Simulation Studies of the Reliability of Computer Aided Model Specification Using the TETRAD, EQS and LISREL Programs

Peter Spirtes, Richard Scheines and Clark Glymour[1]

Department of Philosophy
Carnegie Mellon University
Pittsburgh, Pa. 15213

Abstract

TETRAD n, a fully automated successor to the TETRAD program, is intended to aid in the respecification of underspecified linear causal models, or structural equation models. The performance of TETRAD II is compared with the automatic respecification procedures in the EQS and LISREL VI programs using 360 simulated data sets from nine different linear models containing "latent" or unmeasured variables. For these cases, we find that the TETRAD II program, which uses graph algorithms and heuristic search techniques, is significantly more reliable than either EQS or LISREL VI, which use numerical algorithms and beam search techniques. A detailed analysis of the reasons for these differences is offered. Contrary to those who dismiss automated search techniques as unreliable "ransacking" or "data mining," TETRAD II provides correct information about the true model for 93.8% of the large sample data sets. The need for further simulation tests and the prospects for the development of automated techniques to aid in the initial specification of causal models for nonexperimental data are discussed. An appendix illustrates the application of the published TETRAD program to one of the data sets in this study.

[J]The research reported in this paper was supported by the Office of Naval Research under Contract number N00014-88-K-0194.

1. Introduction

Linear causal models, or structural equation models, are used by market researchers, educational researchers, policy evaluators, social scientists, psychologists, medical researchers, biologists and others who must deal with nonexperimental or quasi experimental data. Models of this kind, which include factor analytic and path analytic models as special cases, are popular for good reasons. Structural equation models permit researchers to express causal hypotheses as linear equations; there are commercial computer programs for estimating the parameters of the models that result, and the estimates of linear coefficients are naturally interpreted as indicating the relative strength of causal effects. The same computer programs permit the application of statistical hypothesis tests to the models thus estimated. Altogether, linear causal models and the associated statistical procedures packaged in programs such as EQS [2] and LISREL [17] seem an appealing framework for the quantification and testing of causal theories in many domains.

Yet the use of linear causal models has come under attack from many quarters.[2] One complaint is that the use of such models involves substantive assumptions—linearity and multinormality, for example—that are rarely tested, and often do not hold. This is not a compelling objection to structural equation modeling procedures, since, first, nothing prevents researchers from performing more careful data analysis, (and programs have been produced, such as PRELIS, to aid in just that endeavor), and second, such analysis does sometimes show that the assumptions of the linear modeling formalism are met to good approximation.

A second complaint is that the formulation of a structural equation model, and the demonstration that it does not fail a statistical test, give almost no reason to believe that the causal assumptions of the model are true or even approximately true. For, on the one hand, even supposing approximate linearity and normality and other assumptions of the linear modeling technique, given a model that passes a statistical test even on a fairly large sample, there may exist thousands, or even millions, of alternative linear causal models that would meet the same criterion, and that might differ in important respects from the model that a researcher advocates. And, on the other hand, even a model that correctly specifies causal relations will fail a chi-square test of goodness of fit on large samples if there are tiny failures of linearity or normality in the process that generated the data.

The natural response to the second objection is that researchers in an area have substantive theoretical knowledge that forms and justifies their causal claims. The structural equation modeling formalism simply enables them to quantify their hypotheses. In practice, however, this is seldom entirely correct. The background knowledge supplied by common sense, by experimental work, or even by a theoretical tradition, usually affords only a fragmentary specification of causal relations in a study, whether in sociology or econometrics or educational research or in other disciplines. This sort of information typically demands that certain causal

[2] See [1,19,11], for example.

relations be postulated, and may also forbid other causal relations, but leaves a great many possibilities indeterminate. Often those possibilities are exactly the point of the research study. Many of the causal hypotheses of a model are therefore often founded on a prior guess that has no special justification, exactly as the second complaint alleges. Those interested in linear modeling have long been aware of the need for techniques that will respond to this difficulty and will provide reliable guides to the respecification of an initial, possibly incomplete, structural equation model.

In consequence, a number of techniques have been proposed for using sample data to modify an initial model, generally by suggesting further causal connections or correlated errors not included in the initial specification of the model. Insofar as these procedures, whether they consist of analysis of residuals, modifications of fitting functions, or other methods, claim to be reliable guides to truth, they constitute a central part of whatever logical response can be given to the second objection to causal modeling, the objection that says the causal hypotheses in structural equation modeling are unsubstantiated. These procedures in turn are often denounced for "ransacking" or "data mining." That is simply name calling. The most important question germane to procedures for inferring modifications to an initial model using sample data is *how reliable are the procedures as guides to the truth?*

In this paper we will compare three fully automated procedures for modifying an initial model using sample data. The procedures used in the EQS and LISREL VI programs are very similar, and seem to be widely used. They employ standard numerical analysis algorithms to find modifications to a fitting function. The third procedure, TETRAD n, is an experimental, fully automatic version of the TETRAD program [12]. TETRAD and TETRAD II use graph analysis algorithms rather than numerical techniques, and rely on heuristic search techniques characteristic of many artificial intelligence programs.

In our study, forty data sets, twenty with a sample size of 200 and twenty with a sample size of 2,000, were generated by Monte Carlo methods from each of nine different structural equation models. All of these were "latent variable" models, chosen because they involve causal structures that are often thought to arise in social and psychological scientific work. In each case part of the model used to generate the data was omitted and the remainder, together in turn with each of the forty data sets for that model, was given to the LISREL VI and TETRAD II programs; only data with the large sample size was given to EQS. A variety of specification errors are represented in the nine cases. Linear coefficient values used in the true models were generated at random to avoid biasing the tests in favor of one or another of the procedures. In addition, a number of ancillary studies were suggested by the primary studies and bear on the reliability of the three programs.

The LISREL and EQS programs involve an automatic search procedure that, given an initial model and data, produce a *unique* recommendation for revision of the initial model. The TETRAD II program, given an initial model and data, produces a set of alternative revisions. Depending on the data and the initial model, the set of alternatives recommended by TETRAD II may vary in size considerably.[3] In our studies the number of alternatives was typically three or four; sometimes a single revision was suggested, and sometimes more than ten alternatives were offered. We find that, for these cases, the information provided by the TETRAD II program is much more reliable than the information provided either by EQS or by LISREL. More importantly, in response to those who condemn attempts at respecifying models by analysis of the data, for the large sample size the information provided by the TETRAD II program was correct for almost 94% of the 180 data sets.

The following section describes the general framework of the TETRAD program, and the modifications made in the TETRAD II program. Section 3 describes the LISREL and EQS model respecification procedures. Section 4 briefly describes some previous simulation studies, the design of our study, the reasons for that design, and the results of our primary studies. Section 5 describes ancillary studies of the reliability of the three procedures separately and conjointly, and analyzes the results of the primary study. Section 6 states our conclusions as to the present best use of these programs in model respecification, describes further studies of the reliability of the programs that we think ought to be conducted, and addresses the prospects for further developments in automatic model construction. An appendix illustrates how the published version of the TETRAD program may be applied to these problems.

2. The TETRAD Programs

2.1. Structural Equation Models and Graphs

Structural equation models are given by a system of linear equations, for example

$$x_1 = a_1 T_1 + \varepsilon_1$$

$$x_2 = a_2 T_1 + \varepsilon_2$$

$$x_3 = a_3 T_1 + \varepsilon_3$$

$$x_4 = a_4 T_1 + \varepsilon_4$$

$$x_5 = a_5 T_1 + \varepsilon_5,$$

together with a specification of the joint distribution on the random variables. The latter conditions may specify, for example, that the joint distribution is multinormal, that the variances of the exogenous variables are not zero, and that certain of the ε variables are correlated with one

[3] The number of alternatives tends to increase as the number of edges that are in the true model, but not the starting model, increases. The number of alternatives also tends to be larger when the edges in the true model, but not the starting model, link together to form paths.

another. **The** tacit specification **is that unless the equations imply, algebraically, a functional dependence of one variable on another, or unless a correlation of error terms is specified, the variables are assumed in the model to be statistically independent For example, one might** specify **that 6j is correlated with 62 and that *e*^ $*^s$ correlated with e_3.**

Such models are commonly accompanied by graphs that represent causal or other dependencies, for example:

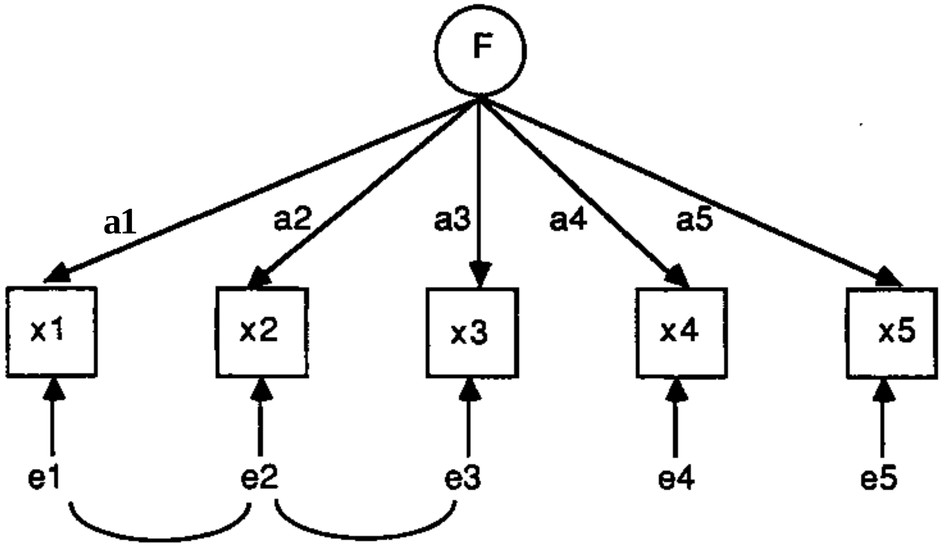


Figure 1:

The undirected lines indicate correlated errors. Important parts of the statistical model can be recovered from the graph alone. The graph encodes the form of the linear equations, and it encodes all of the assumptions of statistical independence that are implicit in the statistical model. The graph does *not* encode the particular numerical values of the linear coefficients, the variances of the independent variables, or the joint distribution family (e.g., multinomial).

The graph is not only a vivid representation of the claims made by a structural equation model; it also determines certain kinds of *statistical constraints*, or *over identifying constraints* that a structural equation model may imply. One such class of constraints consists of *vanishing tetrad differences,* A tetrad difference is just the determinant of a 2 X 2 submatrix of the covariance matrix:

$$\rho_{ij}\rho_{kl} - \rho_{ik}\rho_{jl}$$

One form of constraint on the covariance (or correlation) matrix is obtained by specifying that a tetrad difference vanishes. Such a constraint represents a kind of prediction **a** structural equation model makes about the correlations or covariances in the population it purports to describe. Whether expressed as covariances or correlations, the vanishing tetrad differences implied by a structural equation model are determined entirely by the graph of the model. They do not depend on the variances of the exogenous variables or on the distribution family. In fact, the constraints of this kind that are implied by a model are determined simply by connections in the graph we call *treks. A trek* between two variables Xj and x_k is either an acyclic path in the graph from one variable to the other, or else it is a a pair of acyclic paths from a third variable respectively to each of the variables Xj and x_k.[4] A trek is the graph theoretic representation of either a causal pathway from one variable to another or a common cause acting on two variables. In Model II in Fig. 2, for example, the edges marked a and c form a pair of paths that constitute a trek between the variables j and k, and similarly the edges marked d and b form another trek connecting variables j and k.

A structural equation model may implicitly specify a tetrad constraint in either of two ways. Consider the two models in Fig. 2, (where, for simplicity, we have not drawn the error terms, but the reader should supply them mentally):

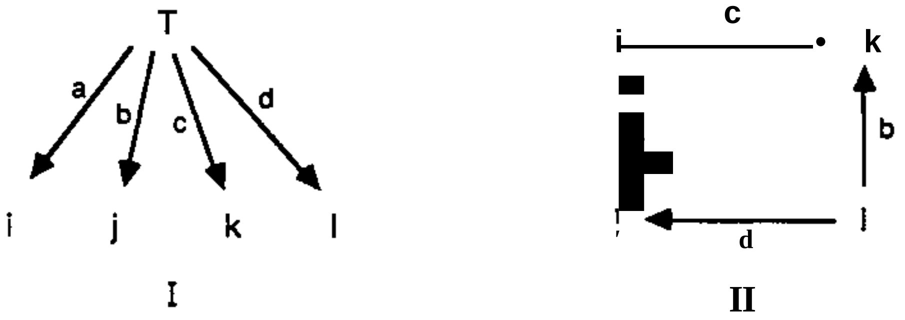


Figure 2: Alternative Causal Structures

In the first case,

[4] We further require that if a trek is a pair of acyclic paths from a third source variable, that the paths not intersect other than in that source variable.

$$\rho_{ij} = abo^2(T)$$

$$\rho_{kl} = cd\,\sigma^2(T)$$

$$\rho_{ik} = aco^2(T)$$

$$\rho_{jl} = bd\,\sigma^2(T)$$

PijPki-PikPji $^{=abcd}$ ^OD-abcd o^4(T) = 0.

In Model I, the tetrad difference vanishes *identically,* since it vanishes for all values of a, b, c, d, and C^2(T).

In the second case,

Pij = a O^2(i)

$$\rho_{kl} = b\,\sigma^2(l)$$

p$_{ik}$ = ca^2(i)

$$\rho_{jl} = d\,\sigma^2(l)$$

Pij Pkl - Pik Pjl = $^{ab\ a\ 2}$ ® o^2*(0*'$^{cd\ a2}$(i) o^2©

In Model n, for particular values of the coefficients, such as a = 2, b = 2, c = 4, and d = 1, py p^ - pfc pjj = 0, and so the tetrad difference vanishes. But this constraint is not *robust* in Model II, because the tetrad difference does not vanish if the non-zero coefficients are varied in that model. For example, in Model II if a = 2, b = 2, c = 4, and d = 2, then py p^ - p^ p$_{j}$j * 0.

