[5] ML or other efficient estimates are preferable for estimating causal effects. We use less efficient estimates for model assessment purposes only.

onstrate several advantages of using a subset of correlations (covariances) to derive path coefficients and a disjoint subset to appraise the model. Our method of estimation is similar to Wright's original method, although we have formulated the problem using matrix algebra.

We have already shown how the model implied by Figure 1C can be expressed in matrix terms using Equation 2. Working through the operations implied by Equation 4 yields the following structure for the predicted correlations among our four hypothetical variables.[6]

$$\hat{r}_{(y_1,y_2)} = bc + bdr_{(x_1,x_2)} + a\,(1 - br_{(y_1,x_1)} + b^2), \quad (6)$$

$$\hat{r}_{(y_1,x_1)} = b, \quad (7)$$

$$\hat{r}_{(y_1,x_2)} = br_{(x_1,x_2)}, \quad (8)$$

$$\hat{r}_{(y_2,x_1)} = ab + c + dr_{(x_1,x_2)}, \quad (9)$$

$$\hat{r}_{(y_2,x_2)} = abr_{(x_1,x_2)} + cr_{(x_1,x_2)} + d. \quad (10)$$

Imagine that we have data on variables x_1, x_2, y_1, and y_2 and we have calculated the correlation matrix shown in Table 1. Our path model includes four unknowns: *a, b, c,* and *d.* We can solve for these unknowns using four of the five polynomials in Equations 6 through 10. To do so, we substitute observed correlations for predicted correlations in a sufficient set of four equations by following our aforementioned selection rule. Specifically, if there is a directed path, c_{ij}, between variables *i* and *j,* we choose r_{ij}. Following this procedure reveals that the four unknown model parameters can be estimated without any consideration of $r_{(y_1,x_2)}$. Solving the polynomials yields the following estimated path coefficients:

$$a = .76,$$
$$b = .43,$$
$$c = -.81,$$
$$d = .69.$$

Because $r_{(y_1,x_2)}$ was not used to estimate the model parameters, we can use this value to assess the internal consistency of the data given the hypothesized model structure. The standardized *RMSr* for the model (using only those correlations not used to estimate the parameters) is as follows:

$$RMSr = \sqrt{(r_{y_1,x_2} - \hat{r}_{y_1,x_2})^2},$$

$$.10 = \sqrt{(.23 - .13)^2}.$$

Of course, with only a single degree of freedom at our disposal, the observed *RMSr* is not particularly informative when it is considered in isolation (although if the *RMSr* was large, we would reject the model no matter how many degrees of freedom were available). However, because our basic philosophy stresses the importance of model comparisons, we do not consider the observed fit index in isolation. Rather, we repeat the above procedures on a series of corrupted path models as described in the previous sections.

Recall that we generate corrupted path models by following the delete 1–add 1 procedure. That is, we delete a path from D* and then add a new path.[7] By deleting and adding the same number of paths (i.e., one path) we keep the number of directed paths in the corrupted and original models equal. This is important because the fit of a model, as indexed by the *RMSr,* can only improve if we add new paths without deleting old paths. (When we add paths without deleting paths, empirical fit will always increase or remain unchanged even though theoretical fit may decrease. See Olsson, Troye, & Howell, 1999, for the important distinction between empirical and theoretical fit in structural equation models.)

Letting Figure 1C represent D*, the delete 1–add 1 rule generates five corrupted path models with identified parameters. The path diagrams for these models are portrayed in diagrams D through H of Figure 1. In all of these diagrams, the new path is labelled *w.*

Following the procedures outlined above, we calculated path coefficients for the five corrupted path

Table 1
Correlation Matrix for Example Path Analysis Model

Variable	1	2	3	4
1. y_1	—			
2. y_2	.50	—		
3. x_1	.43	−.28	—	
4. x_2	.23	.54	.30	—

[6] Some readers may question the accuracy of Equation 6. The equation is correct because $r_{(y_1,x_1)} = b$ in this model and thus the last term of the equation is a model-implied constraint.

[7] Note that after applying the delete 1–add 1 rule, we select a new set of correlations to solve the new set of polynomials. In general, the old and new sets of correlations will differ by one element only.

diagrams displayed in Figure 1, D–H. The *RMSr*s for D* and the five corrupted models are reported in Table 2. Notice that D*, the researcher's preferred model (Diagram C), fits the data substantially better than the alternative models when we quantify fit by *RMSr*. This does not prove that D* is the correct model. In fact, in this hypothetical example we know that D* is not the correct model because the data were simulated to fit Figure 1A. Nevertheless, these findings suggest that D* is a plausible model with verisimilitude. Other plausible models may also exist (in some cases, with high probability), some of which may produce equivalent empirical fit (Lee & Hershberger, 1990; MacCallum, Wegener, Uchino, & Fabrigar, 1993). One of the strengths of our approach is that the subset of equivalent models that are structurally close to D*—specifically, those models that differ by the reflection of a single directed path—will be included among the set of corrupted models. If this subset is large relative to the total set of corrupted models, the evidentiary weight in favor of D* is weakened because many structurally close models fit the data equally well. We consider this to be a positive attribute of a fit measure.

The models depicted in Figure 1 have few variables and few unknown parameters. We can solve for the path coefficients without the aid of a computer if we desire (as Wright, 1921, did many years ago), but even in a small model, this task becomes burdensome and error prone when we consider the potentially large number of corrupted models. More complex models may contain dozens of unknown parameters and generate literally hundreds of corrupted models. Thus, to make the methods described in this article more attractive to researchers, we wrote a Mathematica (Wolfram, 1999) routine called VERIPATH[8] that automates our approach to model appraisal.

