

**NOTICE WARNING CONCERNING COPYRIGHT RESTRICTIONS:**  
The copyright law of the United States (title 17, U.S. Code) governs the making of photocopies or other reproductions of copyrighted material. Any copying of this document without permission of its author may be prohibited by law.

# **Building Latent Variable Models**

by

**Richard Scheines**

**Peter Spirtes**

**Clark Glymour**

February 1991

Report No. CMU-PHIL-19



**Philosophy  
Methodology  
Logic**

**Pittsburgh, Pennsylvania 15213-3890**

# **Building Latent Variable Models<sup>1</sup>**

**Richard Scheines  
Peter Spirtes  
Clark Glymour**

Dept. of Philosophy  
Carnegie Mellon University

LCL Technical Report #91-2

---

<sup>1</sup>This work was funded by the Office of Naval Research contract number N00114-89-J-1964 and the Naval Personnel Research and Development Center. We thank Steve Sorensen, Jan Callahan, and Laurie McDonald for their many valuable suggestions.

## 1. Introduction

Researchers routinely face the problem of inferring causal relationships from large amounts of data, sometimes involving hundreds of variables. Often, it is the causal relationships between "latent" (unmeasured) variables that are of primary interest. The problem is how causal relationships between unmeasured variables can be inferred from measured data. For example, naval manpower researchers have been asked to infer the causal relations among psychological traits such as job satisfaction and job challenge from a data base in which neither trait is measured directly, but in which answers to interview questions are plausibly associated with each trait. By combining background knowledge with an algorithm that searches for causal structure among the unobserved variables, we have created a tool that can reliably extract useful causal information about latent variables from large data bases. In what follows we describe the class of causal models to which our techniques apply, the property that connects the causal structure of such models to measured data, the algorithm that searches for causal structures, its reliability and complexity, and simulation studies that attest to the algorithm's reliability on samples of realistic size.

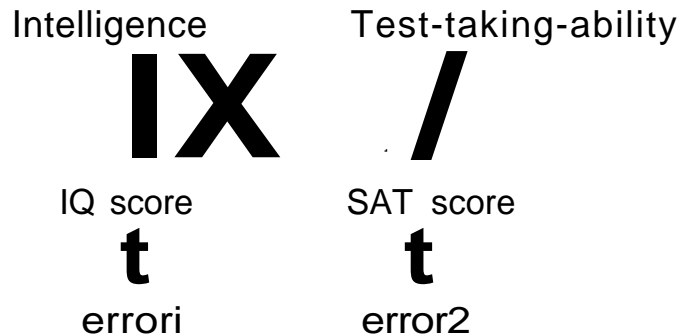
## 2. Causal Models

*2.1 Causal Graphs and Causal Models.* Economists, psychologists, sociologists, and political scientists routinely employ "structural equation models," or "causal models," to represent the causal structure among a set of random variables. These include regression models, factor analytic models, and path models.<sup>2</sup> These models typically are expressed as systems of linear equations among a set of random variables  $V$  along with an intended causal interpretation and distributional assumptions about  $V$ . The causal structure can be represented by a directed graph  $G$ , called a **causal graph**, such that there is an edge from variable  $V_1$  to  $V_2$  in  $G$  just in case  $V_1$  is a direct cause of  $V_2$ . For example, in the causal

---

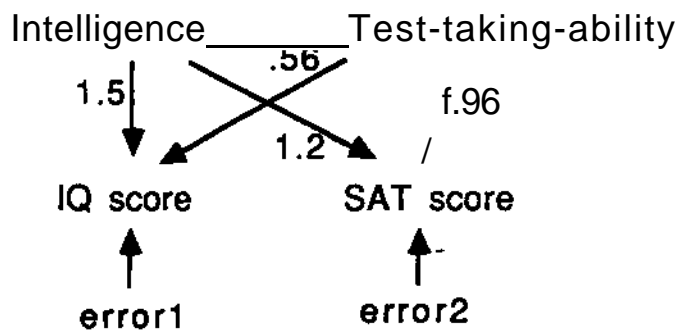
<sup>2</sup>See (Bollen 89) for a good introduction to "structural equation models."

graph in figure 1, both SAT and IQ scores are directly caused by intelligence and test-taking ability.



**Figure 1**

In structural equation models every effect is a linear combination of all of its direct causes. This assumption allows us to represent the equations in such models by attaching labels to the edges in the causal graph, where these labels represent the linear coefficients in the associated equations.