When a structural equation model robustly specifies a vanishing tetrad difference, as with Model 1 above, we say the model *implies* the vanishing tetrad difference.

2.2. TETRAD

The TETRAD program uses both methodological and algorithmic ideas to help in the respecification of structural equation models. One methodological idea is that if constraints, such as vanishing tetrad differences, are found to hold based on the sample data, then insofar as possible the correct model should imply those constraints. A second idea is that the correct model should not imply constraints that are not satisfied empirically. A third methodological idea is that, other things equal, simple models, those with more degrees of freedom, are to be preferred to more complex models. These three ideas can be thought of as the Explanatory Principle, the Falsification Principle, and the Simplicity Principle respectively. Unfortunately, the principles often conflict. We have proved (see [12]) that if M and M' contain the same variables, and the edges in model M are a proper subset of the edges in M\ then the vanishing tetrad differences implied by M are a (not necessarily proper) superset of those implied by M\ Suppose that there are some vanishing tetrad differences implied by M but not by M\ and that some of these hold in the population, while others don't. Then M is superior to M' with respect to the Explanation and Simplicity Principles, but inferior to M' with respect to the Falsification Principle. Is M a better model than M'? Any decision based upon these methodological principles in effect must judge when a loss in simplicity and explanatory power is made up for by a gain in reducing the false implications of a model. The TETRAD program provides information relevant to these principles, but leaves their weighting largely in the hands of the user.

One central algorithmic idea in the TETRAD program is that there are well known, fast algorithms for analyzing directed graphs, algorithms that can be modified to determine the set of all vanishing tetrad differences implied by a model.

The TETRAD program accepts as input correlation or covariance data, and the *graph* of a structural equation model. The graph is given to the program simply by specifying a list of paired causes and effects. The program then subjects each possible vanishing tetrad difference among the measured variables to a statistical test, and forms the set, call it H, of all tetrad equations that pass this test. TETRAD then computes the set, call it I, of all of the vanishing tetrad differences implied by the structural equation model given to the program as input. The intersection of H and I, I&H, is the set of vanishing tetrad differences that pass the statistical test and that are also implied by the initial model; it is a collection of witnesses for the truth of the initial model. I-H is the set of implications of the initial model that do not hold when tested on the sample. The set I-H thus constitutes a collection of witnesses against the truth of the initial model. Adding causal connections that create further treks may result in a revised model that has a different set, call it I\ of implications. In general, of course, different additions to the initial model will have different sets of implied tetrad constraints, but no matter how the initial model is extended, the resulting set, I', of implied tetrad constraints is always a subset of I, the constraints implied by the initial model. So by elaborating the initial model, smaller sets F-H may be obtained.

Among all of the possible elaborations of the initial model, which typically number among the thousands, a few will come as close as possible to having F = I&H (see Fig. 3). These few elaborations will be respecified models that (1) imply all of the empirically correct vanishing tetrad differences implied by the initial model, and (2) imply as few empirically incorrect vanishing tetrad differences as is possible given (1).

A second algorithmic idea in the TETRAD program is an automatic, heuristic search procedure that with great reliability finds all of the *treks* that may be simultaneously added to an initial

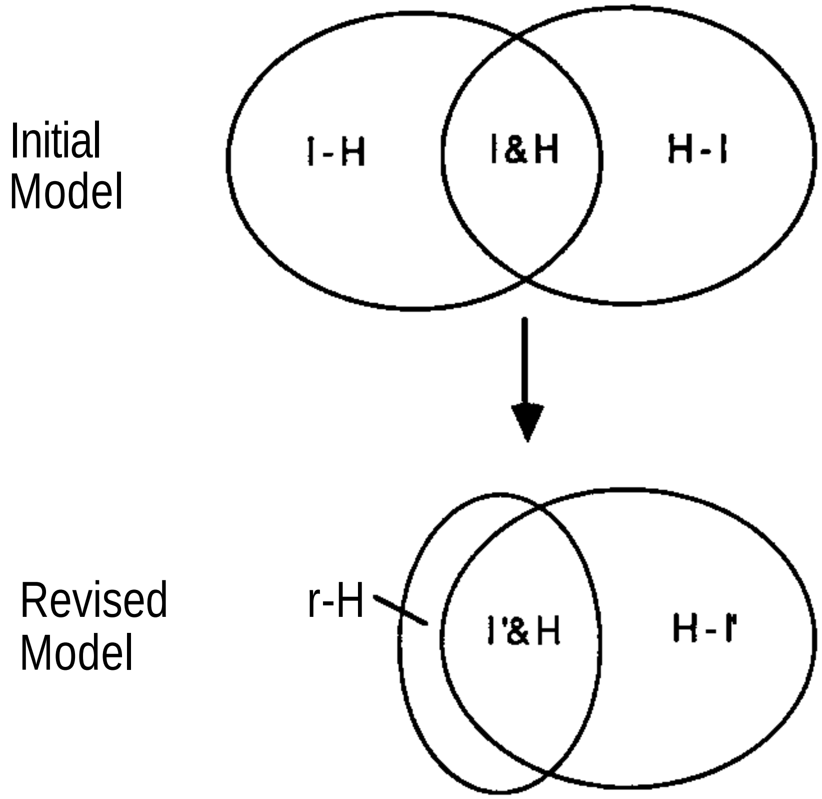


Figure 3: TETRAD's Revision Strategy

latent variable model to form all elaborations satisfying (1) and (2).

The output of the TETRAD program is a list of recommended trek additions parameterized by the significance level of the test applied to the tetrad equations. In addition, the program informs the user about reductions in certain residuals, i.e., about improvements in fit that would be obtained for each causal connection that might be added to the initial model.

The TETRAD program does not produce explicit models as output. Although a set of heuristics are given in *Discovering Causal Structure* for producing explicit causal graphs from the TETRAD output, the procedures require considerable thought and practice to use reliably.

2.3. TETRAD II

TETRAD II is an experimental program that includes a fully automatic respecification search. As with TETRAD, the input is an initial model and a covariance or correlation matrix. The output is a class of alternative best models, a class of alternative second best models, and so on. In each case the model consists of an explicit causal graph rather than a set of treks.

TETRAD II uses the same three methodological principles, Explanation, Falsification, and Simplicity, as does the TETRAD program; it tests each vanishing tetrad difference against the sample data in the same way, and it uses the same fundamental algorithm for computing the vanishing tetrad differences implied by a model. It differs in the procedure used to locate the revisions it recommends. That procedure can be analyzed in two parts, a *scoring function* used to evaluate possible revisions to the initial model, and a *search strategy* used to locate the models with the highest score.

The Scoring Function

We will refer to the scoring function used by TETRAD II as the T-score. The principles described in the previous section are implemented in the T-score, which depends upon two parameters, the significance level and the weight (explained below). For each possible vanishing tetrad difference, t, we calculate the probability of obtaining a tetrad difference as large or larger than the one actually observed, under the assumption that the difference is zero in the population, and that the tetrad differences are normally distributed:[5] we call this the associated probability of t, and denote it by p(t). For a given significance level, if p(t) is larger than the significance level, we say that the vanishing tetrad difference holds in the population; otherwise, we say that the vanishing tetrad difference does not hold in the population. Let $Implied_H$ be the set of vanishing tetrads implied by a given model M that hold in the population, let $Implied\wedge_H$ be the set of vanishing tetrads implied by M that do not hold in the population, let *T* be the score of model M for a given significance level, and let *weight* be a parameter (whose significance is explained below). Then we define

$$T= \underset{tz\ ImpliedJJ}{X}\ pit)- \underset{tt\ Implied\wedge}{X}\ [weight* (l-p(f))) \qquad (1)$$

The first term implements the Explanatory Principle; it gives credit for explaining vanishing residuals that hold in the population. The second term implements the Falsification Principle; it penalizes a model for predicting vanishing residuals that do not hold in the population. The Simplicity Principle is implemented by preferring, among models with identical T-scores, those that have more degrees of freedom.

The weight decides conflicts between the Explanatory and Falsification Principles. It determines the relative importance of explanation versus residual reduction. The lower the weight, the more important explanation is relative to residual reduction. Since a submodel explains at least as many vanishing tetrad residuals as any model containing it, lowering the weight tends to favor

[5] The assumption that the tetrad differences of normal covariates are normally distributed is false, but with large sample sizes holds to a good approximation.

models with fewer edges.

Search

The current search strategy has two parameters. The first is the breadth, which determines how many of the most promising edges are considered to be candidates for addition to the starting model. The second is the depth, which determines the maximum number of edges that can be added to a starting model in the course of the search. However, as explained below, if the results of the search warrant it, the depth is increased and the search is repeated.

The search used in this study begins by computing the score of each one-edge elaboration of the starting model. (We count the addition of a correlated error as a "one-edge" addition to a model.) If the addition of an edge improves the score of the initial model then the edge is added to a list of candidates. When all of the one-edge additions to the initial model have been calculated, the candidate list is ordered from highest to lowest score, and if there are more edges than the specified breadth, the end of the list is discarded.

From this point on, the search proceeds "depth-first": the order in which edges are added is determined by their rank in the candidate list. Search along a branch is cut off either when the addition of an edge fails to improve the score, or when the depth limit is reached. If at the end of a search, the best model has fewer edges than the depth limit, the search is ended. If the best model has the same number of edges as the depth limit, the depth limit is incremented by one, and the search is repeated.

We are not confident that either the scoring function, the search strategy or the criterion for stopping are optimal, and we are experimenting with alternatives. The search does not guarantee finding the model with the highest T-score. For example, it is possible that a model with the highest score contains an edge that is initially removed from the candidate list because there were more than the breadth number of edges with better scores. There is an exponential growth in the time the search takes as the breadth increases.[6]

3. LISREL and EQS

LISREL and EQS are computer programs that provide maximum likelihood estimates of the free parameters in a structural equation model. More precisely, the estimates are chosen to minimize the fitting function

$$F = \log|\Sigma| + tr(S\Sigma^{-1}) - \log|S| - (t)$$

[6] In all cases in the studies reported here, breadth was set at 30 and depth at 3.

where S is the sample covariance matrix, L is the predicted covariance matrix, t is the total number of indicators, and if A is a square matrix then \A | is the determinant of A and Tr(A) is the trace of A. The parameters that minimize the fitting function F also maximize the likelihood of the covariance matrix for the given causal structure.

The EQS program permits the user to enter the structural equations more or less as they would normally be written. The LISREL program, for historical reasons, requires coding the equations of a model in an elaborate and artificial matrix formalism.[7] Each program requires a set of initial values for the parameters, values that are altered in the course of minimizing the fitting function. Both programs give diagnostic information pertinent to the fit of a model, and both contain an automatic procedure for elaborating an initial model. Those procedures are very similar in the two programs.[8] As with TETRAD n, the respecification procedures in EQS and in LISREL VI can be analyzed as a scoring function and a search procedure.

3.1. Scoring

After estimating the parameters in a given model M, LISREL VI calculates the probability of obtaining a discrepancy between the observed and predicted covariance matrices as large or larger than the discrepancy actually observed, under the assumption that M is true. We will call this quantity the L-score. This probability can be used to perform a statistical test of M, where the null hypothesis is that M is true. If the probability is greater than the chosen significance level, the null hypothesis is accepted, and the discrepancy is attributed to sample eiTor, if the probability is less than the significance level, the null hypothesis is rejected, and the discrepancy is attributed to the falsity of M.

It is also possible to do statistical tests on a series of nested models, that is, on a series of models $M_{1?}...,M_k$ in which all models are the same save that for all models Mj in the sequence, the free parameters of Mj are a subset of the free parameters of Mj_{+1} The *difference* between the chi-square values of two nested models also has a chi-square distribution, with degrees of freedom equal to the difference between the degrees of freedom of the two nested models.

For a variety of reasons Joreskog and Sorbom have recommended that L-scores *not* be considered part of a classical hypothesis test. They recommend instead that it be considered a measure of goodness of fit relative to the number of degrees of freedom. Just what this means is unclear. For our purposes, however, it is enough to observe that the L-scores of models are used to evaluate models in LISREL's automated respecification procedure. EQS can be regarded as

[7] An associated program, SIMPLIS, permits the straightforward entry of a model. The causal structures that the program will process are, however, so restricted that we have not found the program of any use. Unfortunately, the restrictions are poorly documented.

[8] EQS also contains a procedure using the Wald test for *removing* causal connections from an initial model. We have not tested the reliability of this procedure.

having the same scoring function.

3.2. Search

The input to LISREL VI's search procedure is a starting model specifying the values of the fixed parameters, starting values for the free parameters, a sample covariance matrix, a list of those parameters that are not to be freed under any circumstance, and a significance level. The search is guided by the "modification indices"; the modification index of a given fixed parameter provides a lower bound on the decrease in the chi-square obtained if that parameter is freed and all previously estimated parameters are kept at their previously estimated values.[9] (Note that if the coefficient for variable A in the linear equation for B is fixed at zero, then freeing that coefficient amounts to adding an edge from A to B to the graph of the model.) LISREL VI calculates the modification indices for all of the fixed parameters[10] in the starting model. The fixed parameter with the largest modification index is freed, and the model is re-estimated. .If the difference in the chi-squares of the starting model and the elaborated model is significant, the parameter is freed, the elaborated model is now the starting model, and the process is repeated. When freeing the fixed parameter with the highest modification index does not result in a model with a chi-square significantly different from the starting model, the parameter is not freed and the search ends. The EQS program involves a different statistical procedure, but uses it in the same way.[11]

4. The Primary Studies

The essential questions about any discovery procedure concern its reliability in the circumstances in which it is meant to be applied. The documents describing the TETRAD, EQS and LISREL programs provide any number of applications of their respective procedures to empirical data sets. Unfortunately, such applications are of little use in judging the *reliability* of the procedures. The reason is obvious: in empirical cases we don't know what the true model is, so we can't judge whether the procedures have found it. We can judge whether the procedures turn up something that isn't absurd, and we can judge whether the procedures find models that

[9] LISREL outputs a number of other measures that could be used to suggest modifications to a starting model, but these are not used in the automatic search. See [7].