Table 2
Root-Mean-Square Residual (RMSr) for D and Corrupted Path Models*

Diagram[a]	*RMSr*
C	.10
D	.32
E	.55
F	.36
G	.59
H	.55

[a] Diagram refers to the path models shown in Figure 1 (e.g., Diagram C here refers to the path model shown in Figure 1C). Diagram C is the researcher's preferred model (D*); Diagrams D–H are the corrupted path models generated by the delete1–add 1 rule.

A Real-Data Illustration

We now turn our attention to a real-data example that demonstrates our method for appraising the internal consistency of an observed correlation matrix given a path diagram. Our example comes from a recent publication by Rettig, Leichtentritt, and Stanton (1999). These authors collected data from 212 noncustodial fathers 3 years following a divorce to examine the predictors of family and life satisfaction in this population. The study was conducted from the perspective of resource theory (Foa & Foa, 1980). Data were collected on eight variables, two of which were considered exogenous causes of satisfaction: (a) perceived economic well-being and (b) social–psychological well-being. The six endogenous measures included (a) cooperative communication during conflict (with one's ex-spouse), (b) cooperative communication during coparenting, (c) low importance of resource deprivation, (d) low frequency of conflict, (e) involvement with children, and (f) a direct measure of family and life satisfaction. Consistent with an important auxiliary assumption of path analysis, all of the variables (questionnaires) had high internal consistency (alpha) reliabilities ($M = .88$, $SD = .08$). Readers are referred to the original publication for a fuller description of the measures and the rationale for the study. A greatly simplified version of resource theory (Foa & Foa, 1980) can be summarized as follows: (a) Individuals with more resources are more likely to give their resources to others, and (b) individuals with more resources are less likely to take resources from others. According to the study authors, cooperative communication during conflict and coparenting between ex-spouses are two forms of positive resource exchange. Table 3 shows the correlations (and standard deviations) among the eight study variables as reported in Table 1 of the original article. Figure 2 displays the authors' model as a path diagram.

In the original article, path coefficients were estimated via multiple regression. We reestimated the path coefficients using LISREL 8.30 (Jöreskog & Sörbom, 2000) by minimizing the ML discrepancy function in Equation 5. The test of exact fit for this model could not be rejected at the .05 significance level, χ^2 (14, $N = 212$) = 22.61, $p = .067$. Al-

[8] VERIPATH can be downloaded from the Web: www.vanderbilt.edu/quantmetheval (click on Downloads; your browser must support frames).

Table 3
Correlation Matrix for Real-Data Example

Variable	1	2	3	4	5	6	7	8
1. Perceived economic well-being	—							
2. Social–psychological well-being	.210	—						
3. Cooperative communication during conflict	.109	.046	—					
4. Cooperative communication during coparenting	.108	.038	.339	—				
5. Low importance of resource deprivation	.035	.152	.089	.275	—			
6. Low frequency of conflict	.066	.244	.203	.095	.372	—		
7. Involvement with children	.048	.082	.226	.498	.081	.184	—	
8. Family and life satisfaction	.289	.043	.182	.471	.344	.342	.235	—
SD	4.53	17.78	5.59	7.53	12.65	3.29	14.42	4.04

Note. From "Understanding Noncustodial Fathers' Family and Life Satisfaction From Resource Theory Perspective," by K. D. Rettig, R. D. Leichtentritt, and L. M. Stanton, 1999, *Journal of Family Issues, 20,* p. 524 (Table 1). Copyright 1999 by Sage Publications. Adapted with permission of Sage Publications, Inc.

though larger than .05, the observed probability value is close to the rejection region. Consequently, as with virtually all researchers who fit structural equation models, we consider alternative ways of assessing model fit. LISREL reports approximately two dozen fit indices, a virtual cornucopia that provides the unknowing (or perhaps unscrupulous) investigator with some latitude to pick and choose the most flattering measures for a given model. On the basis of recent work of Hu and Bentler (1998, 1999), we consider two fit indices before turning our attention to the methods described in this article. These fit indices are the root-mean-square error of approximation (RMSEA) and the standardized root mean square residual (SRMR). For the model displayed in Figure 2, RMSEA = .052 and SRMR = .045. In light of Hu and Bentler's recent Monte Carlo findings, these values indicate moderate-to-close fit of the proposed model.

A limitation of these fit indices—one that constrains our ability to assess the verisimilitude of the model—is that they do not compare the fit of an investigator's preferred model with that of closely related alternative models. We have argued in previous sections that an informative class of alternative mod-