**Figure 2**

In figure 2, for example,  $\text{IQ score} = 1.5 \cdot \text{Intelligence} + .56 \cdot \text{Test-taking-ability} + \text{error1}$ . Let a **linear causal model** associated with a causal graph  $G$  be given by  $\langle D, \theta \rangle$ , where  $D$  is a distribution over the exogenous variables in  $G$  (those that are only causes and not effects) and  $\theta$  is a vector of linear coefficients that correspond to the appropriate edge

labels in  $G$ .<sup>3</sup> Given a causal graph  $G$ , the variance/covariance matrix  $I$  among the measured variables in  $G$  is completely determined by  $\langle D, e \rangle$ .

*2.2 Connecting the Evidence to the Causal Structure: Vanishing Tetrad Constraints.* There are, however, constraints on  $E$  that are determined just by the causal graph  $G$  associated with a causal model. That is, there are constraints on  $Z$  that are satisfied regardless of the value of  $\langle D, e \rangle$ . These constraints include **vanishing partial correlations** and **vanishing tetrad differences**, and they provide a connection between causal structure and measured data. For variables  $X$ ,  $Y$ ,  $W$ , and  $Z$ , there are three possible vanishing tetrad differences, any two of which are independent of each other:

$$P_{XY} * P_{WZ} - P_{XW} * P_{YZ} = 0$$

$$P_{XW} * P_{YZ} - P_{XZ} * P_{YW} = 0$$

$$P_{XZ} * P_{YW} - P_{XY} * P_{WZ} = 0$$

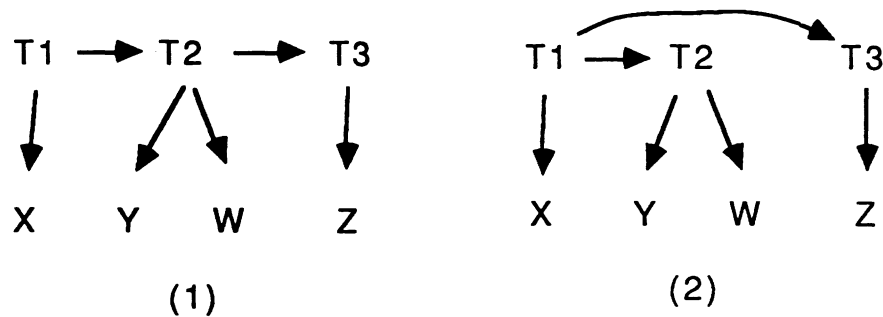
If the exogenous variables in  $D$  have non-zero variance, then we have a purely graphical characterization of the conditions under which a causal graph implies for all values of  $\theta$  a vanishing tetrad difference among measurable correlations.<sup>4</sup> The characterization allows us to calculate the set of tetrad equations implied by a causal graph extremely quickly. We search for the causal graph that implies the set of vanishing tetrad differences that most closely matches the set of vanishing tetrad differences judged to vanish in the population.<sup>5</sup>

---

<sup>3</sup>We also assume that each variable has a unique associated "error" variable of non-zero variance, that is all variables of non-zero indegree are the effect of a variable of indegree 0 and outdegree 1.

<sup>4</sup>See (Spirtes 89).

<sup>5</sup>The statistical tests we use to judge whether a tetrad difference vanishes in the population, and the metric that we use to measure how closely a set of tetrad differences implied to vanish match the set of tetrad differences judged to vanish in the population are described in detail in (Spirtes 91), and (Scheines 91). It is possible that a tetrad difference vanishes in the population for a given graph for some but not all values of  $\theta$ ; however if a graph does not entail that a given tetrad



**Figure 3**

Tetrad equations allow us to discriminate among different latent variable models where partial correlations do not.<sup>6</sup> For example, suppose that in graphs (1) and (2) in figure 3 only X, Y, Z, and Z are measured. Neither graph implies any vanishing partial correlations, but

graph 1 implies:

$$\rho_{X1,X3} * \rho_{X2,X4} = \rho_{X1,X4} * \rho_{X2,X3}$$

$$\rho_{X1,X4} * \rho_{X2,X3} = \rho_{X1,X2} * \rho_{X3,X4}$$

$$\rho_{X1,X3} * \rho_{X2,X4} = \rho_{X1,X2} * \rho_{X3,X4}$$

graph 2 implies:

$$\rho_{X1,X2} * \rho_{X3,X4} = \rho_{X1,X3} * \rho_{X2,X4}$$

If the results of statistical tests indicate that only  $\rho_{XW} * \rho_{YZ} = \rho_{XY} * \rho_{WZ}$  in the population, then graph 2 is a better model of the data, because it entails the only tetrad difference judged to vanish in the population, and it does not entail any vanishing tetrad differences that are judged not to hold in the population.