[10] As long as they are not in the list of parameters not to be freed.

[11] EQS allows the user to specify several different types of searches. In the search that we used, EQS performs univariate Lagrange Multiplier tests to determine the approximate separate effects on the chi-square value of freeing each fixed parameter in a set specified by the user. The program then repeats the procedure on the model obtained by freeing the parameter that most decreases the chi-square value. It stops when no freed parameter provides a significant decrease in chi-square. It should be noted that both LISREL VI and EQS are by now quite complicated programs, with less than optimal documentation. An understanding of their flexibility can only be obtained through experimentation with the programs. Since the EQS automatic search program is quite new, we are less confident than in the case of LISREL VI that we have fully exploited its resources.

pass statistical tests, but neither of these features is of central importance. What is of central importance is whether or not the automated model revision procedures find the truth. Empirical tests can sometimes be obtained of models produced by the automatic searches, and they may provide some evidence of the reliability of the procedures. In general, however, such tests have rarely been obtained, and they cannot be relied upon since one does not know whether the initial model given to the program is empirically correct It is possible to do mathematical analyses of the power of a discovery procedure to distinguish or identify alternative structures in the limit, as the sample size grows without bound. Some results of this kind pertinent to TETRAD methods are implicit in *Discovering Causal Structure*, and we have subsequently proved a number of other limiting properties of the TETRAD procedures. Limit results do not, however, address the behavior of automated discovery procedures on samples of realistic sizes, and it is that which ought most to concern empirical researchers.

The best solution available is to apply Monte Carlo methods to assess the reliability of model respecification procedures. Using a random number generator, data for a specified sample size can be generated from a specified structural equation model. Part of the model used to generate the data is then given to the procedures, and we see with what reliability the procedures can recover information about the missing parts of the models used to generate the data. In this way, the reliability of the procedures can be tested in nearly ideal circumstances: the true structural equation model is known, the sampling is random, and distribution assumptions are satisfied to a good approximation. The manuals documenting the LISREL and EQS programs contain no tests of their model respecification procedures on simulated data. Three such tests are reported in *Discovering Causal Structure.*

4.1. Previous Simulation Studies

There have been a number of studies of the effectiveness of various features of LISREL.

Fornell and Larcker [9] performed a study that was intended to demonstrate problems LISREL IV has in evaluating (not correcting) models with large parameter values. They claimed that causal models with small linear coefficients have smaller chi-square values than do models with large linear coefficients. They gave as input to LISREL "true" models and their associated correlation matrices. However, they did not generate their correlation matrices by Monte Carlo methods; in many cases their matrices were altered by hand from other correlation matrices. Their results are misleading since in some cases the associated correlation matrix they use is extremely improbable given the model they assume.

MacCallum [21] studied the reliability of LISREL's automatic search on "structural models,"[12] and how it is affected by the sample size, the starting model, and by prior restrictions. Not

[12] The "structural" part of a causal model includes causal connections among latent variables only. The "measurement" part of a model includes all other causal connections.

surprisingly, he found that the reliability of the search decreased as the sample size decreased, the number of specification errors in the starting model increased, and the number of prior restrictions on the search decreased. MacCallum generated Monte Carlo data for mo different "true" models. For each true model, he gave LISREL the covariance matrix generated by that model together with a variety of starting models that varied both by the number of edges they were missing and by the inclusion of edges that were not present in the true model. In MacCallum's study the parameter values of the "true" models were atypical, however. Each coefficient connecting latent variables was given one of only two different values, and each coefficient connecting a latent with a measured variable was also given one of only two values. These restrictions had the effect of imposing constraints not implied by the causal structure of the models. Although his study was systematic and thorough, it was a test only of LISREL's reliability in correcting errors in the structural pan of a causal model. Localizing the error in this way drastically reduced the number of possible corrections LISREL had to consider. In one of his cases, for example, there were 297 possible corrections to the full causal model, but only 3 possible corrections to the structural part.

Costner and Herting [7] did an extensive and useful study of the ability of LISREL V to correct mis-specifications through manual examination of the modification indices and standardized residuals. Although they did not run the automatic search procedure in LISREL VI, their descriptions of the modification indices allow us to infer to a large extent what the automatic search procedure would have done. They considered a wide variety of different kinds of mis-specifications. While their parameter values were not randomly generated, in some cases they did systematically vary the magnitudes of the parameters. The results of their study, which we discuss below, were generally corroborated by our own. Their study did, however, have important limitations. They were unable to detect the effects of sample size on the reliability of LISREL, since they did not use Monte Carlo simulations to generate their covariance matrices. Instead, LISREL was given the population covariance matrices. There were edges that Costner and Herting wished to, but could not, add to their starting model, because of the elaborate restrictions that LISREL places on additions to the starting model. (For example, an edge from one indicator to another indicator cannot be added.) There are methods of re-describing models in the LISREL formalism which would have allowed Costner and Herting to determine the effects of adding these "forbidden" edges to their starting models, but they were not used.[13] Finally, since Costner and Herting used the population covariance matrices (which produce a chi-square of zero if the correct model is input), and did not run the automatic search available in LISREL VI, they were unable to determine how often LISREL would mistakenly add too many edges to the starting model.

[13] Which is certainly no fault of Costner and Herting, since the recodings required are not documented in any of the LISREL manuals.

4.2. Requirements for Comparative Simulation Studies

An ideal study of the reliability of the automatic respecification procedures in TETRAD n, LISREL VI, and EQS, under conditions where the general structural equation modeling assumptions arc met, would examine the effects of the following factors by varying them independently.

- The causal structure of the true model.
- The magnitudes and signs of the parameters of the true model.
- How the starting model is mis-specified-that is, the structure of omitted dependencies.
- The sample size.

In addition, an ideal study should:

- Compare fully algorithmic procedures, rather than procedures that require judgment on the part of the user. Procedures that require judgment can only adequately be tested by carefully blinding the user to the true model; further, results obtained by one user may not transfer to other users. With fully algorithmic procedures, neither of these problems arises.
- Examine causal structures that are of a kind postulated in empirical research, or that there are substantive reasons to think occur in real domains.
- Generate coefficients in the models randomly. Costner and Herting showed that the size of the parameters affects LISREL's performance. Further, the reliability of TETRAD II depends on whether vanishing tetrad differences hold in a sample because of the particular numerical values of the coefficients rather than because of the causal structure, and it is important not to bias the study either for or against this possibility.
- Ensure in so far as possible that all programs compared must search the same space of alternative models. For example, LISREL VI cannot consider any models that elaborate an initial model by postulating a direct effect from an endogenous variable to an exogenous variable in the initial model. For a fair comparison, EQS and TETRAD II must therefore be relieved of the responsibility of considering such models.

4.3. Study Design

Selection of Causal Graphs:

The nine causal structures studied are illustrated in Figure 5, 6, and 7 below. (For simplicity of depiction we have omitted uncorrelated error terms in the figures, but such terms were included in the linear models.). The heavier directed or undirected lines in each figure represent relationships that were included in the model used to generate simulated data, but were omitted from the models given to the three programs; i.e., they represent the dependencies that the programs were to attempt to recover. The starting models are shown in Figure 4.

The models studied include a model with five measured variables and one latent factor, seven models each with eight measured variables and two latent variables, and one model with three latent variables and eight measured variables.

One factor models commonly arise in psychometric and personality studies (see [18]); two latent factor models are common in longitudinal studies in which the same measures are taken at different times (see [23]), and also arise in psychometric studies; the triangular arrangement of latent variables is a typical geometry (see [28]).

Selection ofConnections to be Recovered:

The connections to be recovered include

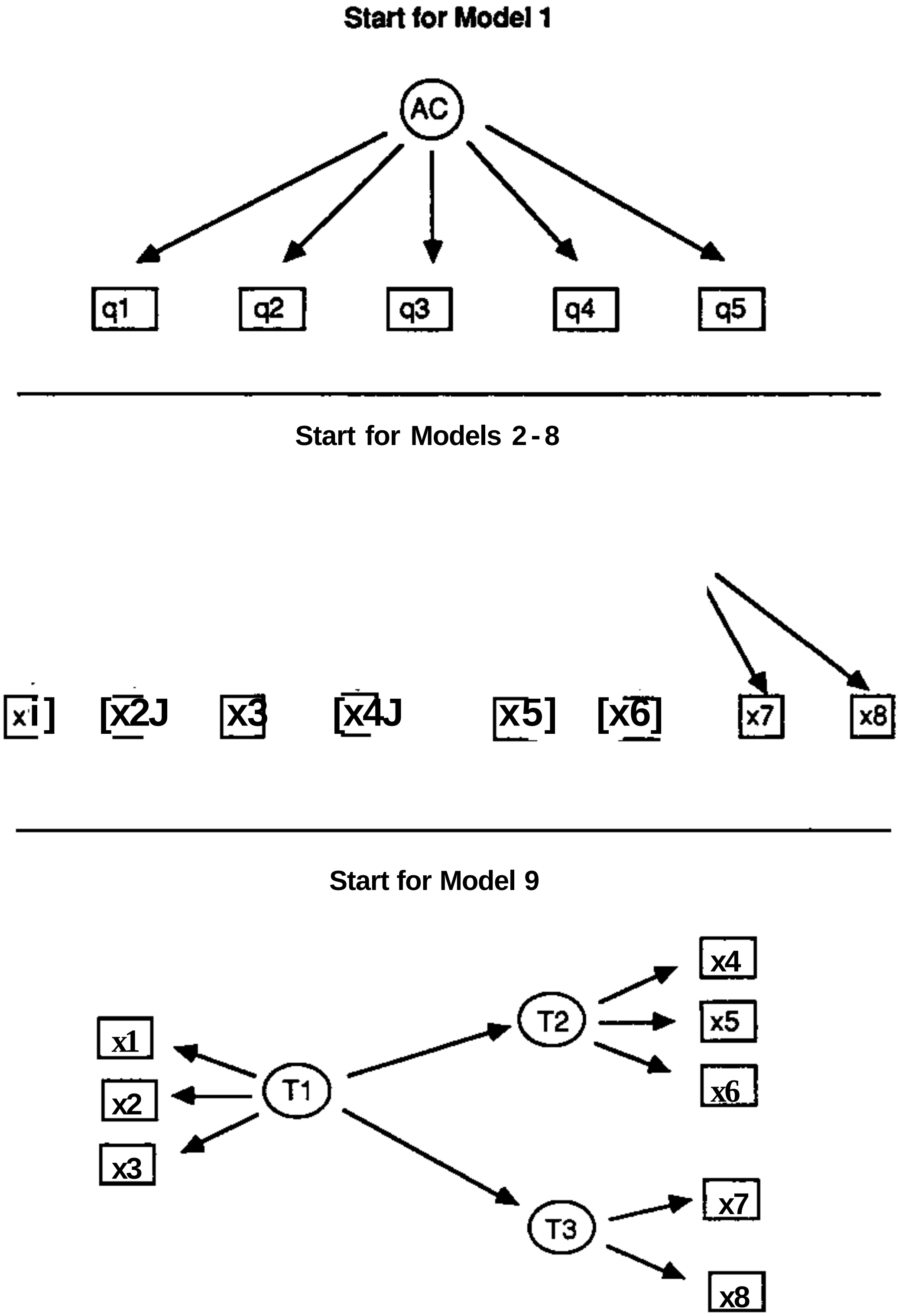
- Directed edges from latent variables to latent variables; relations of this kind are often the principal point of empirical research. See [22], for example.
- Edges from latent variables to measured variables; connections of this kind may arise when measures are impure, and in other contexts. See [6], for example.
- Correlated errors between measured variables; relationships of this kind are perhaps the most frequent form of respecification.
- Directed edges from measured variables to measured variables. Such relations cannot obtain, for example, between social indices, but they may very well obtain between responses to survey or psychometric instruments (see [5]), and of course between measured variables such as interest rates and housing sales.

One limitation of the primary study is that we consider only one case in which there is a mis-specification in the structural part of a model. Such models are typically large, and the time required by each of the programs to search through the space of alternatives to complex models for hundreds of different data sets was beyond the scope of this preliminary study. For example, LISREL VI on the Compaq 386 PC is slow enough to make certain procedures infeasible. We had initially planned to include a tenth case in our primary studies. The model to be studied was hierarchical, with ten measured variables serving as indicators of four latent variables that in turn depend on two second order latent variables. The omitted edge to be recovered connected one of the first order latent variables to another such variable. We found that LISREL VI required roughly one and one half hours to process this model on the Compaq 386, and that because the program often did not converge, batch processes were frequently stopped. The attempt to study the case was abandoned after running five data sets through LISREL VI. We are reasonably confident that the results for LISREL VI and for TETRAD II would, however, resemble those of case 9. We plan to investigate the reliability of each of the programs in correcting this kind of mis-specification more thoroughly in a future study.

We also have not included in the primary study cases that we know beforehand cannot be recovered by one or another of the programs. We have not, for example, used data generated by a model containing an edge directed into an exogenous variable in the starting model, because we know beforehand that LISREL VI cannot detect such edges. Neither have we included cases in which the true model implies the *same* set of vanishing tetrad differences as a proper submodel. We have not, for example, used starting model 3, and data generated from a model that extends starting model 3 with the addition of a T2 -> Tl edge. Such models cannot be distinguished by the TETRAD procedures.