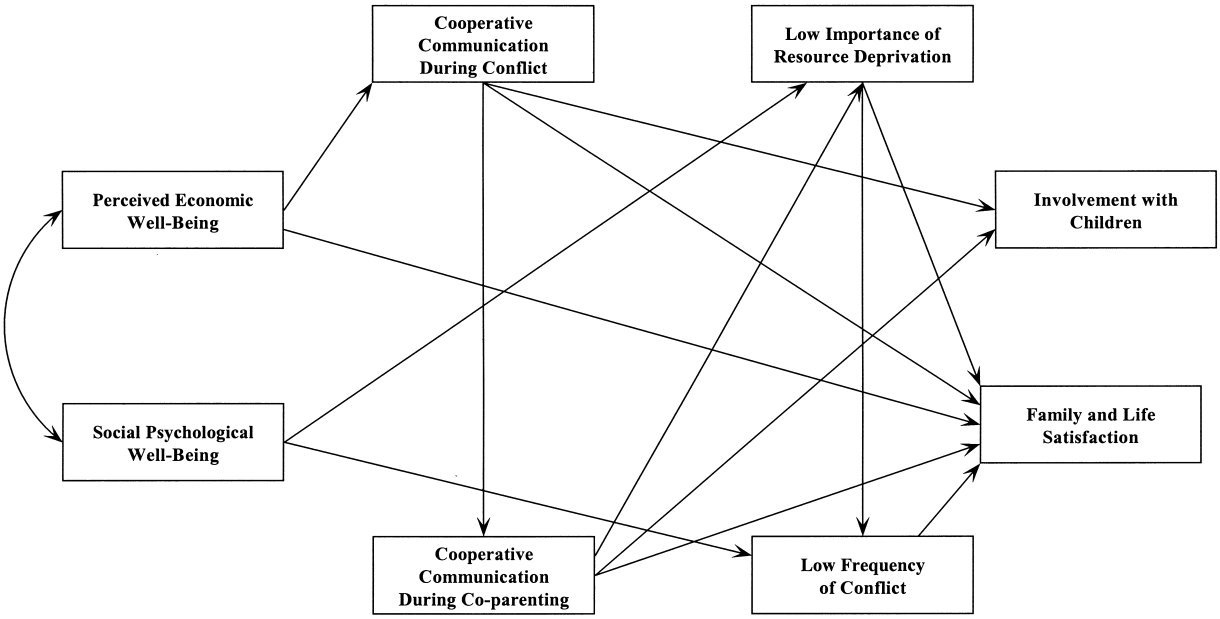

Figure 2. A path diagram of noncustodial fathers' family and life satisfaction. Adapted from "Understanding Noncustodial Fathers' Family and Life Satisfaction From Resource Theory Perspective," by K. D. Rettig, R. D. Leichtentritt, and L. M. Stanton, 1999, *Journal of Family Issues, 20,* Figure 2. Copyright 1999 by Sage Publications. Adapted with permission of Sage Publications, Inc.

els can be generated by the delete 1–add 1 rule. An investigator applying this rule presumes that the substitution of a random path in D* for a theoretically implied path will decrease the model's ability to account for the observed pattern of correlations.

VERIPATH was used to fit the example path analysis model with a subset of correlations from Table 3. The 13 path coefficients of this model were estimated by solving 13 inhomogeneous polynomials. With eight observed variables, there are $8 \times 7/2 = 28$ nonredundant correlations. The model includes 1 exogenous correlation and 13 path coefficients. Thus, we can compute a measure of the internal consistency of the data (i.e., the correlations) that is conditioned on the proposed causal structure by calculating *RMSr* on the 14 nonused correlations. Doing so yields *RMSr* = .07. As we have stressed repeatedly in this article, however, the *RMSr* of D* is only part of the story. To more completely assess D*'s badness of fit, we compare $RMSr_{D^*}$ to that obtained from the class of alternative models that are generated by the delete 1–add 1 rule. For the model displayed in Figure 2, application of this rule resulted in the generation of 233 identified corrupted models. The mean *RMSr* of these models was .10 (*SD* = .02; 20% trimmed mean = .09).

What have we learned after fitting 234 models? Do our findings provide strong support for D*, or have we failed to corroborate the model? This is a difficult question that we are unable to answer given our scanty knowledge of the substantive research domain. Researchers more familiar with resource theory and the measures used in this study are in a better position to address this question and to determine the number of corrupted models that $\{D_T\}$ tolerates. For the data reported in Table 3, D* produced an *RMSr* that was smaller than 89% of those produced by the (233) corrupted models. Stated otherwise, 11% of the corrupted models produced an *RMSr* that was smaller than or equal to that calculated from D*.

Ideally, prior to fitting a model, the investigator considers all paths in D* to determine which paths, if any, can be eliminated or added without jeopardizing the core tenets of theory T (Lakatos, 1970). Notice that this viewpoint recognizes that some paths in a model are more central to a theory's edifice than are other paths. The totality of diagrams (path models) that are tolerated by T compose the set $\{D_T\}$. We propose that an investigator's theory receives greatest corroboration when D* and the models in $\{D_T\}$ are among those with the smallest badness of fit as quantified by *RMSr*. Our rudimentary knowledge of the theory behind the present path model constrains our ability to specify the members of $\{D_T\}$ for this example. Nevertheless, we have shown that it is still possible to compare the relative fit of D* to a large class of structurally close corrupted models. Although they are not ideal, these model comparisons may be the best that an investigator can do in the face of incomplete knowledge or weak theory. The severity of our test is directly proportional to our willingness and ability to specify the models in $\{D_T\}$. Thus, to obtain the strongest corroboration for T, researchers should delineate the core and peripheral causal relations in their models (i.e., define the set $\{D_T\}$). We do not wish to sound perfectionistically hard on researchers, holding them to impossible demands of specificity. On the contrary, investigators have leeway to say as much or as little about what T motivates, allows, and forbids. However, there is a price to pay for being too vague about the tolerated substitutions in the diagram or defining $\{D_T\}$ too broadly. If D* is the sole element of $\{D_T\}$, the researcher who "wins" wins big, but it is at the risk of T being slain because of theoretical intolerance. When the researcher takes less risk, by permitting many substitute arrows into D*, and "wins" (i.e., if D* and $\{D_T\}$ are among the models with the smallest *RMSr*), T gets less credit.

One of the benefits of our approach to model evaluation is that researchers can inspect VERIPATH output to determine how many of the corrupted models are statistically equivalent to D* (Lee & Hershberger, 1990; MacCallum et al., 1993). Returning to our example, without background knowledge of the research domain, it is pointless to argue whether 89% is a cogent indication of model fit or the last nail in the model's coffin. Quantitative measures of model fit, like all numbers, are not self-interpreting; they must be embedded in an interpretive text that takes into consideration background knowledge and a model's auxiliary assumptions.

Design Constraints and Alternative Models

In our final illustration, we demonstrate how design constraints can restrict the range of corrupted models that are generated by the delete 1–add 1 rule. The most common type of design constraint is due to time. In true experiments, objects are assigned to treatment or control groups via randomization prior to the assessment of experimental effects. Thus, randomization can also be conceptualized as a design constraint that is due to time (cf. Holland, 1986).