---

difference vanishes for all values of  $\theta$ , the Lebesgue measure of the set of  $\theta$ 's for which the tetrad does vanish is zero.

<sup>6</sup>For example, vanishing partial correlations. Although vanishing partial correlations of any order give us as good a connection to causal structure as can be had when all the common causes of measured variables are themselves measured (Spirtes 91), they cannot distinguish among the two graphs in Figure 3, unlike tetrad constraints.

### 3. The Strategy

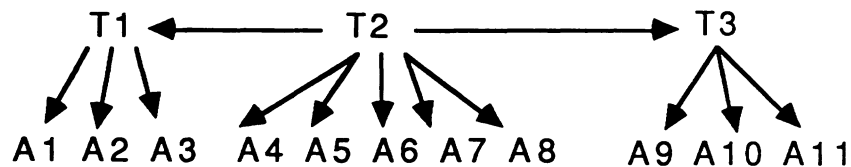
Structural equation models that involve latent variables are sometimes presented as containing two parts: the "measurement model", and the "structural model." The structural model involves only the causal connections among the latent variables, and the measurement model is the rest of the causal connections. Roughly, our strategy is to use certain tetrad equations to find a *pure measurement model*, i.e. one in which each measured variable is directly associated with only one latent. Having such a measurement model, we can then use different tetrad equations to constrain the structural model.

#### 3.1 Pure Latent Variable Models.

A model is a **pure latent variable model** if and only if

- i) each latent variable has at least two measured variables as its direct effects,
- ii) each measured variable is the cause of no variable, and
- iii) each measured variable is the direct effect of exactly one latent variable and a unique error term.<sup>7</sup>

For example:



**Figure 4**

where the *As* represent measured variables and the *Ts* represent unmeasured variables, and for convenience the error terms have been omitted. The important fact is that *in a pure latent variable model the*

<sup>7</sup>Models like these are also called "multiple indicator models."



*correlations and vanishing tetrad constraints partially determine the causal structure among the unmeasured variables.*

What if we are given a set of variables that do not form a pure latent variable model? It is possible that while a given set of variables do not form a pure latent variable model, some subset of the variables do. *It is possible to use vanishing tetrad differences to reliably select a subset of the variables that form a pure latent variable model (if one exists), even if the causal structure among the latent variables is unknown.* We then use vanishing tetrad constraints to partially determine the causal structure among the latent variables in the pure latent variable model. We will explain each of these procedures in more detail.

### *3.3 Finding Pure Latent Variable Models.*

A measurement model is pure just in case it is the measurement model of a pure latent variable model, i.e., each measured variable is not a cause and is the direct effect of exactly one latent variable and its own independent error term. In actual research the set  $V$  of measured variables in a data base is often chosen to measure a particular set of latent variables  $T$ . That is, for each latent variable  $T_i$ , a subset of  $V$  is intended to measure  $T_i$ . We suppose the investigator can, by whatever means, form mutually exclusive subsets of  $V$  such that each subset measures a latent variable  $T_i$ , i.e. the researcher can partition  $V$  into  $V_{T_i}$ , such that for each  $i$  the variables in  $V_{T_i}$  are direct effects of  $T_i$ .<sup>8</sup> The resulting model may not be a pure latent variable model, however, because some of the measured variables may be causes of other variables, or because some measured variables may be caused directly by more than one latent variable in  $T$ . To summarize, we assume that a researcher attempting to discover the causal relations among a set of latent variables  $T$  from measured variables  $X$ , can

- 1) Identify  $T$ , and

---

<sup>8</sup>That is, there is an edge from  $T_i$  to each  $V$  in  $V_{T_i}$ .

- 2) successfully partition the variables in  $X$  into groups that at least measure each  $T_j$ .