Selection ofStarting Models:

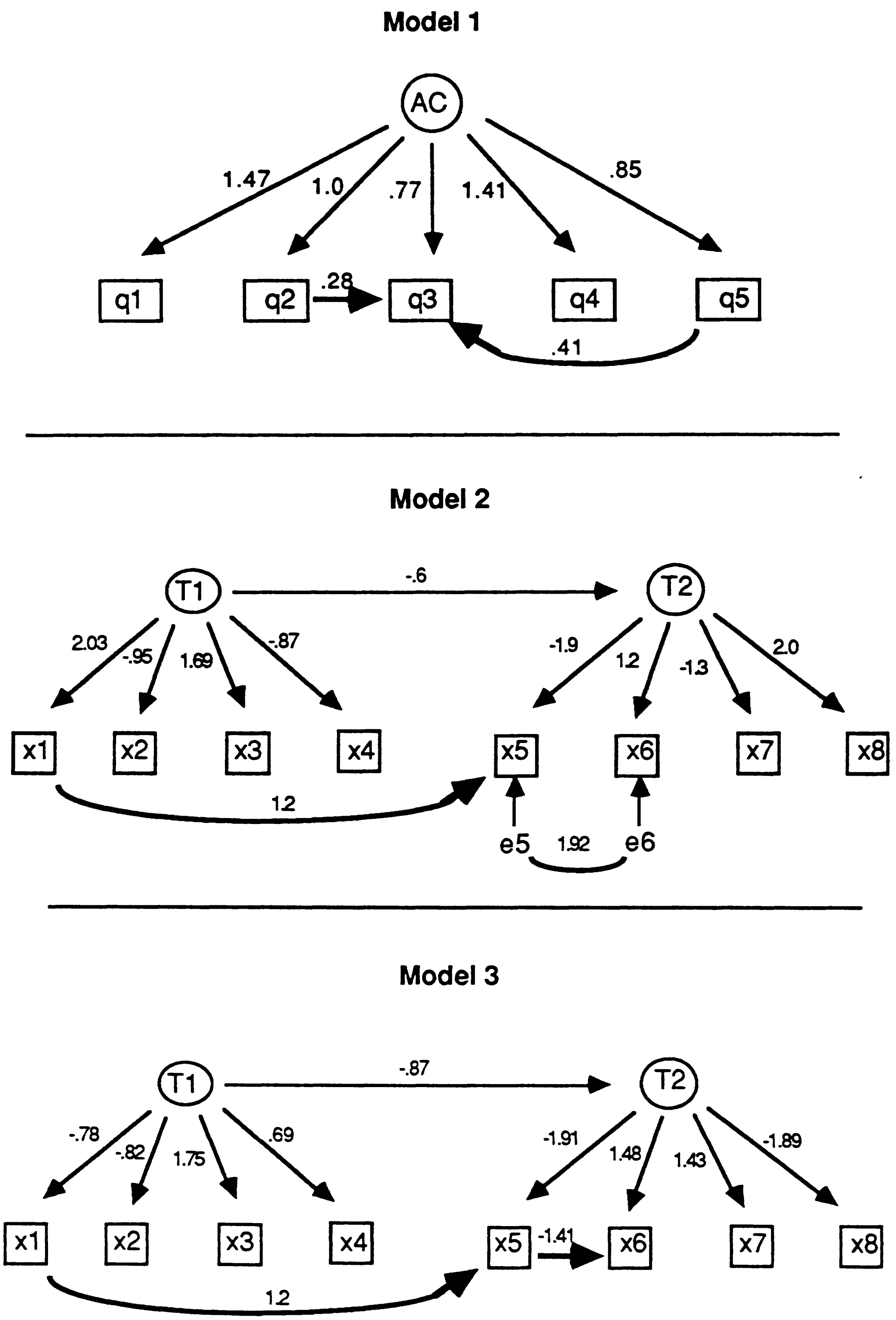
Only three starting models were used in the nine cases (see Figure 4). The starting models are, in causal modeling terms, pure factor models or pure multiple indicator models. In graph theoretic terms they are *trees.* In every empirical case we have read in which sample data are used to elaborate an initial latent variable model, the initial model is of this kind.


Start for Model 1

Start for Models 2-8

Start for Model 9

Figure 4: The Starting Models

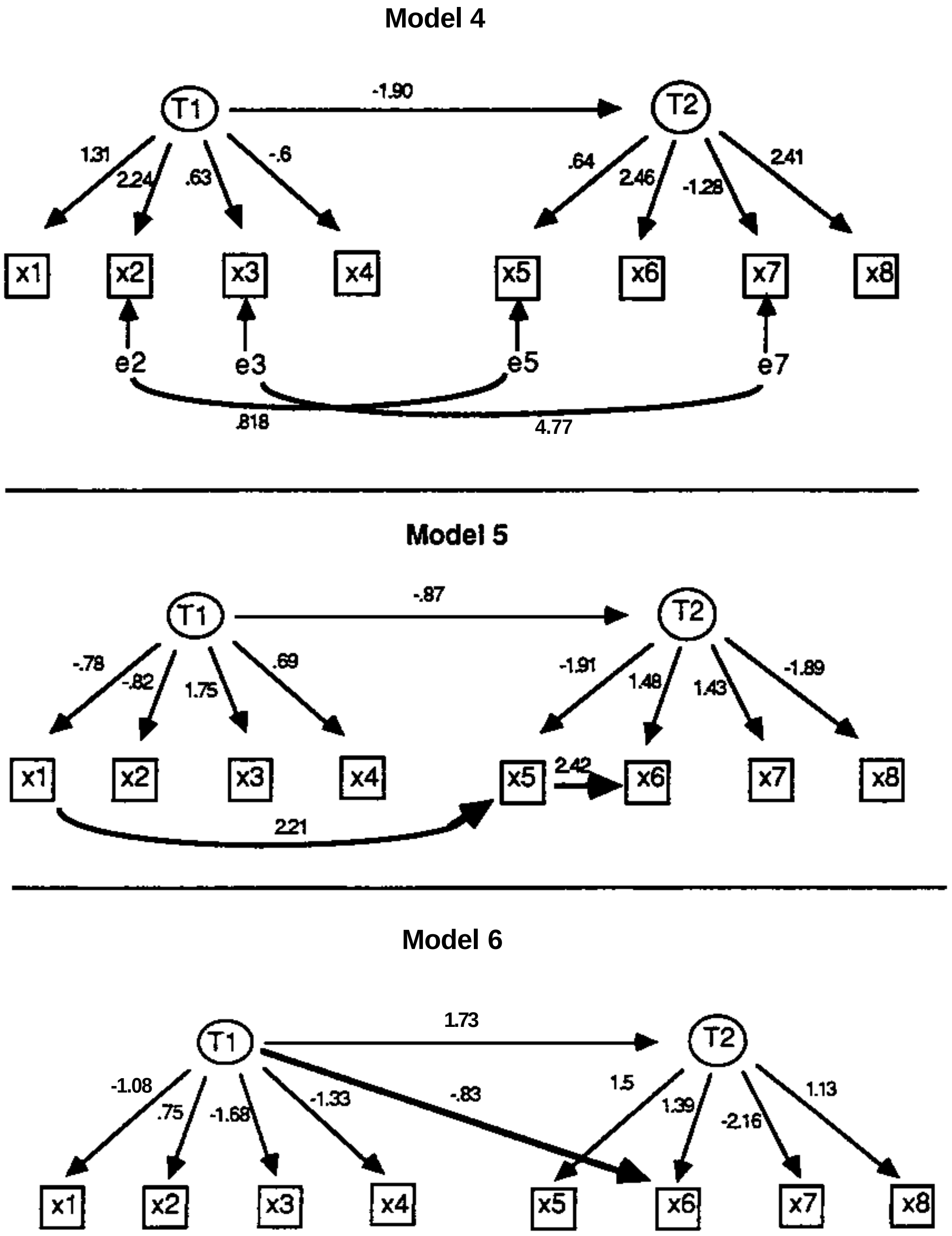


Model 1

Model 2

Model 3

Figure 5: True Models 1, 2, and 3


Model 4

Model 5

Model 6

Figure 6: True Models 4, 5, and 6

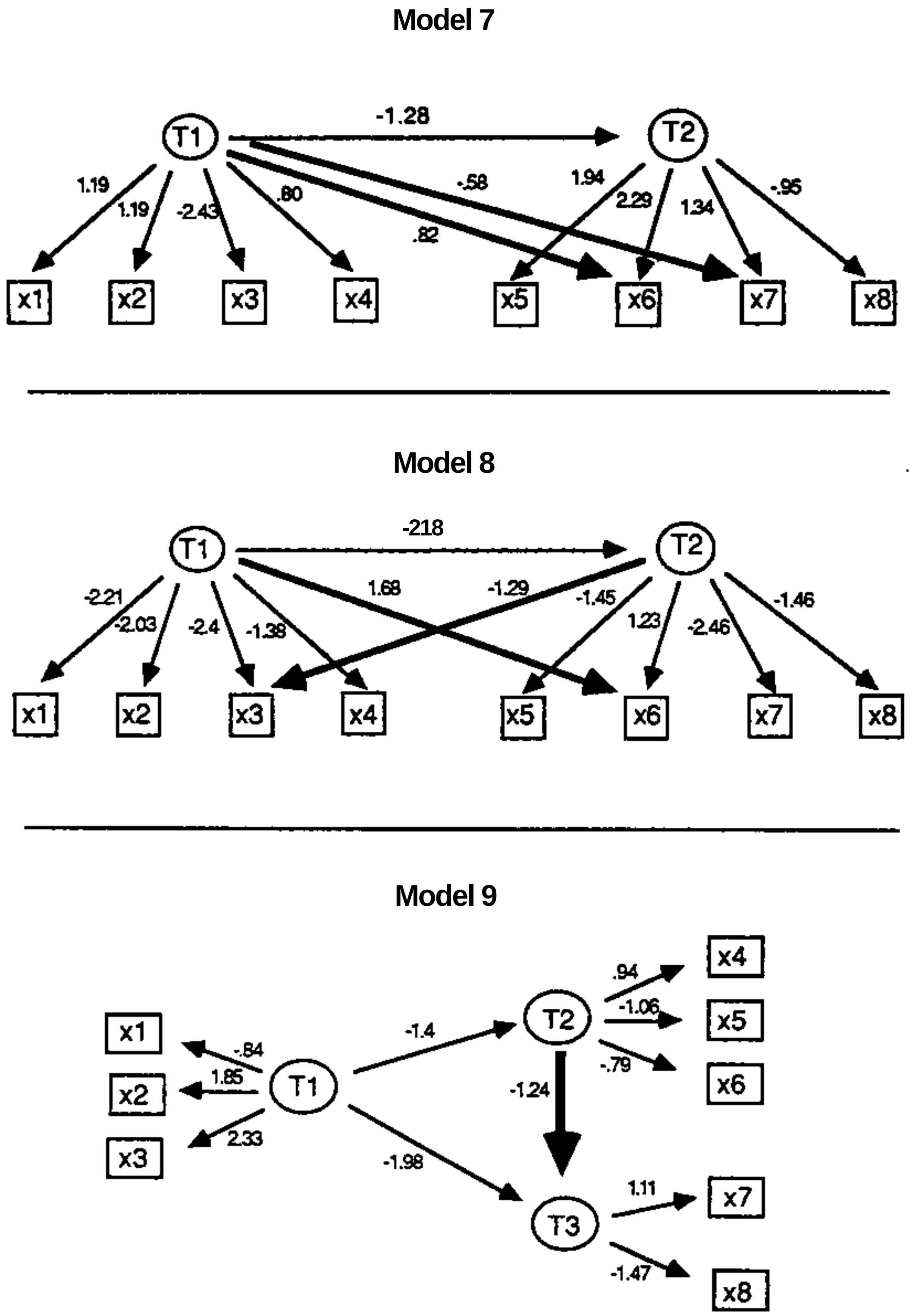


Figure 7: True Models 7, 8, and 9

Selection of Parameters

In the figures showing the true models the numbers next to directed edges represent the values given to the associated linear coefficients. The numbers next to undirected lines represent the values of specified covariances. In all cases, save for models 1 and 5, the coefficients were chosen by random selection from a uniform distribution between .5 and 2.5. The value obtained was then randomly given a sign, positive or negative. The minimum value of .5 was chosen to ensure the statistical significance of all parameters.

In model 1, all linear coefficients were chosen from the .5--2.5 interval, but all signs were made positive. The values of the causal connections between indicators were specified non-randomly. The case was constructed to simulate a psychometric or other study in which the loadings on the latent factor are known to be positive, and in which the direct interactions between measured variables are comparatively small.

Model 5 was chosen to provide a comparison with model 3 in which the coefficients of the measured-measured edges were deliberately chosen to be large relative to those in model 3.

Generation of Data

For each of the nine cases, twenty data sets with n=200 and twenty data sets with n=2,000 were obtained by first generating values for each of the exogenous variables (including error variables) with a random number generator giving a standard normal distribution[14], and then calculating the value of each endogenous variable as a linear function of its immediate causes. Correlated errors were obtained in the simulation by introducing a *new* exogenous common cause, and fixing the coefficients so that their product equals the covariance of the error terms.

Implementation:

The LISREL VI runs in the primary study were performed with the personal computer version of the program, run on a Compaq 386 computer. This machine is many times faster than IBM AT machines. EQS runs were performed on an IBM XT clone with a math coprocessor. Comparison LISREL VI and EQS runs (discussed in the section on ancillary studies) were performed on a VAX mainframe. All TETRAD II runs were performed on Sun 3/50 workstations.