By definition, cause–effect relations unfold over

time. Thus, unless we are prepared to reverse "time's arrow," we should not allow the delete 1–add 1 rule to violate the time ordering of a causal structure. We can illustrate this notion by considering a model with time-ordered observations.

The model we examine was developed Rodgers and Maranto (1989) to elucidate the causal factors underlying publication productivity. More bluntly, the authors wished to determine why some academicians publish whereas others seemingly perish. To address this question, they mailed questionnaires to a probability sample of 932 members of the American Psychological Association. Two hundred forty-four completed questionnaires were returned. However, because of missing data and various exclusion criteria (e.g., an individual did not have a PhD in psychology), only 162 questionnaires (from 86 men and 76 women) were included in the final sample.

Although the authors considered several models in their study, we focus only on their most comprehensive and preferred model. This model includes seven variables of academic training and achievement, some of which are composites of other variables that we do not examine. The seven variables, with their abbreviated names and intercorrelations (as reported by Rodgers & Maranto, 1989) are given in Table 4. Figure 3 displays a path diagram of the theory under consideration (as before, we have omitted paths from the residual variances to avoid diagram clutter).

The variables in Figure 3 are defined as follows: sex (gender), ab (ability, intellectual resources), gpq (graduate program quality), pre (quality-weighted publications before PhD), qfj (quality of first academic job), pubs (number of publications), and cites (number of citations). A glance at Figure 3 and a moment's reflection about the variables reveals that there is a nonignorable structure to the data that is due to time. For example, an individual's gender is determined prior to attending graduate school. A first academic job occurs after graduation from a university. Articles are cited only after they are written.

Table 4
Correlations Among Seven Determinants of Publication Productivity

Variable	1	2	3	4	5	6	7
1. ab	—						
2. gpq	.62	—					
3. pre	.25	.09	—				
4. qfj	.16	.28	.07	—			
5. sex	−.10	.00	.03	.10	—		
6. cites	.29	.25	.34	.37	.13	—	
7. pubs	.18	.15	.19	.41	.43	.55	—

Note. Adapted from "Causal Models of Publishing Productivity in Psychology," by R. C. Rodgers and C. L. Maranto, 1989, *Journal of Applied Psychology, 74,* p. 643. Copyright 1989 by the American Psychological Association. Adapted by permission of the author. ab = intellectual resources; gpq = graduate program quality; pre = quality-weighted publications before PhD; qfj = quality of first academic job; sex = gender; cites = number of citations; pubs = number of publications 6 years post PhD.

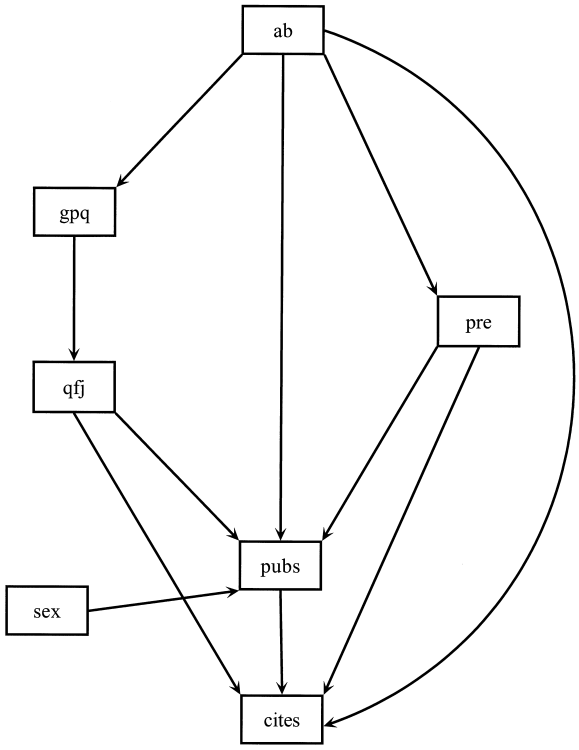

Figure 3. The Rodgers and Maranto (1989) model of publication productivity in academic psychologists. ab = intellectual resources; gpq = graduate program quality; pre = quality-weighted publications before PhD; qfj = quality of first academic job; pubs = number of publications 6 years post PhD; sex = gender; cites = number of citations.

In a recent book on causation, prediction, and automatic model generation, Spirtes, Glymour, and Scheines (2000) also used the Rodgers and Maranto (1989) data in one of their examples. These authors proposed that the seven variables of academic training and achievement could be logically ordered within six time waves: Time 1, sex (gender); Time 2, ab (intellectual resources); Time 3, gpq and pre (graduate program quality and quality-weighted publications before PhD); Time 4, qfj (quality of first academic job); Time 5, pubs (number of publications); and Time 6, cites (number of citations).

For the sake of illustration, we assume that the

aforementioned order is correct and we honor the implied constraints when generating corrupted models by the delete 1–add 1 rule. For example, we do not consider models with the path cites → pubs. To assess the verisimilitude of the authors' theory (represented in Figure 3), we used VERIPATH to test D* and to automatically generate and test the class of structurally close corrupted models.

It is noteworthy that if VERIPATH ignores the time ordering in these data, then the delete 1–add 1 rule generates 128 path models. However, if the time ordering is honored, only 87 path models are generated. Thus, by logical considerations alone, we have effectively reduced the class of corrupted models that are central to the combinatorics of our corroboration procedure. In the present example, numerous statistically equivalent structures were banished from consideration. As a general rule, we prefer the class of corrupted models to be as large as possible—but no larger than what is dictated by the design constraints of our theory.

So how did the model fare in terms of close fit rather than exact fit? We believe that it fared pretty well. The *RMSr* was only .055, a value smaller than 93% of those produced in the 87 alternative models. We view this percentile as a nonparametric estimate of the model's verisimilitude—one that was calculated from simple combinatorics.