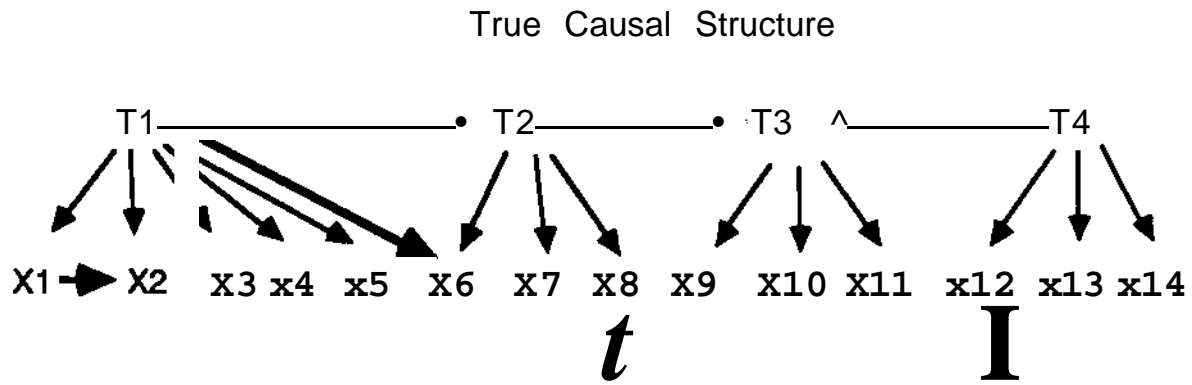


Figure 5

For example, suppose that the true causal graph  $G$  is the one depicted in figure 5. Then we assume a researcher can identify  $T = \{T_1, T_2, T_3, T_4\}$  and can successfully partition  $X = \{X_1 - X_{14}\}$  into clusters as we show in figure 6.<sup>9</sup>

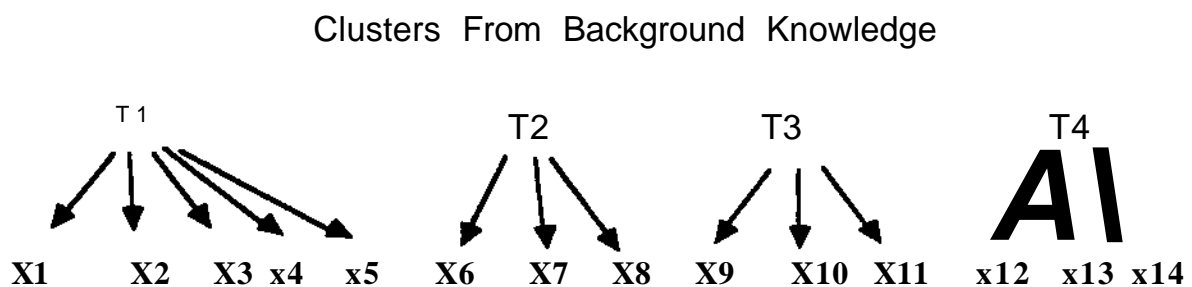
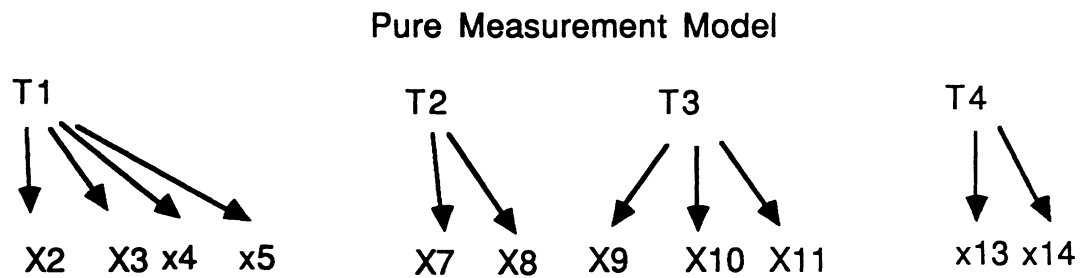


Figure 6

$G$  (figure 5) is not a pure latent variable model because  $X_1$  causes  $X_2$ ,  $X_6$  is caused by both  $T_1$  and  $T_2$ , and  $X_{12}$  causes  $x_8$ . Consequently, the set of measured variables  $X_1$  through  $X_{14}$  do not constitute a pure measurement model of  $T$ . There are subsets of  $\{X_1 - X_{14}\}$  that can constitute a pure measurement model of  $T$ , however, and figure 7 shows one.

<sup>9</sup>The clusters in figure 6 are not uniquely correct. Another correct initial clustering might group  $X_6$  with the  $T_1$  cluster.



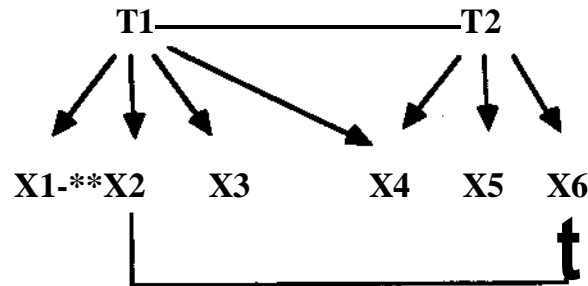
**Figure 7**

We have designed an algorithm that can reliably find a subset of variables that form a pure latent variable model, *regardless of what the (up to this point unknown) causal connections among the latent variables are*. We call this part of the procedure and a later part the SCALES procedure. SCALES eliminates those members of  $V_{T_i}$  that are impure measures of  $T_i$ , either because they are also the effects of some other unmeasured variable  $T_j$  or because they are also causes or effects of some other measured variable.