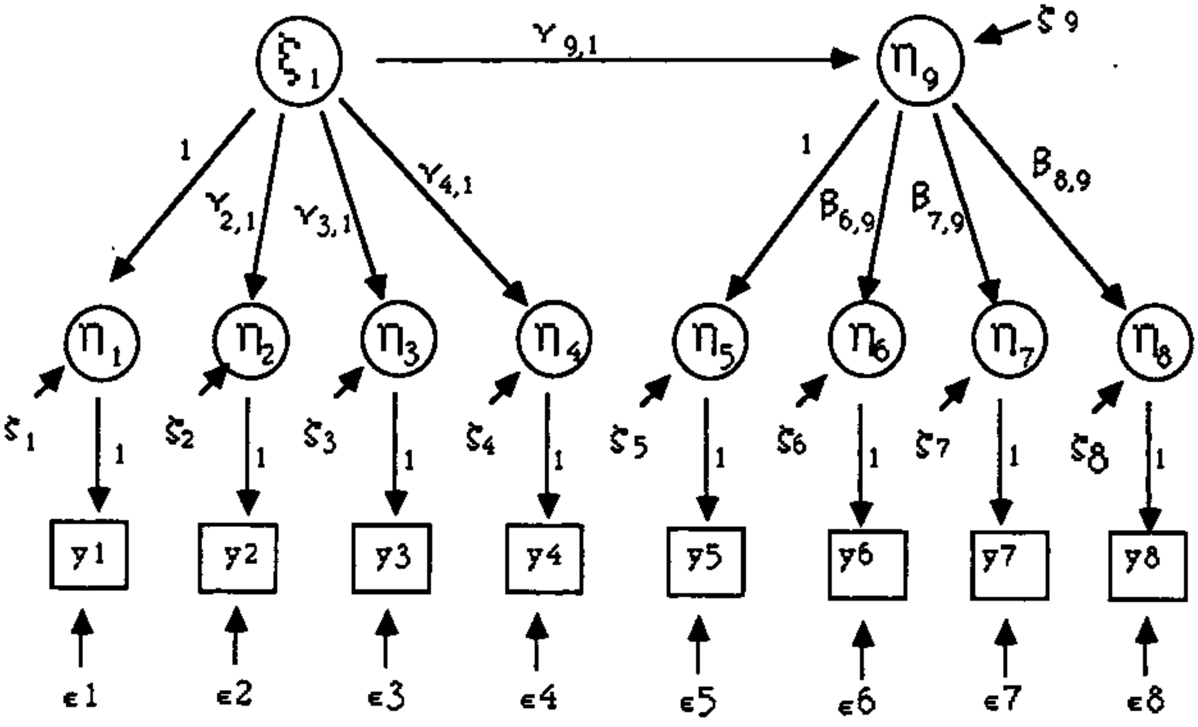
Specifying Starting Models in LISREL VI

LISREL VI, like previous editions of the program, requires the user to formulate models in such a way that variables are represented in distinct matrices according to whether they are

[14] First a number was pseudo-randomly selected from a uniform distribution by calling the "random" system call on the UNIX operating system. Then this number was input to a function which turned it into a pseudo-random sample from a standard normal distribution.

exogenous, endogenous but unmeasured, measured but dependent on exogenous latent, measured but dependent on endogenous latent, and so forth. Variables in certain of these categories cannot have effects on variables in other categories. When formulated as recommended in the LISREL manual, LISREL VI would be in principle unable to detect many of the effects considered in this study. However, these restrictions can in most cases be overcome by a system of substitutions of dummy variables in which measured variables are actually represented as endogenous latent variables.[15]

The following diagram depicts the LISREL VI model that corresponds to the starting model for cases 2 - 8 in Fig. 4. We include an example of a LISREL VI input file in the appendix.



$e_e = 0$

Figure 8: Starting Model for Cases 2-8: in LISREL Notation

[15] For LISREL IV, the details of this procedure are described in *Discovering Causal Structure*. The same procedure works for LISREL VI with the exception that we were not able to get LISREL VI to accept changing £ variables to *T*\ variables, as we recommended in the book. This had the unfortunate effect that LISREL would not consider adding any edges into Tl (represented by the £ variable).

LISREL VI and EQS contain a parameter for the significance level of the chi-square significance test used in their searches. In LISREL VI the default value (.01) of the parameter was used. For comparability, in EQS we used .01 rather than the default value (.05). Lowering the significance level tends to reduce the number of edges added in a LISREL or EQS search.

Specification ofTETRAD II Parameters

TETRAD II requires that the user set a value of the weight parameter (see section 2.3), and a value for the significance level used in the test for vanishing tetrad differences. A significance level of .05 was used for all samples; each sample was run with the weight parameter set at .1 and also with the weight parameter set at 1. In all cases, the breadth of the search was 30 and the initial depth was 3.

4.4. Results

For each data set and initial model, TETRAD II produces a set of best alternative elaborations. In some cases that set consists of a single model; typically it consists of three or four alternatives. EQS and LISREL VI, when run in their automatic search mode, produce as output a single model elaborating the initial model. The information provided by each program is scored "correct[11] when the output contains the true model. Since, however, it is important to see how the various programs err when their output is not correct, we have provided a detailed classification of various kinds of error. We have classified the output of TETRAD II as follows (where a model is in TETRAD'S top group iff it is tied for the highest T-score, and no model with the same T-score has more degrees of freedom):

Correct-xht true model is in TETRAD'S top group.

Width-the number of best alternatives (recorded only for searches
that contained the true model in TETRAD'S top group.)

Errors:

Overfit—TETRAD's top group does not contain the true model but
contains a model that is an elaboration of the true model.

Underfit—TETRAD'S top group does not contain the true model but
does contain a model of which the true model is an elaboration.

Other—none of the previous categories apply to the output.

We have scored the output of the LISREL VI and EQS programs as follows:

Correct--tht true model is recommended by the program.

Errors:

In TETRAD'S Top Group-the recommended model is not correct, but is among the best alternatives suggested by the TETRAD II program for the same data.

Overfit—the recommended model does not contain the true model but contains a model that is an elaboration of the true model.

Underfit—the recommended model does not contain the true model but does contain a model of which the true model is an elaboration.

Right Variable Pairs--the recommended model is not in any of the previous categories, but it does connect the same pairs of variables as were connected in the omitted parts of the true model.

Other-none of the previous categories apply to the output.

The results are summarized in Fig. 9. The solid black bars in the figures represent the performance of TETRAD II when the weight = .01; the shaded bar represents the program's performance when weight = 1.0.

For large sample sizes, TETRAD II is significantly more reliable for the kinds of cases we considered than either LISREL VI or EQS. TETRAD II was able to correct the mis-specification in almost 94% of the cases with weight = .1, whereas LISREL VI was able to do so less than 23% of the time and EQS a little more than 16%. For small sample sizes, the performance of all three programs was considerably worse,[16] although TETRAD II still significantly outperformed LISREL VI. TETRAD II corrected the mis-specification 37.2% of the time at weight = .1 and 41% of time at weight = 1.0, while LISREL VI corrected the mis-specification 12.8% of the time. A detailed case-by-case breakdown of the results is presented in the appendix. It should be noted that in these cases the performance of the LISREL VI and EQS programs is not impressive even if one includes among the correct responses those in which the appropriate pairs of variables were connected (but, of course, *incorrectly* connected).

LISREL VI (and EQS) output a single recommended model, whereas TETRAD II outputs a set of models. Output of a single model would be more informative if the single model selected could be relied upon to be correct, but our results strongly suggest that this is not the case. The automatic output of LISREL VI and EQS contains no information about whether there are alternatives to the model they suggest that have equal, or nearly equal, L-scores. One of the major advantages of TETRAD II is that it outputs the set of all of the alternatives that have the same or nearly the same T-score.

[16] We did not run EQS on all 180 data sets at sample size 200, but in the several cases we did its performance was still slightly inferior to LISREL VI.

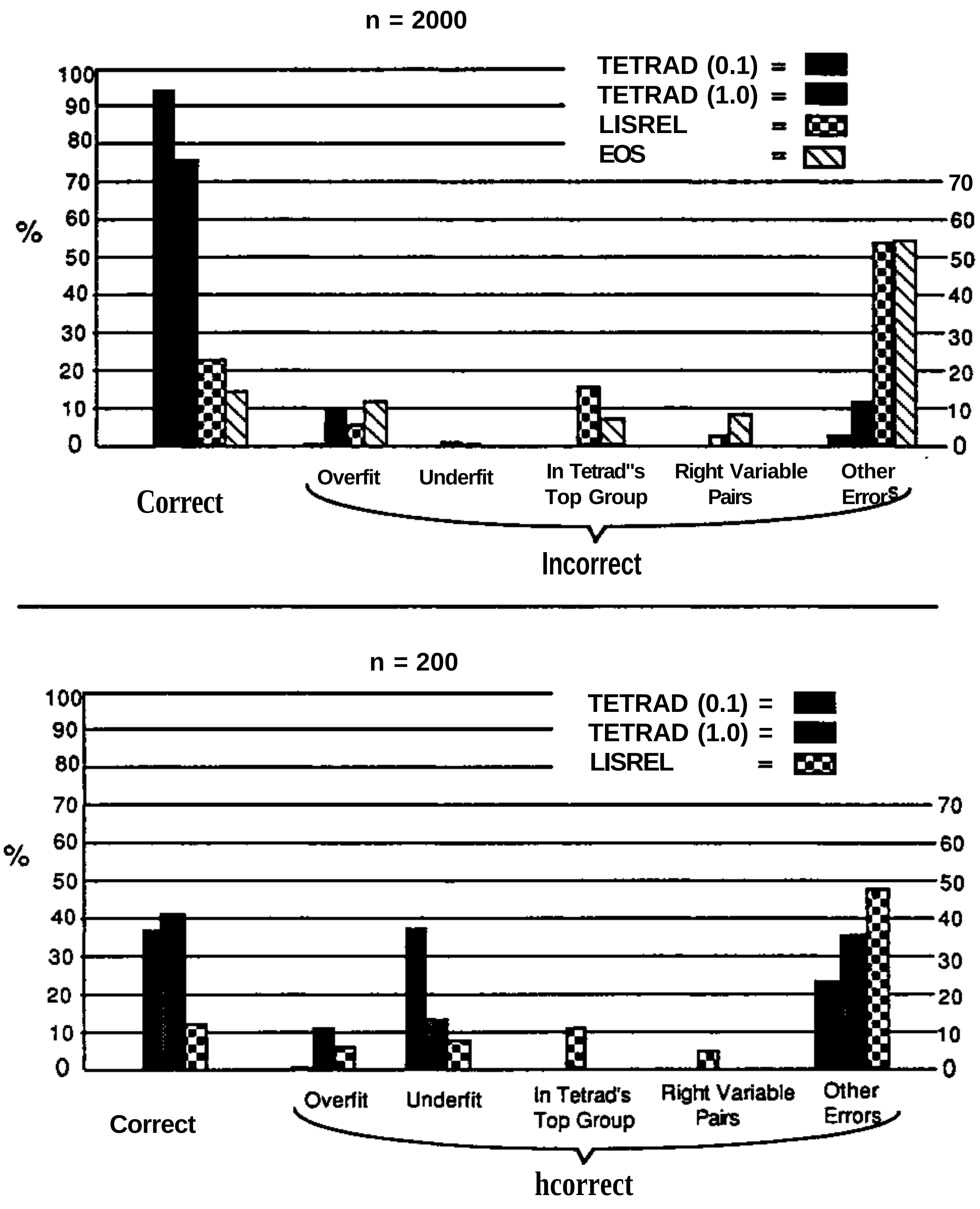


Figure 9: Summary of Results

This point is illustrated by a data set from case 1, for which TETRAD II gives the following four alternative additions to the initial model (a "*" marks the true model).

1. q2->q3 q5->q3 *
2. q2->q3 q5C q3
3. q2 C q3 q3 C q5
4. q2 C q3 q5 -> q3

All of these models receive identical *L-scores* from EQS and LISREL VI (that is, they have the same p values on a chi-square test), but the output of a LISREL VI or EQS search is at best only one of these models, and generally not the correct one. The user is not (save in the case that the modification involves a single edge), told of the alternatives. The fact that there are other models receiving exactly the same L-score should be highly relevant to a researcher's confidence in the recommended model.

5. Analysis and Ancillary Studies

5.1. Parameter Estimation Problems

The main difference in basic approach between TETRAD II and either LISREL VI or EQS is that TETRAD II does not involve any parameter estimation.

LISREL VI and EQS, unlike TETRAD II, require initial estimates of the signs and magnitudes of the free parameters. LISREL VI takes the initial estimates and then performs a two stage least squares estimate of parameters, before calling its maximum likelihood estimation procedure. Perhaps that step accounts for LISREL's slightly better performance.

In the starting models for our searches, we always fixed the parameter of one indicator of each latent variable at 1; all of the other edges were free parameters with .5 as a starting value. The variances of the latent variables were also free parameters with .5 as the starting values in the LISREL VI studies, and 1.5 in the EQS studies.

If the original estimates are too far from the actual values, the parameter estimates will either fail to converge or converge to the wrong values. In 3, 5, and 9, both LISREL VI and EQS repeatedly had serious convergence problems. In those cases, we changed the initial estimates so that they had the same signs as the corresponding parameters in the true model; the results of the searches when the initial estimates had the coiTect signs are reported in both the tables and the bar graphs. As can be seen from the very low success rates for **LISREL** VI and EQS on these models, this charity caused little **or** no improvement **on the proportion of correct outputs for data** sets from such models (except in case 9 at sample size 2000 for EQS).

How serious the parameter estimation **problem** is **depends** upon several factors.

In many cases convergence was obtained **when** the initial starting values had the correct signs,[17] but this requires more knowledge than a researcher may have. In many studies the sign of a crucial causal link *is* the operative question.[18]

Some causal structures are considerably more sensitive to incorrect starting values than other models. For example, when EQS was given all positive initial parameter estimates, we found that on each case either the program converged to approximately correct parameter estimates on all 20 data samples, or failed to do so on all 20 data samples.

If the original starting values are inadequate, it is always possible to search for starting values that will allow the parameter estimates to converge. However, in a large model, there are so many alternative sign configurations that such a search may not be feasible. For example, in case 9, for the first data sample, we randomly chose 20 different sign assignments to the free parameters in the starting model. In EQS, for only one of those 20 sign assignments did the parameter estimates of the starting model converge. For some models, in the absence of strong background knowledge about the free parameters in the starting model, finding a set of starting values that converge can be quite difficult and time consuming. Relevant prior knowledge may often be available. In psychometric studies, for example, it is often assumed that all dependencies of measured variables on latent variables involve linear coefficients of positive sign. In longitudinal studies in which the same battery of questions or indices are repeatedly applied, it is often reasonable to assume that the signs of the coefficients are unchanged in two or more administrations of a survey instrument.