Finally, we draw your attention to the (admittedly inefficient) path coefficients that we obtained by solving the polynomials and the (originally reported) unweighted least squares estimates (Rodgers & Maranto, 1989). We call these estimates the *VERIPATH* and *least squares* estimates, respectively. For ease of comparison, we have assembled both sets of estimates in Table 5.

Table 5
Comparison of VERIPATH and Least Squares Estimates

Coefficient	VERIPATH	Least squares
ab → gpq	.62	.62
ab → cites	.13	.13
ab → pre	.25	.25
ab → pubs	.12	.14
sex → pubs	.45	.41
gpq → qfj	.28	.28
qfj → pubs	.39	.34
qfj → cites	.17	.16
pre → pubs	.15	.12
pre → cites	.22	.22
pubs → cites	.42	.42

Note. ab = intellectual resources; gpq = graduate program quality; cites = number of citations; pre = quality-weighted publications before PhD; pubs = number of publications 6 years post PhD; sex = gender; qfj = quality of first academic job.

Perhaps the most important point to draw from Table 5 is that VERIPATH estimates can be very similar to more efficient estimates (in this case, least squares estimates) even though the VERIPATH estimates are often calculated using 30% to 50% fewer pieces of information (i.e., correlations). In the present example, only 57% (12 of the 21) of the nonredundant correlations were used to generate the VERIPATH estimates.

Discussion

How does our method avoid Freedman's (1987, 1997) trenchant criticisms of path analysis and structural equation models? What is the difference between our approach and the conventional approach that immunizes ours from Freedman's attack? The answer lies in the sequence of inferences, in our choice of the epistemic path. The difference is not a matter of mathematics, or even of "logic" in the narrow sense (what is deducible from what), but of inductive logic, or of methodology.

In explicating this somewhat subtle and unfamiliar matter, it is necessary to parse the semantics of the tricky word *assumption.* There are several meanings of this term that, in its applications, are importantly different and whose conflation is methodologically dangerous. The generic meaning is clear enough, namely, a statement is "assumed" if it appears as a premise in an argument, narrowly a deductive argument as in a syllogism, and more broadly as whatever is being offered in support of a hypothesis to be proved. Hence, observational facts are often said to be "assumed" when we offer them in support of a scientific theory (broader meaning). When we turn to what are called "assumptions" in applications of mathematical statistics, matters become complicated despite their falling under the above generic meaning. Without claiming sharp distinctions, we can demarcate roughly four sorts of assumptions that satisfy the broad definition as "something that is being used to infer something else"; these differ in terms of one's knowledge situation and one's chosen epistemic path.

First, a main substantive causal theory of interest is "assumed" *arguendo,* in order to derive the factual consequences that, if confirmed by our observations, are taken as corroborators of the theory and, if disconfirmed, as falsifiers of it.

When we scrutinize real cases of theory testing, we always find that the observational test involves other, auxiliary assumptions, which makes both the corroboration and discorroboration of the theory of interest problematic (Meehl, 1997). Looking closer, we see that these auxiliary *conjectures* (a better term here than *assumptions,* because they are not known for certain any more than is the substantive theory of interest) can in turn be sorted into three categories on the basis of our knowledge situation. One category is conjectures that we only hope are true (or nearly so) and for which we have no strong support, either from background knowledge or within the data of our study. We are forced to rely on empirical extrapolation, or extension of a plausible but not highly corroborated theory, or intuition, or clinical experience, or common sense. We do not pluck this auxiliary out of thin air, but we realize that if it is challenged we do not have a solid answer; the auxiliary is not "safely" assumed. For example, we administer a spelling achievement test and find the raw score distribution to be markedly skewed. We consider the obvious explanations (e.g., the students are in an upper-middle-class suburban school, and their IQs and academic achievement run higher than their school grade level; or the test supplier erred in the age or school grade level). We fall back on the rough generalization that achievement tests tend to be nearly normally distributed and thereupon make an appropriate nonlinear transformation of the raw scores in terms of the Gaussian integral. If this plausible but not strongly supported assumption is incorrect, the inferences we draw from observed relations between the transformed scores and other variables or manipulations will be distorted.

A second kind of auxiliary conjecture that is relied on in making statistical calculations and inferences is susceptible of more or less direct statistical test within the data. For example, justifying use of the Pearson product–moment correlation as an adequate descriptive statistic requires that the relation between the variables should be linear. We say "more or less" direct, because whatever test of linearity we use, because real data points will never fall exactly on the fitted straight line, what significance level should we set up for concluding nonlinearity as an arbitrary element? Whatever convention is adopted for this intermediary step, the investigator must keep in mind that failure to detect falsity of the auxiliary conjecture does not guarantee its correctness, so that theoretical inferences drawn from statistical inferences made from tests and estimates predicated on the auxiliary are problematic.

Finally, a largely neglected type of auxiliary conjecture—in the life sciences the most common—is that in which the assumption is neither a pious hope nor directly testable but is tested indirectly, by virtue of its essential role in derivation chains to predictive statements of relations between inferred quantities. In such situations, the initial plausibility of the substantive theory T of interest and of the auxiliary conjecture may not be very different, and competent scientists may differ as to which one deserves more credence. For example, in the period between the end of the Victorian age and World War I, the existence of molecules was problematic to physicists and chemists, as was the nature of the newly discovered X ray. Scientists differed as to which was more probable on the evidence. X-ray crystallography's accurate prediction of the quantitative properties of refraction using various crystals jointly supported both inferences as to the number of molecules in a mole of sodium chloride making up the crystal's lattice, and the conjecture that X rays were electromagnetic waves with wavelengths in the spectral region shorter than ultraviolet. A conventional psychostatistician might complain that some physicists concluded for the reality of molecules (relying on the wave theory of X rays), and others concluded in favor of the electromagnetic theory of X rays (because they already believed in molecules and assumed a certain value for Avogadro's number)—which is being relied on to prove which? The proper answer is, "Both, at once," and without vicious circularity (cf. Cronbach & Meehl, 1955). Only with these distinctions among uses of the word *assumption* in mind can one properly analyze the inferential structure, the epistemic path, in different uses of path analysis.