### *Impure Indicators*

A measured variable can be an impure measure for three reasons, which are exhaustive but not exclusive.

- 1) If  $V_i$  measures  $T_i$ , but is also causally connected to a latent variable  $T_j$  in some way not mediated by  $T_i$ , then we say that  $V_i$  is **latent-measured impure**.
- 2) If a pair of measured variables  $V_1, V_2$  from the same cluster  $V_{T_i}$  are causally connected in some way not mediated by  $T_i$  then we say  $V_1$  and  $V_2$  are **intra-construct impure**.
- 3) If a pair of measured variables  $V_1, V_2$  from distinct clusters  $V_{T_i}$  and  $V_{T_j}$  are causally connected in some way not mediated by either  $T_i$  or  $T_j$  then we say  $V_1$  and  $V_2$  are **cross-construct impure**.



**Figure 8**

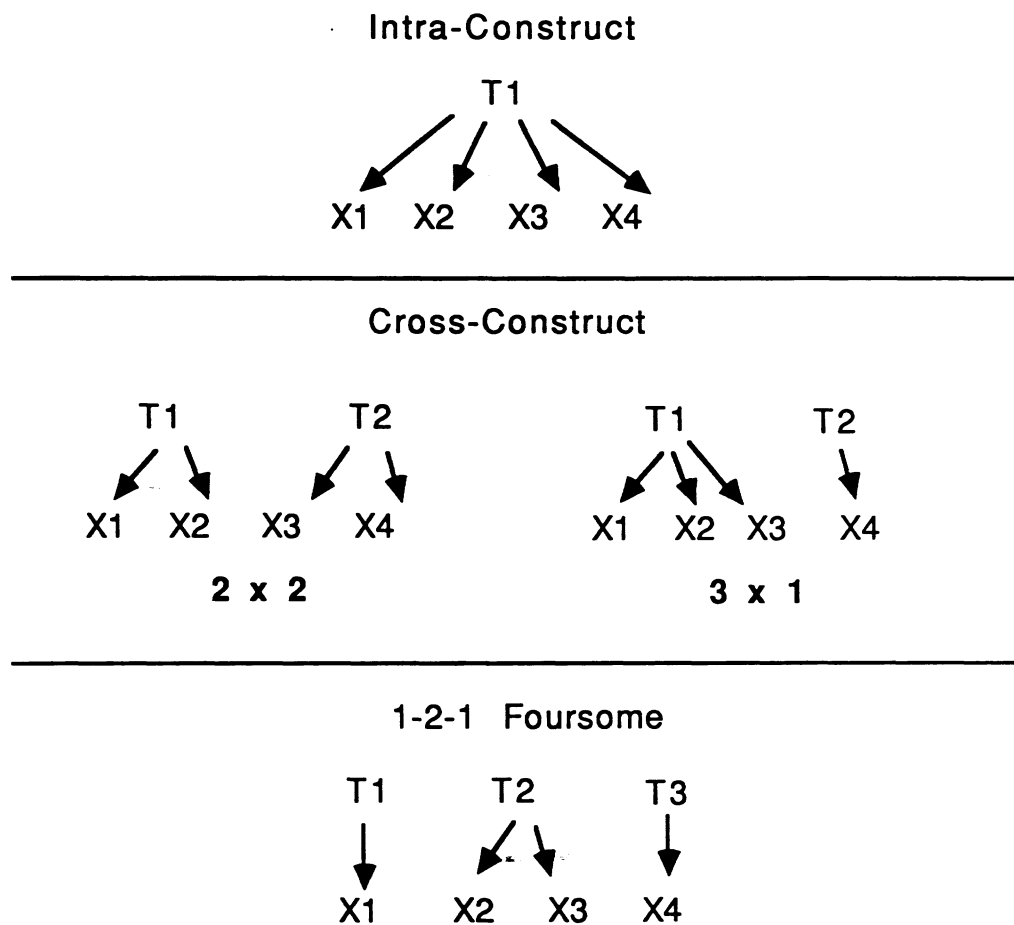
For example, in figure 8, X1, X2, and X6 are impure. The breakdown is as follows.

Intra-construct impure: { X1, X2}  
 Cross-construct impure: { X2, X6}  
 Latent-measured impure: { X4 }

The strategy for finding subsets of  $V$  that form a pure measurement model is to use tetrad equations of one type to eliminate those measured variables that are intra-construct impure, to use tetrad equations of a different type to eliminate those that are cross-construct impure, and finally those that are latent-measured impure. Provided the initial clusterings are correct, the strategy will correctly identify a subgraph that is a pure latent variable model (if one exists).