5.2. Failure of Calculation in the Search

TETRAD IPs search procedure never halted because of calculation problems. However, even when LISREL VFs parameter estimation for the starting model converged, the LISREL VI search quite often failed either because it was unable to estimate the parameters of some elaboration of the starting model or because it could not calculate the modification indices. For

[17] It should be noted that in linear modeling with latent variables, the practice of fixing for each latent variable the coefficient of one indicator variable at unity has the consequence that, even if the causal stnicture is correct, it cannot be concluded that the true signs of the coefficients are as in the estimated model. Conventionally fixing the coefficients at -1 rather than 1 would typically reverse many signs.

[18] For example, see our discussion of Timberlake and William's model of the effect of foreign capital investment on the level of repression in third world countries, [12], chapter 8.

example, in case 9 at sample size 2000, even when the starting model contained true initial parameter estimates for the edges in the starting model, LISREL VI's automatic search ended 6 times out of 20 with a failure to calculate parameter estimates or modification indices. In case 3 at sample size 2000, 6 of the 18 searches in which LISREL VI succeeded in estimating the parameters of the starting model ended with a similar failure. (If LISREL attempted to add an edge to a model, but was unable to re-estimate the parameters or to calculate the modification indices, we still counted the edge that it attempted to add as part of the recommended model.)

Costner and Herting found much the same problem in their study. In particular, when a direct connection between a latent variable and an indicator shared by two latent variables was left out of the starting model, LISREL VI tended to add an incorrect edge to the model, and then calculate that some variable had a negative variance. This happened especially when the missing latent - indicator edge had a large coefficient. Similar problems occurred when the missing edge was a latent - latent edge. They found that going back to the starting model and adding the edge with the second highest modification index usually solved the calculation problem and also gave the correct answer.

5.3. Breaking Ties

In a large number of cases, there were several edges tied for the highest modification index. It is not clear from the documentation how LISREL VI chooses which of these edges to add to the model, but whatever it is, it is not reliable. For example, in case 1 at sample sizes 200 and 2000, the edges x3 -> x5, x3 C x5, and x5 -> x3 all had identical, or very nearly identical modification indices. At sample size 2000, LISREL VI added was x5 -> x3 (the correct edge) eight times, x3 -> x5 ten times, and x3 C x5 one time. EQS was similarly unreliable in breaking ties among edges that would cause identical or nearly identical estimated decreases in the chi-square.

5.4. Choosing the Wrong Edge

There are several reasons why the edge with the highest modification index might not be the correct edge to add to a model. Since the modification index places only a lower bound on the decrease in the chi-square, it is possible that freeing a fixed parameter that does not have the highest modification index could result in the greatest decrease in the chi-square. And since at each stage of the search LISREL VI chooses to free only the fixed parameter with the highest modification index, it could easily miss freeing other fixed parameters that would produce larger decreases in the chi-square. Also, the reliability of the modification indices is highly dependent upon the particular values of the parameters freed. This is documented in [7], and illustrated in several of the Monte Carlo studies we ran.

We found that it was common that all of the edges with the highest modification indices were erroneous. A number of factors affected how often this happened.

The first factor was the causal structure. When the true model contained a pair of edges that linked together to form a path, and the starting model did not contain those edges, LISREL VI performed worst. For example, in case 5, which has linked edges, LISREL VI always chose the wrong edge at both sample sizes 200 and 2000 (when it converged). In contrast, when the true model had correlated errors, LISREL VI performed the best. For example, in case 4 (which did not have linked edges), at sample size 2000 the edge with the highest modification index was correct 19 out of 20 times. (Costner and Herring also found that LISREL VI was most reliable in detecting correlated errors. In a wide variety of different cases with correlated errors, LISREL VI was always able to detect the missing correlated errors.)

The second factor was the parameter size. When the linked edges have linear coefficients that are large compared to the other coefficients in the model, LISREL VI often chose the wrong edge. The only difference between case 3 and case 5 was that in case 5 the linear coefficients of the linked edges were large compared to the other coefficients in the model, and in case 3 they were not In case 5 at sample size 2000, when LISREL converged, at the first stage of the search, the edge with the top modification index didn't even connect the correct pair of variables in any of the 20 samples. In case 3 at sample size 2000, the overall performance of LISREL VI was still poor. However, at least the edge with the largest modification index connected the correct pair of variables (although often in the wrong way) in all of the 19 cases in which the parameter estimates converged. It should be noted that while LISREL VI and EQS have special difficulties when the coefficients of omitted edges are large, there is no such effect with the TETRAD II program. In practice, the more important such an effect is in the true process generating the data, the less likely LISREL and EQS are to find it.

5.5. Search Strategy

Some of the comparatively poor performance of the LISREL VI and EQS programs on the kinds of cases that we considered in our study can be traced to their search strategy. With LISREL VI, for example, when several alternative models have the same or virtually the same (e.g., differing by round off error) modification indices, LISREL VI picks one of them by no principle we can find documented.[19] All of the other alternatives are discarded, and the program proceeds to look for further modifications to the one selected best alternative. This *beam search* strategy produces many errors. The TETRAD II performance would be better approximated by LISREL VI and by EQS if, when several models with virtually identical maximal scores are found, their searches were to branch (as does that of TETRAD II), and the program were to report all of the alternatives found. A branching search is computationally costly however, and would, of course, require at least an order of magnitude more time. Moreover in the worst case the extra time required for a branching search would grow exponentially with the number of alternatives.

[19] The PC and mainframe versions of LISREL VI will in some cases make different choices given the same data and initial model.

5.6. Computational Demands

Since the three programs were run on different machines, no direct comparison of their computational demands is possible. For the breadths and depths used in these studies the average running time **for TETRAD II was about 10 minutes on Sun 3/50 work stations. Although TETRAD II is not implemented on MS-DOS, TETRAD is, and runs in about the same** time on **the Compaq 386 machine (used for the USREL VI runs in the primary study) as on the Sun 3/50. Thus the run times of USREL VI and TETRAD II may be indirectly** compared: The run time for **USREL** VI is roughly twice **that of TETRAD II. EQS** runs **were done on** still a different machine, and we have in this case no Rosetta stone to compare speeds with the other two programs.[20]

In many cases the difference in time for TETRAD II and for LISREL VI processing would be relatively inconsequential were the latter program otherwise reliable. It becomes important in considering a branch strategy using modification indices. Typical runs with USREL·VI on the Compaq 386 required 15 to 25 minutes, depending on the model.[21] A branching procedure that retained three alternatives at each stage stopped on all branches after freeing two parameters in the initial model would increase the time required by about a factor of 7. In many cases any reasonable branching would be much broader. For example, in case 1 of our primary study, after LISREL VI added the first edge, on some data sets there were 11 alternative further additions indistinguishable to within four of the five significant figures LISREL reported. In general, the time required for a LISREL branch search would increase exponentially as the number of alternatives considered at each stage increases.

5.7. Stopping Criteria

LISREL VI stops searching when adding the edge with the highest modification index does not cause a significant drop in the chi-square of the model. In practice, this stopping criterion often does not work. For example, at sample size 2000, TETRAD II overfit or underfit a total of 5 times; LISREL VI underfit or overfit 12 times. In many other samples that did not fall into the overfit or underfit categories, LISREL VI found a model that was in TETRAD ITs top group, but then continued to add further edges to the model.

[20] If the reader wonders why the processing was done on so many different machines, the answer is simple: because of repeated convergence difficulties which stop batch processes, LISREL VI is inefficient to run even in batch. (Although EQS also suffers from convergence difficulties it is not as inefficient to run in batch, since it does not require manual intervention when convergence difficulties occur.) Processing 360 data sets for LISREL and 180 for EQS required hundreds of hours. Run in parallel the study took us about two months. Run sequentially on the Compaq 386, for example, it would have taken us at least half again as long. TETRAD II was run on Sun 3/50 workstations because the Andrew system facilities of Carnegie Mellon University enabled us to run the program simultaneously on several workstations.

[21] The version of LISREL VI that runs on large VAXes under the VMS operating system is considerable faster.

5.8. Using LISREL VI and EQS as Adjuncts to a TETRAD II Search

Models that receive the same T-score do not necessarily receive the same L-score. In some cases, the L-scores do make useful distinctions among models that receive the same T-score, particularly in those cases where two models contain the same treks, but different pairs of variables are directly connected.

A sensible strategy might be to use TETRAD II to generate a list of alternative revisions of an initial model, and then test the revisions with LISREL or EQS, discarding those alternatives that have very low, or comparatively low L-scores. Unfortunately, there are often convergence problems which limit the usefulness of this strategy. For example, at sample size 200, in cases 3, 5, 6, 7, 8, and 9 there was at least one model in the top group output by TETRAD II which caused convergence problems in LISREL VI. Even when the free parameters in each of the models were given the correct starting values, there were convergence problems at sample size 200 in cases 3, 7, and 8. A model that causes convergence problems cannot be dropped from consideration, even if other models do not have convergence problems. For example, in case 7, the true model had convergence problems, and a false model did not.

To systematically test the merits of this strategy we performed the following study. For each case, we chose a sample data set (n = 200) upon which TETRAD ITs search was successful. We then constructed two LISREL and EQS input files for each of the models in TETRAD'S top group[22], one in which all the parameters had starting values of .5, and the other in which all parameters had correct starting values. In case 4, where TETRAD'S top group contained only the correct model, this strategy is obviously not necessary. In these cases we ran LISREL and EQS on two data sets upon which TETRAD IPs search was successful. The tables shown below list the results.

[22] For cases 6,7 and 8, we only constructed input files for a subset of the models in TETRAD II's top group.

Statistical Results on Models in TETRAD's Top Group: EQS

* = correct model
s.p. = serious problems (e.g. failure to converge)

		Parameter Starting Values			
		All = .5		True	
	D.o.f.	X^2	$p(X^2)$	X^2	$p(X^2)$
Model 1: data set 3, width = 4					
1) q2 -> q3, q3 C q5	3	1.19	.75	1.19	.75
2) q2 C q3, q3 C q5	3	1.19	.75	1.19	.75
3) q2 -> q3, q5 -> q3 *	3	1.19	.75	1.19	.75
4) q2 C q3, q5 -> q3	3	1.19	.75	1.19	.75
Model 2: data set 9, width = 2					
1) x1 -> x5, x5 C x6 *	17	32.11	.015	32.11	.015
2) x1 -> x5, x6 -> x5	17	s.p.		32.11	.015
Model 3: data set 20, width = 3					
1) x1 -> x5, x1 -> x6	17	s.p.		22.5	.166
2) x1 -> x5, x5 -> x6 *	17	s.p.		9.8	.911
3) x1 -> x6, x6 -> x5	17	s.p.		s.p.	
Model 4: width = 1					
data set 2					
x2 C x5, x3 C x7 *	17	s.p.		9.15	.935
data set 3					
x2 C x5, x3 C x7 *	17	s.p.		18.06	.385
Model 5: data set 1, width = 3					
1) x1 -> x5, x1 -> x6	17	s.p.		103.7	<.001
2) x1 -> x5, x5 -> x6 *	17	s.p.		12.87	.744
3) x1 -> x6, x6 -> x5	17	s.p.		14.48	.633
Model 6: data set 2, width = 2					
data set 2					
1) T1 -> x6 *	18	22.28	.22	22.28	.22
data set 4					
2) T1 -> x6 *	18	10.58	.911	10.58	.911

EQS

		Parameter Starting Values			
		All = .5		True	
	D.o.f.	X^2	$P(X^2)$	X^2	$P(X^2)$
Model 7: data set 17, width = 12					
1)X7->X6,T1->X7	17	28.81	.036	28.81	.036
2) x6 -> T2, Tl -> x7	17	s.p.		68..02	<.001
3)Tl->x6,x7->T2	17	s.p.		s.p.	
4) Tl -> x6, x6 -> x7	17	11.9	.803	11.95	.803
5) Tl -> x6, Tl -> x7 *	17	s.p.		11.99	.800
6) x6 C x7, x7 -> T2	17	s.p.		s.p.	
Model 8: data set 8, width = 13					
1) x3 -> x6, x3 -> T2	17	26.6	.064	26.6	.064
2) T2 -> x3, Tl -> x6 *	17	s.p.		14.64	.621
3) x6 -> x3, x3 -> T2	17	28.25	.042	s.p.	
Model 9: data set 5, width = 3					
1) T2 -> T3 *	17	s.p.		12.6	.762
2) T2 C T3	17	s.p.		12.6	.762
3) T3 -> T2	17	s.p.		12.6	.762

Statistical Results on Models in TETRAD'S Top Group: LISREL

* = correct model
s.p. = serious problems (e.g. failure to converge)

		Parameter Starting Values			
		All = .5		True	
	D.o.f.	X^2	$P(X^2)$	X^2	$P(X^2)$
Model 1: data set 3, width = 4					
Dq2->q3,q3Cq5	3	1.19	.75	1.19	.75
2)q2Cq3, q3Cq5	3	1.19	.75	1.19	.75
3)q2->q3,q5->q3*	3	1.19	.75	1.19	.75
4)q2Cq3,q5->q3	3	1.19	.75	1.19	.75
Model 2: data set 9, width = 2					
1) xl -> x5, x5 C x6 *	17	32.13	.014	32.13	.014
2) xl -> x5, x6 -> x5	17	32.13	.014	32.13	.014
Model 3: data set 20, width = 3					
1) xl -> x5, xl -> x6	17	s.p.		22.47	.167
2) xl -> x5, x5 -> x6 *	17	s.p.		9.8	.912
3) xl -> x6, x6 -> x5	17	s.p.		s.p.	
Model 4: width = 1					
data set 2					
x2 C x5, x3 C x7 *	17	9.15	.935	9.15	.935
data set 3					
x2 C x5, x3 C x7 *	17	18.09	.383	18.09	.383
Model 5: data set 1, width = 3					
1) xl -> x5, xl -> x6	17	81.52	.000	81.52	.000
2) xl -> x5, x5 -> x6 *	17	12.84	.747	12.84	.747
3) xl -> x6, x6 -> x5	17	s.p.		14.57	.627
Model 6: data set 2, width = 2					
data set 2					
l)Tl->x6*	18	s.p.		22.24	.222
data set 4					
2) Tl -> x6 *	18	s.p.		10.59	.911