The strongest kind of assumption is that referred to in Wright's (1921) first paragraph, where we know the causal relations. In that case, the structure (topology, connections) of D* is given, and the path coefficients are literally deduced from the observed correlations, the only source of error being random sampling error. Unfortunately, this delightful knowledge situation rarely obtains in psychology.

If, following Wright's (1921) second paragraph, we treat D* as a conjecture subject to empirical test, then there are patterns of correlations that are incompatible with certain path diagrams no matter what numerical values are assigned to the path coefficients. This leads

to the notion of corroborating a preferred D* by failing to refute it while killing off its competitors, a common procedure of the kind that Freedman (1987, 1997) disapproves. We have seen psychological studies in which a table of a dozen chi-squares is presented, some of which are significant and others not, and the investigator chooses one of the "good ones." Whatever may be said about the purely statistical merits of causal modeling from the epistemological standpoint, this kind of thing has to be recognized epistemologically as a form of *ad hoc*ery. Because the substantive causal theory taken literally is incomplete, and all of the path diagrams taken literally are incorrect, and the conjunction of the seven auxiliary statistical conjectures is almost certainly false, then all of these chi-squares would show an inadequate fit, given sufficient power.

Our method treats the auxiliary conjectures not as assumptions relied on to make some other inference deducible, whereupon doubts concerning them lead to discarding the inferred conclusion. Rather, the preferred diagram D* that T motivates and the broader class of diagrams $\{D_T\}$ that the theory tolerates are explicitly taken to be literally false. Similarly, the conjunction of the seven highly problematic auxiliaries is recognized as being almost certainly false. We are therefore not interested in any so-called exact test refuting these theoretical falsities. What, then, are we trying to get at instead? We are asking whether D* is better than the whole class of competitor diagrams or, failing that, whether a class $\{D_T\}$ (all the diagrams that would be compatible with T) does better than the class $\{-D_T\}$ (all the diagrams that are incompatible with T or at least not motivated by T).

For comparative purposes, it is noteworthy that other methods of model corroboration have taken a very different approach from our own. For instance, the recent work of Judea Pearl (2000) and the TETRAD group (Spirtes et al., 2000) putatively ascertains whether a given structure (path diagram, structural equation model) is literally true or false using a minimally sufficient set of statistical tests of model-implied independence relations (in linear models, "these implications are perfectly encoded by a set of partial correlations—as implied by D-separation rules," J. Pearl, personal communication, July 31, 2001; see Pearl, 2000, pp. 16–17, for a definition of D-separation rules). Contrary to this view, our approach focuses on the "truth-likeness" rather than the literal "truth" of a model.

Verisimilitude in Science and Life

We accept as a fundamental metaconcept the notion of *verisimilitude* (closeness to the truth, truth-likeness). In common life, in business affairs, in courts of law, in evaluating biographies or newspaper stories, laypeople as well as scholars take it for granted that some accounts are more accurate than others. Scientists regularly invoke the idea of verisimilitude in talking about scientific theories. A scientist who says that T_1 is a better description of reality than T_2 is comparing verisimilitudes. A scientist who amends a theory to try to make it better is implicitly using the concept of verisimilitude. Given the concept of verisimilitude, our line of thought is straightforward. We know that all psychological theories are incomplete (and almost all of them contain postulates that are literally false), and we see nothing special about path analysis in this respect. We know that T is incomplete and that the diagrams $\{D_T\}$ derived from it are literally false. The literal falsity of D* and its permissible substitutes in $\{D_T\}$, when combined with the nearly certain falsity of the auxiliary conjectures, guarantees that a conjunction of the diagram and the observed correlations will be mathematically inconsistent.

A weak theory T says that certain causal influences are powerful and that others are not. It goes on to say that some causal influences are middling and may be large enough to run ahead of some of those we have classified as powerful. It also says there is a large class of other diagrams which T, despite its looseness, will not tolerate. The degree of incoherence in the conjunction of $\{D_T\}$ and the observed correlations is our epistemic path to appraising the verisimilitude (not literal truth) of T. In order to give some sort of numerical value to the epistemic risk, we proceed nonparametrically, using the combinatorics of path diagrams.

One's philosophy of science seldom makes a difference in doing science, but when it does, as in this instance, it makes a big difference. Given the undisputed formalism (as per Wright, 1921) and the clear deductive relations within it, the difference between our approach and the conventional approach lies in our philosophy of science. Combining an emphasis on verisimilitude (rather than exact truth) and the necessity for severe tests prescribes an epistemic path through the formalism that is radically different from the usual. Although each step in the derivations is deductive—each set of numbers being entailed by the

preceding set—the inductive logic is different from what is usually practiced. If it were appropriate in the life sciences to rely on Wright's (1921) first paragraph, one could properly say that the path coefficients are deducible from the conjunction of the theoretical model and the observed correlations but our subject matter does not permit us that option. If we treated the auxiliaries as required assumptions, we would deduce the path coefficients from the correlations plus the topology of the diagram—but knowing that both the diagram and the auxiliaries are problematic, we confront a set of alternative inferred diagrams and causal weights from which we must choose on a methodological basis that is not deductive, giving rise to the *ad hoc*ery condemned above.