### *Foursomes*

Tetrad equations involve foursomes of measured variables. Based on the clustering from background knowledge: call a foursome of measured variables an **intra-construct foursome** if all of the measured variables are effects of the same latent variable (see Figure 9). Call a foursome in which at least two measured variables are measures of two different latent variables, and in which there are exactly two latent variables, a **cross-construct foursome**. Call a foursome in which one indicator is from one latent variable, two indicators are from a second latent variable, and another indicator is from a third latent variable a **1-2-1 foursome**.



**Figure 9**

Tetrad equations among intra-construct foursomes identify intra-construct impurities. Those among 2x2 cross-construct foursomes identify cross-construct impurities, and those among 3x1 cross-construct foursomes identify measured variables that are latent-measured impure.

*Tetrad Equations and Statistics*

Let  $T_p$  be a population tetrad difference, say:

$$\rho_{X1,X2} * \rho_{X3,X4} - \rho_{X1,X3} * \rho_{X2,X4}.$$

Let  $T_s$  be the corresponding tetrad difference in the sample. Let  $p(T) =$  the probability that a sample tetrad difference  $> T_s$ , given that  $T_p = 0$ . We calculate such a probability by using the Wishart statistic<sup>10</sup> for the variance of a tetrad difference under the assumption of multivariate normal distributed variables. Since the sample distribution of the tetrad difference is asymptotically normal, we can calculate  $p(T)$  straightforwardly. In the TETRAD II program, the decision to reject or accept the hypothesis  $T_p = 0$  is based on a user set significance level. In what follows, we heuristically assume that the higher  $p(T)$ , the more likely that  $T_p = 0$ . To capture this intuition, we assign a score<sup>11</sup> to each tetrad equation:

If  $p(T) >$  significance level, then Score = Score +  $p(T)$ ,  
 else Score = Score -  $(1 - p(T))$

### *Intra-Construct Foursomes*

An intra-construct foursome  $X_1$ - $X_4$  implies all three tetrad equations:

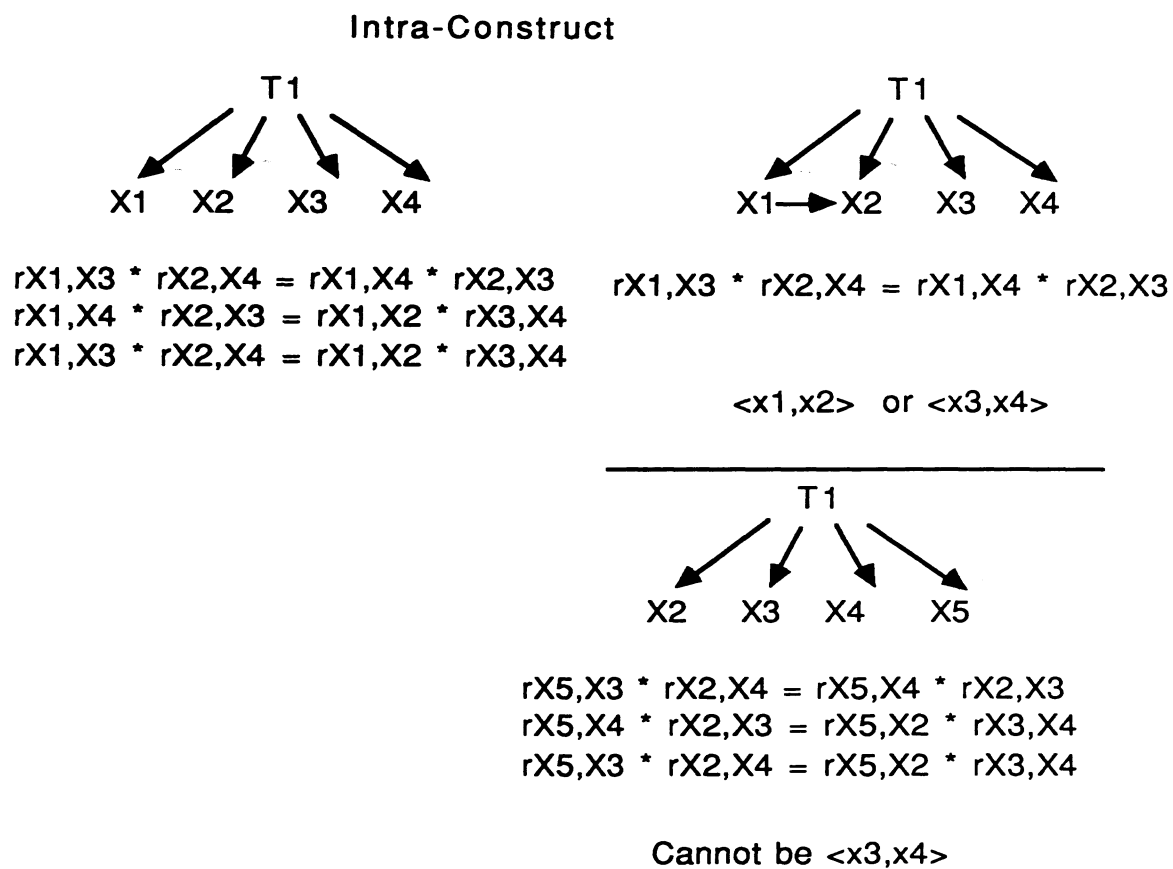
$$PX_{1.X2} * PX_{3.X4} = PX_{1.X3} * PX_{2.X4} = PX_{1.X4} * PX_{2.X3}$$