LISREL

		Parameter Starting Values			
		All = .5		True	
	D.o.f.	**X^2**	$P(X^2)$	X2	$P(X^2)$
Model 7: data set 17, width = 12					
1) X7->X6, T1->X7	17	**s.p.**		**28.85**	**.036**
2) x6->T2, Tl->x7	17	**s.p.**		**s.p.**	
3) Tl -> x6, x7 -> T2	17	12.0	.80	s.p.	
4) Tl -> x6, x6 -> x7	17	**s.p.**		11.95	.803
5) Tl -> x6, Tl -> x7 *	17	**s.p.**		12.0	**.800**
6) x6 C x7, x7 -> T2	17	**s.p.**		28.85	**.036**
Model 8: data set 8, width = 13					
1) x3 -> x6, x3 -> T2	17	**26.65**	**.063**	**26.65**	**.063**
2) T2 -> x3, Tl -> x6 *	17	14.66	.620	14.66	.620
3) x6 -> x3, x3 -> T2	17	s.p.		**s.p.**	
Model 9: data set 5, width = 3					
1) T2 -> T3 *	17	s.p.		12.62	.761
2) T2 C T3	17	**s.p.**		12.62	.761
3) T3 -> T2	17	**s.p.**		12.62	.761

In cases 5, and 8 TETRAD II's top group contains some models that receive unacceptably low L-scores. Eliminating these models from consideration reduces the top group without introducing any errors. It is clear, however, that with arbitrary starting values, EQS and LISREL VI are rarely of use in reducing the alternatives produced by TETRAD DL One model in case 8 and one in case 5, and no others, are eliminated.[23]

5.9. Limitations of TETRAD II

There are connections in some models that, if omitted, TETRAD II cannot possibly recover. If a starting model and a true model imply the same set of tetrad equations, even though the true model contains additional dependencies not included in the starting model, then TETRAD II

[23] The convergence of either the LISREL VI or EQS programs in almost all cases when given the correct coefficients as starting values indicates that the difficulties the programs have with arbitrary starting values are not due to identification problems.

cannot discover those dependencies. The causal connections that TETRAD II cannot discover often occur in unidentifiable models, and in these cases LISREL VI will not discover the connections either. For example, if to the third starting model of our primary study we add a correlated error between variables Tl and T2, the result is a model that implies the same set of vanishing tetrad differences as does the starting model. Given the starting model and data generated by the elaborated model, TETRAD II cannot discover the correlated error between Tl and T2. Neither can LISREL VI. To verify that claim we generated 20 data sets from such an elaborated model (with positive coefficients chosen randomly between .5 and 1.5, and the Tl T2 correlation equal to .89) and gave the starting model and the data sets (n=2000) to LISREL VI. LISREL VI recommended no revisions to the starting model for 8 of the 20 data sets; it recommended connections between irrelevant pairs of variables for 6 of the data sets, and it failed to converge for the remaining 6 data sets.

There certainly are cases in which, with sufficient prior knowledge, LISREL VI, and presumably EQS as well, will with some reliability locate connections that TETRAD II cannot detect. MacCallum's study provides an example. MacCallum generated twenty data sets (n=300) from the LISREL VI model shown in Fig. 10.

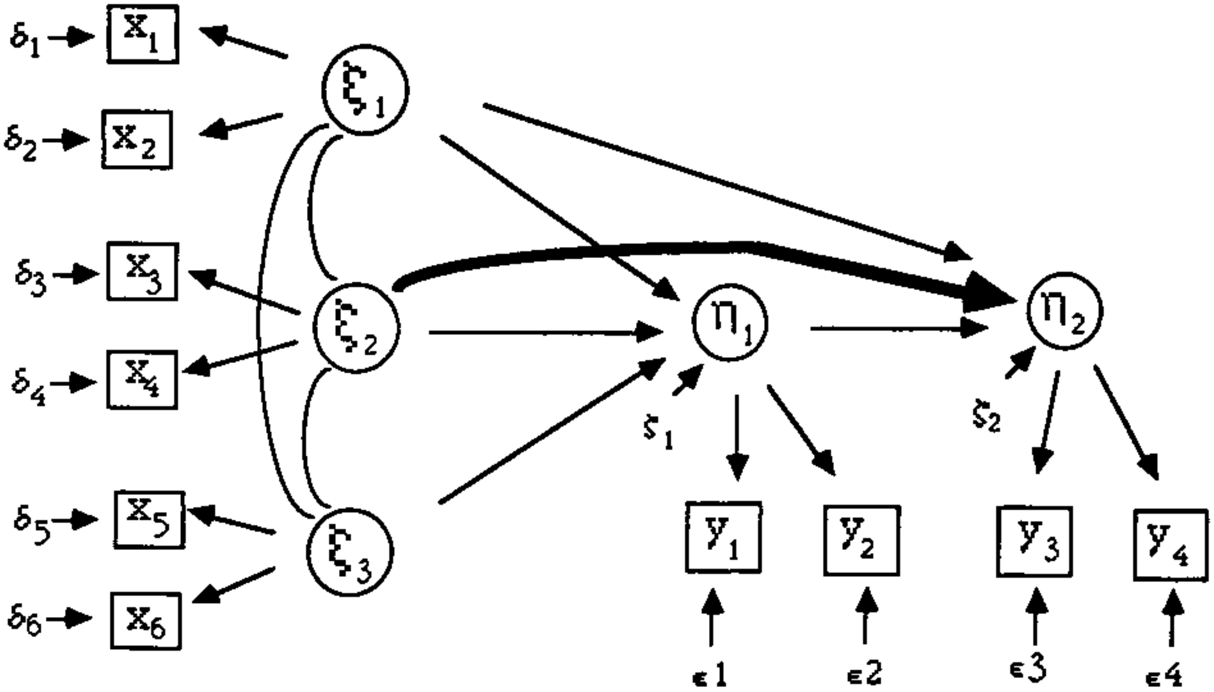


Figure 10: MacCallum's True Model

The darkened connection was omitted in the model given to LISREL VI. The automatic search procedure correctly located the missing connection for 17 of the 20 data sets. TETRAD II cannot discover the omission. The difference is less impressive, however, when one remembers all of the restrictions that were placed upon the LISREL search. There are 297 different one-parameter modifications that could be made to the initial model. MacCallum's LISREL search considered only direct causal connections among latent variables. Since no ξ variable can have an edge directed into it, this leaves only three alternative one-parameter modifications: an edge from ξ_2 to η_2, an edge from ξ_3 to η_2, or an edge from η_2 to η_1. These three modifications, together with the hypothesis of no modification, were the initial hypothesis space the program had to consider. We do not know how LISREL VI would perform if the search space were not artificially restricted in this way, but our experience with model 9 of the primary study suggests that its reliability would be greatly decreased. MacCallum himself had the same suspicion, which is why he restricted the search so drastically. For correcting errors in the measurement model, he suggested using a method devised by Costner and Schoenberg (in Costner73), which is similar in spirit to the method used by TETRAD II.

5.10. Using TETRAD

A legitimate practical concern is whether the TETRAD II results could be reproduced with the published TETRAD program. To answer this question unequivocally would require another large study which we do not propose to undertake. Our opinion, however, is that if the TETRAD program were used on these data following the heuristics described in *Discovering Causal Structure*, the results would be satisfactory in cases with large samples, but not so good as those obtained with the TETRAD II program.

An illustration of the application of TETRAD to one of these problems is given in the appendix.

6. Conclusions and Prospects

Based on our study, we put little confidence in the automatic model elaboration procedures in EQS and LISREL VI when applied to latent variable models of the kind we considered. In producing a unique respecification, the programs attempt more precision than can be reliably obtained. Further, their numerically based architecture makes the programs unstable and makes a more adequate search strategy unpromising for computational reasons.

In contrast, over a range of latent variable structures, the TETRAD II procedures are quite reliable in large samples. There are, however, special cases in which TETRAD II cannot possibly locate missing connections. Those are cases in which the missing connections, when added to the initial model, have no effect on the implied tetrad constraints. TETRAD will also fail when a large number of tetrad constraints appear to hold in the sample because of the numerical values of the linear coefficients rather than because of the causal structure. In large samples, for any natural probability distribution over the coefficients, cases of the second sort will be rather rare (a point confirmed by our primary study).

If the automatic procedures of EQS and LISREL VI are too bold in attempting to produce a unique extension of the initial model, TETRAD II may be too timid in the set of alternatives it generates. Reducing that set of alternatives by using LISREL VI or EQS to test the alternatives is a harmless strategy, so far as we can judge, but it may not be very effective. Fortunately we believe there is an alternative strategy that may be more effective. Besides vanishing tetrad differences, the graph of a structural equation model, *along with the signs of certain coefficients,* may imply *tetrad inequalities.* That is, it may imply that a tetrad difference is greater than or equal to zero. It is computationally inefficient to compute all possible sign assignments to coefficients and the corresponding implied tetrad inequalities in the course of the TETRAD II search. It may be useful and feasible, however, to perform such computations on the models in the TETRAD II output. Some, but not others, may admit sign assignments to coefficients that are consistent with prior knowledge and imply the tetrad inequalities reflected in the data.

What holds for latent variable models may not hold for path models without latent variables, or for mixed models in which latent variables occur as effects rather than causes of measured variables. TETRAD II can be adapted to elaborate path models using vanishing partial correlations and partial correlation inequalities rather than tetrad equations and inequalities. We have not yet implemented such a procedure, but we plan to do so and to carry out a similar battery of simulation tests.

An important question that this study has not addressed concerns the robustness of the search procedures under failures of the general modeling assumptions. Two obvious sources of concern are the assumptions of multinormality and linearity. We plan to test the reliability of all three programs under failures of both assumptions. Our expectation is that, for large samples, TETRAD II will be quite robust under failures of the normality assumption.

The omission of causal relationships or correlated errors is not the only kind of specification error in causal modeling, and perhaps not even the most frequent or most important kind. The initial model provided to any of the programs considered here may have errors of commission, or it may simply have the wrong clusterings of measured variables with latent variables. The EQS program admirably attempts to address part of this problem through a statistical test for errors of inclusion. We have not attempted to assess the reliability of the EQS procedure, but we are pursuing a much more radical line of work.

Researchers who use linear modeling techniques often have quite fragmentary knowledge about the processes that gave rise to their data. If they are forced to produce a specific model they will inevitably fill out that fragmentary knowledge with assumptions that may or may not be true; typically, they may have a difficult time conceiving of many alternatives, although logically and mathematically, a great many alternatives may exist consistent with their fragmentary knowledge. We believe that *when the general structural equation modeling assumptions are met* there are procedures that will,

- consistent with the background knowledge, determine whether or not the data are generated by latent common causes (see [27,24]).
- consistently with the background knowledge construct all models having graphs that are trees and implying the tetrad constraints found in the data;
- elaborate each tree into a set of models that best explain the data;
- reliably produce as output a set of alternative models that includes the true model.

We possess algorithms for the first three items; we plan to implement and test them in the coming year. If our belief is correct, the procedures should considerably extend the scope and reliability of automated aids to the construction of causal models.

Appendix I

Results for the Primary Study

TETRAD: Weight = .1, n = 2000

Model	1	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Correct	17	**20**	**18**	**19**	**19**	20	20	18	**18**
-Width	4	**2**	**3**	**1**	**3**	2	10.4	13	**3**
Incorrect	**3**	**0**	**2**	**1**	1	0	**0**	2	2
-Overfit	**0**	**0**	**1**	**0**	**0**	**0**	**0**	1	**0**
-Underfit	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	0
-Other	**3**	**0**	**1**	**0**	1	**0**	**0**	1	2

TETRAD: ^Weight =1.0, ri = 200C>

Model	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Correct	17	15	16	**17**	**16**	14	**16**	13	15
-Width	4	2	3	**1**	**3**	2	**12**	13	**3**
Incorrect	3	5	4	**3**	**4**	6	**4**	7	**5**
-Overfit	1	**2**	3	**3**	**3**	**3**	**1**	1	1
-Underfit	0	**0**	0	0	0	**0**	**0**	**0**	0
-Other	2	**3**	1	0	1	**3**	**3**	**6**	4

Table 1-1: TETRAD at sample size = 2,000

LISREL, n== 2000									
Model	**1**	2	**3**	**4**	5	6	**7**	8	9
Correct	7	5	**3**	**15**	0	5	**0**	5	**1**
Incorrect	13	15	**17**	5	20	**15**	20	15	19
-Overfit	0	3	0	3	0	**3**	0	**1**	**0**
-Underfit	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**
-In Tetrad's Top Group	**1**	6	**0**	0	**0**	**4**	5	**10**	**1**
-Right Var. Pairs	**3**	**0**	**0**	**2**	**0**	**0**	**0**	**0**	**0**
-Other	**9**	6	**17**	**0**	20	8	**15**	**4**	**18**

EQS, n = 2000									
Model	**1**	**2**	**3**	**4**	5	6	**7**	**8**	9
Correct	**2**	**3**	0	15	0	3	0	**0**	**6**
Incorrect	**18**	**17**	20	5	**20**	**17**	**20**	20	**14**
-Overfit	**0**	**0**	**0**	**1**	**3**	**1**	**3**	**7**	**5**
-Underfit	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
-In Tetrad's Top Group	7	**2**	0	0	0	**4**	**2**	**0**	**0**
-Right Var. Pairs	8	**4**	**0**	**4**	**0**	**0**	**0**	0	0
-Other	**3**	**11**	**20**	**0**	**17**	**12**	**15**	**13**	**9**

Table 1-2: LISREL & EQS at sample size = 2,000

TETRAD: Weight = 0.1, n = 200									
Model	1	2	3	4	5	6	7	8	9
Correct	0	0	5	2	17	6	0	18	19
-Width	-	-	3	1	2.9	2	-	13	3.1
Incorrect	20	20	15	18	3	14	20	2	1
-Overfit	0	0	0	0	1	0	0	0	0
-Underfit	18	0	0	17	0	14	20	0	0
-Other	2	20	15	1	2	0	0	2	1