In our approach, beginning with the unquestioned concession that T is incomplete, its associated D* is incorrect, and the seven statistical auxiliary assumptions are incorrect, our deductions take place within an idealization—perhaps a rather poor one, but within that idealization they are rigorous. This is, in fact, the universal practice in all of the empirical sciences, even physics. If the idealization is bad enough, the verisimilitude very poor, then our method will not "work" in its predictive task of diagram selection. Solving k equations and k unknowns is deductive, and plugging the solved-for path coefficients into the polynomials for the remaining correlations is deductive. We simply bypass the sampling errors to which the correlations are subject, probably the most shocking deviation from convention. The next step in the reasoning is also deductive, relying solely on the combinatorics of the path diagrams. However, the final inferential step, after contemplating the probability number yielded by that combinatorial analysis, is methodological. It is fundamentally the same as all corroboration of empirical theories, namely, inference to the best explanation, choosing that explanation rather than attributing an antecedently unlikely numerical result—antecedent selection of a small subset of path diagrams—to Salmon's (1984) highly improbable coincidence. We are aware that this crucial step used in all empirical scientific reasoning has yet to be adequately explicated by philosophers (witness the continued Bayesian controversy), but we submit there is nothing special about our proposal in that respect. It is the reasoning required in all applications of significance tests and confidence intervals, all successful (or "close") point predictions, and all surprising qualitative observations, so we are taking the principle for granted (Meehl, 1990).

Comparison With Conventional Approaches to Path Model Testing

How does the procedure presented here compare in strengths and weaknesses with conventional methods? It is subject, along with them, to error arising from near-certain falsity of the auxiliary conjectures. As with other methods, it suffers inaccuracy from random sampling error in the measured individuals, to be ameliorated by insisting on sufficiently large samples (without which one should not be conducting a path analysis in the first place). Selecting (or, more commonly, possessing only) a deficient list of variables is a problem in all path analysis. There is no easy solution—certainly no mechanical rule can exist—to deal with lack of scientific wisdom, lucky intuition, or availability of appropriate indicators. This problem occurs also in experimental research, but in correlational studies it is usually more dangerous.

Appraising a theory by correlational data involves two inferential transitions. First, we estimate population values (parameters) from sample statistics. This transition is subject to random sampling error, which we try to control by statistical procedures based on the theory of probability. Thanks to the mathematical statistician, we possess powerful algorithms for doing this. Second, given our statistical inferences, we make a transition to a substantive (causal or compositional) theory. It is imperative to distinguish this epistemic transition from the purely statistical one. The causal relations are from T to H^* (hypothesis) to a sample statistic s. The inferential relations are in the reverse direction. A mathematically justified confidence in the $s \rightarrow H^*$ transition tells us nothing about the strength of the $H^* \rightarrow$ T movement. Focusing all attention on the former, as is done in statistics texts, misleads as to its epistemic importance. Thesis: The more important source of error in empirical science is the $H^* \rightarrow$ T movement, not the $s \rightarrow H^*$ movement (Meehl, 1990, Figure 2, p. 116). Error in the $s \rightarrow H^*$ inference can be controlled by sample size, reliable measurement, and logical design. We can make its malignant influence negligible in comparison with the dangers in the $H^* \rightarrow$ T inferential step. The latter dangers we aim to reduce by making risky predictions instead of ad hoc adjustments. Logicians have not provided an algorithm leading to a numerical value for this and may consider it impossible. But that severe tests—observation results whose successful prediction is improbable absent the theory—are a basic feature of the developed sciences is not in dispute.

There are two sources of numerical error in our

method that prima facie speak in its disfavor. One—not unique to our approach—is our treatment of small path coefficients as negligible (literally as if they were zero, which they are surely not). Our justification for this is that an admittedly imprecise path coefficient of, say, $c_{ij} = .05$ rarely confirms or refutes an interesting theory in sociology, political science, or behavior genetics. The second source of error is the use of a minority of correlations to infer the path coefficients. Conventional statistical thinking considers it foolish not to use all of the available information. In fact, we are using "all of the information," but in a different way than statisticians usually think of doing it. Why is it rational for us to proceed as we do? In empirical science, portions of the total quantitative information may be used in different ways, serving different epistemological purposes. We begin by emphasizing the quasi-certain inaccuracy of the inferred causal weights, whatever method is used and however much of the information is used. In stressing that obvious truth, we do not argue that increments in unavoidable existing imprecision are (in general) acceptable, absent countervailing considerations. We are mindful of the sententious advice given by the sergeant major to French Foreign Legion recruit John Smith: "When things are bad, do not make them worse, for they will be quite bad enough" (Wren, 1925). However, there is a countervailing consideration. We accept an increment in imprecision (which, we conjecture, is relatively small compared with the imprecision generated by the unavoidable error sources) for a weightier purpose, namely, saving out a mass of data to be predicted from the inferred weights. We hold that numerical precision (somewhat illusory) is less important than the epistemic leverage provided by substituting risky numerical prediction for the conventional ad hoc adjustment.

This attitude, stemming from philosophy of science rather than mathematical statistics, seems more warranted given our realistic statement of a modest epistemic aim for path analysis in the behavioral sciences: to ascertain, with some confidence, which causal influences are strong and which are weak. We do not expect enthusiastic approval from the conventional statistician, whose stock in trade is, appropriately, trying to achieve accurate numerical estimates of population values from samples.

Does the addition of these particular error sources to the aforementioned error sources common to all path analyses vitiate our method? Given our modest epistemic aim of distinguishing the strong causal influences from the weak, vitiation would mean here that numerous malestimates of causal weights will prevent the theory-preferred diagrams from performing well in the predictive phase, and which diagrams outperform others will degenerate into a matter of luck. We doubt that any general analytic derivation as to the size of this sort of malign influence is possible, but, fortunately, the matter need not be left to conjecture. The final empirical step, resting on no dubious assumptions, relies on mere combinatorics of path diagrams. If the error sources were vitiating, then $\{D_T\}$ would not show predictive superiority (except by a low-probability accident). Because it succeeds, the hypothesis of such vitiating error is strongly refuted, *modus tollens,* the standard move in all statistical inference testing. It is essential to grasp the simple logic of this argument, because it is the factual reply to worries about what might conceivably distort results.