if at least three of  $X_1$ - $X_4$  are intra-construct impure. If a *pair* among  $X_1$ - $X_4$ , say  $\langle X_1, X_2 \rangle$  as in the top right of figure 10, is intra-construct impure, then the tetrad equations involving  $p_{xi, X_2}$  are not implied by the model. Unfortunately, both equations that involve  $p_{xi, X_2}$  also involve  $p_{x_3, X_4}$ , so we cannot, just from a single intra-construct foursome, distinguish which among two pairs is impure. Thus, in figure 10, we show in the top right the model that generates the data, the tetrad equation that implied by that model, and the two pairs  $\langle X_1, X_2 \rangle$  and  $\langle X_3, X_4 \rangle$  that we can postulate as candidates for intra-construct impure measures.

---

<sup>10</sup>See(Glymour87).

<sup>11</sup> For more detail on the scoring function, see (Spirtes 90a)



**Figure 10**

We can distinguish which pair is impure by checking other intra-construct tetrad equations that involve one pair and not the other. For example, if we check the foursome  $\{X2, X3, X4, X5\}$ , then only the pair  $\langle X3, X4 \rangle$  is involved. If the two tetrad equations in this foursome involving  $\rho_{X3, X4}$  hold, then we cannot eliminate  $\langle X3, X4 \rangle$  from consideration. In essence, we use redundancy to overcome the underdetermination unavoidable from single foursomes. The algorithm for eliminating intra-construct impurities is as follows.

Repeat

For each indicator  $V_i$  in  $V_{T_i}$ ,

For each intra-construct foursome  $F$  involving  $V_i$ ,

For each tetrad equation  $T$  in  $F$ ,

If  $p(T) > \text{significance level}$ , then  $\text{Score} = \text{Score} + p(T)$ ,

else  $Score_{Vi} = Score_{Vi} - (1 - p(T))$ .

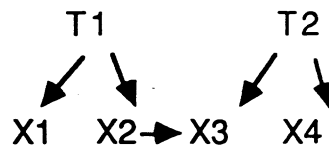
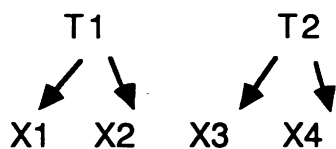
$V_{Ti} := V_{Ti} - \{Vi \text{ with lowest score}\}$ .

Until  $(\text{Min}(Score_{Vi}) > \text{User set cutoff}) \text{ Or } |V_{Ti}| \leq 4$

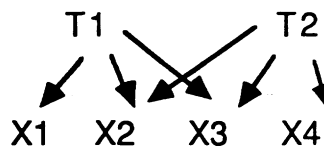
*Cross-Construct Foursomes*

*2x2 Foursomes*

Since intra-construct foursomes involve only one latent variable, the causal connections among the latent variables has no effect on this part of the algorithm. Cross-construct foursomes involve two latent variables, say T1 and T2. Luckily, the causal structure between T1 and T2 has no effect on the tetrad equations applied among cross-construct foursomes involving T1 and T2.



$$r_{X1,X3} * r_{X2,X4} = r_{X1,X4} * r_{X2,X3}$$



<x1,x3> or <x2,x4> or  
<x1,x4> or <x2,x3>

**Figure 11**

In the left side of figure 11 we show a 2x2 cross-construct foursome from a pure measurement model, and under it the tetrad equation it implies regardless of the nature of the causal connection between T1 and T2. This equation is not implied by a model in which any of the four possible cross-construct pairs are impure. On the right side we show two



ways in which one of the cross-construct pairs,  $\langle X_2, X_3 \rangle$ , is impure. Below the models on the right we show what we can learn from this foursome if either of these models is correct. Again, to narrow the blame we look at overlapping foursomes. The algorithm for 2x2 foursomes is as follows.