TETRAD: Weight = 1.0, n = 200									
Model	1	2	3	4	5	6	7	8	9
Correct	2	2	5	17	8	13	1	9	16
-Width	4	2.5	2.4	1	3	2	12	13	3.1
Incorrect	18	18	15	3	12	7	19	11	4
-Overfit	0	0	6	2	7	2	0	3	0
-Underfit	9	1	0	0	0	5	10	0	0
-Other	9	17	9	1	5	2	9	8	4

LISREL, n = 200									
Model	1	2	3	4	5	6	7	8	9
Correct	3	9	0	4	0	3	0	4	0
Incorrect	17	11	20	16	20	17	20	16	20
-Overfit	0	1	3	5	0	1	2	1	0
-Underfit	4	3	0	0	2	4	2	0	1
-In Tetrad's Top Group	1	4	0	0	0	4	0	9	3
-Right Var. Pairs	2	0	0	8	0	0	0	0	0
-Other	10	3	17	3	18	8	16	6	16

Table I-3: TETRAD & LISREL at sample size = 200

Example of LISREL VI Input File for Starting Models 2 - 8

Starting Model for 2 - 8

```
DA NI=8 NO=2000 MA=CM
LABELS

•xl* 'x2' 'x3' 'x4' 'x5' 'x6' 'x7' 'x8'
CM

  1.64
  0.68  1.73
 -1.41 -1.47 4.13
 -0.58 -0.61 1.23 1.46
  0.68 -0.57 1.19 0.43 6.63
 -0.02  1.81  -3.78  -1.44-13.08 28.43
  0.94 0.98 -2.11 -0.77 -3.51 8.50 4.38
 -1.26  -1.40  2.80  1.12  4.72-11.46  -4.55  7.09
MO NK=1 NX=0 NE=9 NY=8 LX=ZE TE=ZE TD=ZE PS=SY,FI BE=FU,H GA=FU,FI
PABE

000000000
000000000
000000000
000000000
000000000
000000001
000000001
000000001
000000000
MA BE

000000000
000000000
000000000
000000000
000000001
00000000.5
00000000.5
00000000.5
000000000
MALY
*
100000000
010000000
001000000
000100000
000010000
000001000
000000100
000000010
```

```
FR PS(1,1) PS(2,2) PS(3,3) PS(4,4) PS(5,5) PS(6,6) PS(7,7) PS(8,8) PS(9,9)
STl.OPH(U)
MAPS
*
.5
0.5
00.5
000.5
0000.5
00000.5
000000.5
0000000.5
00000000.5
PAGA

01 1 100001
MAGA

1.5.5 .5 0 0 0 0.5
NFLY(1,1)-LY(8,9)
NFGA(1,1)BE(5,9)
OU AM TO MI
```

A Sample TETRAD Analysis

Consider the following model and data (the same as is in the above LISREL VI input file), taken from the first data set for Model 3 of our case study:

Tl=e9
T2 = -0.87 Tl + elO
xl=-0.78Tl+el
x2 = -0.82Tl+e2
x3 = 1.75Tl+e3
x4 = 0.69Tl+e4
x5 = 1.20x1+-1.91 T2 + e5
x6 = -1.41x5 +1.48 T2 + e6
x7 = 1.43 T2 + e7
x8 = -1.89T2 + e8

n = 2000

xl	x2	x3	x4	x5	x6	x7	x8
1.64							
0.68	1.73						
-1.41	-1.47	4.13					
-0.58	-0.61	1.23	1.46				
0.68	-0.57	1.19	0.43	6.63			
-0.02	1.81	-3.78	-1.44	-13.08	28.43		
0.94	0.98	-2.11	-0.77	-3.51	8.50	4.38	
-1.26	-1.40	2.80	1.12	4.72	-11.46	-4.55	7.09

The exogenous variables are distributed normally and the error terms have zero mean.

Suppose, erroneously, we start with the initial model

Tl=e9
T2 = bTl+elO
xl=alTl+el
x2 = a2 Tl + e2
x3 = a3 Tl + e3
x4 = a4 Tl + e4
x5 = a5 T2 + e5
x6 = a6 T2 + e6
x7 = a7 T2 + e7
x8 = a8 T2 + e8

where the a and b terms represent unknown coefficients. The initial model omits two causal connections that occur in the model that generated the data. How well can these connections be recovered from the data using TETRAD?

The procedure recommended in *Discovering Causal Structure,* which is meant for the published version of the TETRAD program, leads to two alternative models, including the correct model, in about ten minutes including the time to prepare input files. The LISREL VI program run on a Compaq 386, with the input file shown above, takes about 12 minutes and returns the wrong model.

After preparing TETRAD'S input file, which simply lists the covariance of each pair of measured variables, and gives the sample size, we create a file describing the graph of the initial model. It consists of a simple list of directed edges,

```
Tlxl
Tlx2
Tlx3
Tlx4
T1T2
T2x5
T2x6
T2x7
T2x8
```

Next we enter TETRAD and begin the search. The program is interactive, and from this point on we need not leave TETRAD until the search is completed. With the one command "suggested" (which prompts the user for the covariance file name, the initial graph file name, and the name of the file to which the output is to be written) we obtain TETRAD'S suggested trek additions:

```
The graph analyzed in this example is:
T1->x1 Tl->x2 Tl->x3 Tl->x4 T1->T2 T2->x5 T2->x6 T2->x7 T2->x8


The significance level is:    0.0500

The sample size is:  2000

Sets of suggested treks at significance level = 0.0000
-------------------------------------------------------

{xl-x5 xl-x6 x5-x6 Tl->x6 x7-x8 }
{xl-x5 xl-x6 x5-x6 Tl->x5 x7-x8 }
{xl-x5 xl-x6 x5-x6 Tl->x5 Tl->x6 }
{xl-x5 xl-x6 Tl->x5 Tl->x6 x7-x8 }
```

The output involves 6 different treks, two of them specified as directed edges from a latent to a measured variable. No other suggested trek additions are suggested until p > .8, so we consider only these. A further command gives the Rttr chart for the initial model. The parts of the chart that involve these six treks are shown below

```
   Edge        Rttr      D(I-H)    I(H-I)    Pi
-------------------------------------------------------
```

```
xl->x5      6.749    39     0     0.8002
x5-> xl     5.065    27    12     0.7520
xlCx5       4.256    21     0     0.7288


xl-> x6     5.093    39     0     0.7528
x6-> xl     4.218    27    12     0.7277
xlCx6       3.082    21     0     0.6952


x5-> x6     4.870    18     0     0.7464
x6-> x5     4.870    18     0     0.7464
x5Cx6       4.870    18     0     0.7464

Tl-> x5     4.641    24     0     0.7398

Tl-> x6     3.590    24     0     0.7097


x7-> x8     3.184    18     0     0.6981
x8-> x7     3.184    18     0     0.6981
x7Cx8       3.184    18     0     0.6981
```

Although the charts are lengthy, they are easily scanned. What matters are the second and fourth columns. Each number in column 2 measures how much adding the edge in column 1 to the initial model reduces the false implications of the model. Thus it is a measure of an increase in fit, and a large number indicates a better fit. Each number in column 4 is the number of equations that hold in the data that are no longer implied when the edge in column 1 is added to the initial model. Thus it is a measure of a decrease in explanatory power, and a small number is desirable.

Almost all of these connections preserve the explanatory power of the initial model (indicated by a 0 in column 4), but xl —> x5 and xl —> x6, in that order, do most to improve the fit (indicated by the largest numbers in column 2). Alternatively add each of these connections to the initial model (one command, "changegraph," for each case) and with a further command ("rttr") obtain the two Rttr charts for the two modifications of the initial model. In the first case, with xl — > x5 added to the initial model, the Rttr chart says that x5 —> x6 is the preferred further addition. In the second case, with xl —> x6 added to the initial model, the Rttr chart says that x6 —> x5 is the preferred further addition. The relevant parts of the two Rttr charts are:

```
The graph analyzed in this example is:
 xl->x5 Tl->xl Tl->x2 Tl->x3 Tl->x4 T1->T2 T2->x5 T2->x6 T2->x7
 T2->x8


The significance level is:    0.0500
The sample size is:  2000

Base Model:   Edges: 10   Fixed edges: 2   TTR:   5.3137

  Edge        Rttr     D(I-H)   I(H-I)   Pi
--------------------------------------------------


x5-> x6      4.834     30       0     0.8917
x6-> x5      2.390     10       0     0.8252
x5 C x6      2.390     10       0     0.8252

*******************************************************************

The graph analyzed in this example is:
 x1->x6 T1->x1 T1->x2 T1->x3 T1->x4 T1->T2 T2->x5 T2->x6 T2->x7
 T2->x8

The significance level is:   0.0500

The sample size is:  2000

Base Model:   Edges: 10   Fixed edges: 2   TTR:   6.9696

  Edge        Rttr     D(I-H)   I(H-I)   Pi
--------------------------------------------------


x5-> x6      2.985     10       0     0.7963
x6-> x5      6.490     30       0     0.8917
x5 C x6      2.985     10       0     0.7963
```

We thus obtain two elaborations of the initial model, and a further check shows that they are indistinguishably good explanations of the data according to TETRAD's measures.

The search now stops, because the Rttr chart for either of these models no longer suggest an x7--x8 trek, or any other connection. The fourth column in either of the resulting Rttr chart shows that any addition to the modified model will prevent the explanation of a number of tetrad equations, and do rather little to improve fit.

References

[I] Baumrind, D.
Specious Causal Attributions in the Social Sciences.
Journal ofPersonality and Social Psychology 45:1289 -1298,1983.

[2] Bender, P.
Theory and Implementation ofEQS: A Structural Equations Program.
BMDP Statistical Software Inc., 1964 Westwood Boulevar, Suite 202, Los Angeles, California 90025,1985.

[3] Blalock, H. M.
Causal Inferences in Nonexperimental Research.
The Univ. of North Carolina Press, Chapel Hill, North Carolina, 1961.

[4] Blalock, Hubert M, ed.
Causal Models in the Social Sciences.
Aldine Publishing Company, New York, 1971.

[5] Campbell, D., Schwartz, R, Sechrest, L., and Webb, E.
Unobtrusive Measures: Nonreactive Research in the Social Sciences.
Rand McNally, Chicago, 1966.

[6] Costner, H., Schoenberg, R.
Diagnosing Indicator Ills in Multiple Indicator Models.
In Goldberger, A., Duncan, O. (editors), *Structural Equation Models in the Social Sciences.* Seminar Press, New York, 1973.

[7] Costner, H. and Herring, J.
Respecification in Multiple Indicator Models.
In Blalock, H. (editor), *Causal Models in the Social Sciences: 2nd Edition,* pages 321-393. Aldine Publishing Co., New York, 1985.

[8] Duncan, O.
Introduction to Structural Equation Models.
Academic Press, New York, 1975.

[9] Fornell, C, and Larcker, D.
Evaluating Structural Equation Models with Unobservable Variables and Measurement Error.
Journal ofMarketing Research 18:39-50,1981.

[10] Fox,J.
Linear Statistical Models and Related Methods.
John Wiley and Sons, New York, 1984.

[II] Freedman, D. and Navidi, W.
Regression Models for Adjusting the 1980 Census.
Statistical Science, forthcoming.

[12] Glymour, C, Scheines, R., Spirtes, P., and Kelly, K.
Discovering Causal Structure.
^cademic Press, San Diego, CA, 1987.

[13] Goldberger, A., Duncan, O. (editors).
Structural Equation Models in the** Social **Sciences.
Seminar Press, New York, 1973.

[14] Harary, F., Norman R.f Cartwright, D.
Structural Models: An Introduction to the Theory ofDirected Graphs.
Wiley, New York, 1965.

[15] Heise, D.
Causal Analysis.
Wiley, New York, 1975.

[16] James, L., Mulaik, S., and Brett, J.
Causal Analysis: Assumptions, Models and Data.
Sage Publications, Beverley Hills, California, 1982.

[17] Joreskog, K. and Sorbom, D.
USREL VI User's Guide
third edition, Scientific Software, Inc., Mooresville, Indiana, 1984.

[18] Kohn,M.
Class and Conformity.
Dorsey Press, Homewood, II., 1969.

[19] Ling, Robert
Review of "Correlation and Causation" by David Kenny.
Journal of the American Statistical Association 77:489-491,1983.

[20] Long, J. Scott.
Quantitative Applications in the SocialSciences. Volume 34: *Covariance Structure Models.*
Sage Publications, 1983.

[21] MacCallum, R.
Specification Searches in Covariance Structure Modeling.
Psychological Bulletin 100:107-120,1986.

[22] Maruyama, G., and McGarvey, B.
Evaluating Causal Models: An Application of Maximum Likelihood Analysis of Structural Equations.
Psychological Bulletin 87:502-512,1980.

[23] McPherson, J.M, Welch, S., and Clark, C
The Stability and Reliability of Political Efficacy: Using Path Analysis to Test Alternative Models.
American Political Science Review 71:509-521, 1977.

[24] Scheines, R.
Automating Creativity.
In Fetzer, J. (editor), *Aspects ofArtificial Intelligence.* Kluwer Academic Publishers, Dordrecht, Holland, 1988.

[25] Simon, H.
Causal Ordering and Identifiability.
In *Models ofDiscovery,* pages 53-80. D. Reidel, Dordrecht, Holland, 1953.

[26] Simon, H.
Spurious Correlation: A Causal Interpretation.
In *Models ofDiscovery,* pages 93-106. D. Reidel, Dordrecht, Holland, 1954.

[27] Spirtes, P., and Glymour, C.
Latent Variables, Causal Models and Overidentifying Constraints.
Journal ofEconometrics -, forthcoming.

[28] Wheaton, B., Muthen, B., Alwin, D. and Summers, G.
Assessing Reliability and Stability in Panel Models.
In D. Heise (editor), *Sociological Methodology 1977,* pages 84-136. Jossey-Bass, San Francisco, 1977.

[29] Wishart, J.
Sampling Errors in the Theory of Two Factors.
British Journal ofPsychology 19:180-187,1928-29.