References

Aigner, D. J., Hsiao, C., Kapteyn, A., & Wansbeek, T. (1984). Latent variable models in econometrics. In Z. Griliches & M. D. Intriligator (Eds.), *Handbook of econometrics* (Vol. 2, pp. 1321–1393). Amsterdam: North-Holland.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411–423.

Arbuckle, J. L. (1995). Amos for Windows. Analysis of moment structures (Version 3.5) [Computer software]. Chicago: SmallWaters.

Austin, J. T., & Wolfe, D. (1991). Theoretical and technical contributions to structural equation modeling: An updated annotated bibliography. *Structural Equation Modeling, 3,* 105–175.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology, 31,* 419–456.

Bentler, P. M. (1995). *EQS Structural Equations Program manual.* Encino, CA: Multivariate Software.

Berry, W. D. (1984). *Nonrecursive causal models.* Newbury Park, CA: Sage.

Bielby, W. T., & Hauser, R. M. (1977). Structural equation models. *Annual Review of Sociology, 3,* 137–161.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Breckler, S. J. (1990). Applications of covariance structure

modeling in psychology: Cause for concern? *Psychological Bulletin, 107,* 260–273.

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.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research, 18,* 115–126.

Clogg, C. C., & Haritou, A. (1997). The regression method of causal inference and a dilemma confronting this method. In V. R. McKim & S. P. Turner (Eds.), *Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences* (pp. 83–112). Notre Dame, IN: University of Notre Dame Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105,* 317–327.

Cudeck, R., & Henley, S. J. (1991). Model selection in covariance structure analysis and the "problem" of sample size: A clarification. *Psychological Bulletin, 109,* 512–519.

Duncan, O. D. (1975). *Introduction to structural equation models.* New York: Academic Press.

Epstein, R. J. (1987). *A history of econometrics.* Amsterdam: Elsevier.

Foa, E. B., & Foa, U. G. (1980). Resource theory: Interpersonal behavior as exchange. In K. J. Gergen, M. S. Greenberg, & R. H. Willis (Eds.), *Social exchange and advances in theory and research* (pp. 77–94). New York: Plenum.

Freedman, D. A. (1987). As others see us: A case study in path analysis. *Journal of Educational Statistics, 12,* 101–128.

Freedman, D. A. (1997). From association to causation via regression. In V. R. McKim & S. P. Turner (Eds.), *Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences* (pp. 113–161, 177–182). Notre Dame, IN: University of Notre Dame Press.

Holland, P. (1986). Causal inference, path analysis, and recursive structural equations models. In C. Clogg (Ed.), *Sociological methodology* (pp. 449–484). Washington, DC: American Sociological Association.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model-misspecification. *Psychological Methods, 3,* 424–453.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models and data. Studying organizations: Innovations in methodology* (Vol. 1). Beverly Hills, CA: Sage.

Jöreskog, K. G., & Sörbom, D. (2000). *LISREL 8.30 user's reference guide* [Computer software manual]. Chicago: Scientific Software.

Karlin, S., Cameron, E. C., & Chakraborty, R. (1983). Path analysis in genetic epidemiology: A critique. *Journal of Human Genetics, 35,* 695–732.

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin, 96,* 201–210.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). New York: Cambridge University Press.

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

Li, F., Harmer, P., Duncan, T. E., Duncan, S. C., Acock, A., & Boles, S. (1998). Approaches to testing interaction effects using structural equation modeling methodology. *Multivariate Behavioral Research, 33,* 1–39.

Linhart, H., & Zucchini, W. (1986). *Model selection.* New York: Wiley.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51,* 201–226.

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.

Mayo, D. G. (1991). Novel evidence and severe tests. *Philosophy of Science, 58,* 574–585.

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

McKim, V. R., & Turner, S. P. (1997). *Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences.* Notre Dame, IN: University of Notre Dame Press.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant using it. *Psychological Inquiry, 1,* 108–141, 173–180.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals

and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393–425). Mahwah, NJ: Erlbaum.

Niles, H. E. (1922). Correlation, causation and Wright's theory of "path coefficients." *Genetics, 7,* 258–273.

Niles, H. E. (1923). The method of path coefficients—An answer to Wright. *Genetics, 8,* 256–260.

Olsson, U. H., Troye, S. V., & Howell, R. D. (1999). Theoretic fit and empirical fit: The performance of maximum likelihood versus generalized least squares estimation in structural equation models. *Multivariate Behavioral Research, 34,* 31–58.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge, MA: Cambridge University Press.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed). New York: Harcourt Brace.

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

Popper, K. (1977). *The logic of scientific discovery.* New York: Basic Books. (Original work published 1959)

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.

Rodgers, R. C., & Maranto, C. L. (1989). Causal models of publishing productivity in psychology. *Journal of Applied Psychology, 74,* 636–649.

Rogosa, D. (1987). Causal models do not support scientific conclusions: A comment in support of Freedman. *Journal of Educational Statistics, 12,* 185–195.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world.* Princeton, NJ: Princeton University Press.

Shipley, B. (2000). *Cause and correlation in biology: A user's guide to path analysis, structural equations and causal inference.* Cambridge, MA: Cambridge University Press.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search.* Cambridge, MA: MIT Press.

Wolfram, S. (1999). *The Mathematica book* (4th ed.) [Computer software]. Champaign, IL: Wolfram Media; and New York: Cambridge University Press.

Wren, P. C. (1925). *Beau Geste.* New York: Stokes.

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

Wright, S. (1923). The theory of path coefficients—A reply to Niles's criticism. *Genetics, 8,* 239–255.

Received March 15, 2001
Revision received December 13, 2001
Accepted January 31, 2002 ■