Repeat

For each indicator  $V_i$  in  $V$ ,

For each 2x2 cross-construct foursome  $F$  involving  $V_i$ ,

For each tetrad equation  $T$  in  $F$ ,

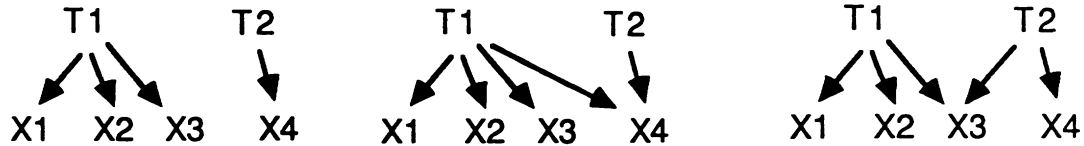
If  $p(T) > \text{significance level}$ , then  $\text{Score} = \text{Score} + p(T)$ ,  
 else  $\text{Score}_{vi} = \text{Score}_{vi} - (1 - p(T))$ .

$V := V - \{V_i \text{ with lowest score}\}$ , provided no  $|V_{ji}| \leq 3$ .

Until  $(\text{Min}(\text{Score}_{vi})) > \text{User set cutoff}$

### 3x1 Foursomes

After the 2x2 procedure, we will have eliminated all cross-construct impurities except for a single latent-measured impurity, if there is one. This is because a 2x2 foursome implies one tetrad equation if it is pure or if exactly one of its indicators is latent-measured impure. In the 3x1 cross-construct foursome the situation is better. Whereas the two measurement models on the left side of figure 12 imply all three tetrad equations no matter what the causal connection between  $T_1$  and  $T_2$ , the model on the right implies only one. The difference is that in the middle model the indicator on the 1 side of a 3x1 (the singleton) is latent-measured impure, whereas on the right one of the indicators on the 3 side of a 3x1 is impure. In the latter case the tetrad equations involving a correlation between the impure indicator and the singleton are not implied by the model. Thus if one equation holds, we can uniquely identify the impure indicator.



$$\begin{aligned}
 r_{X1,X3} * r_{X2,X4} &= r_{X1,X4} * r_{X2,X3} \\
 r_{X1,X4} * r_{X2,X3} &= r_{X1,X2} * r_{X3,X4} \\
 r_{X1,X3} * r_{X2,X4} &= r_{X1,X2} * r_{X3,X4}
 \end{aligned}$$

$$\begin{aligned}
 r_{X1,X3} * r_{X2,X4} &= r_{X1,X4} * r_{X2,X3} \\
 &<T2 \rightarrow x3>
 \end{aligned}$$

**Figure 12**

Again, however, we use a strategy of redundancy to detect latent-measured impurities. The algorithm for 3x1 foursomes is as follows.

Repeat

For each indicator  $V_i$  in  $V$ ,

For each 3x1 cross-construct foursome  $F$  involving  $V_i$ ,

For each tetrad equation  $T$  in  $F$ ,

If  $p(T) >$  significance level, then  $\text{Score} = \text{Score} + p(T)$ ,

else  $\text{Score}_{V_i} = \text{Score}_{V_i} - (1 - p(T))$ .

$V := V - \{V_i \text{ with lowest score}\}$ , provided no  $|V_{T_i}| \leq 2$ .

Until  $(\text{Min}(\text{Score}_{V_i}) > \text{User set cutoff})$

After the 3x1 procedure we have a subset of  $V$  that constitutes a pure measurement model. We can now search for facts about the causal connection among the latent variables.

### 3.4 Finding the Structural Model.

If a set of measured variables  $V$  is causally sufficient, i.e., if every common cause of two variables in  $V$  is also in  $V$ , then  $V_1$  and  $V_2$  (in  $V$ ) are causally adjacent if and only if  $V_1$  and  $V_2$  are dependent conditional on every subset in  $V$  that doesn't include  $V_1$  or  $V_2$ . If we begin with an undirected graph in which every pair  $V_1$  and  $V_2$  is adjacent, as soon as we

find a conditioning set  $S$  such that  $V_1$  and  $V_2$  are independent conditional on  $S$ , we can remove the undirected edge between  $V_1$  and  $V_2$ .

If a pair  $\langle V_1, V_2 \rangle$  is independent conditional on a set  $S$ , and the cardinality of  $S$  is  $n$ , we say that  $V_1$  and  $V_2$  are  $n$ th-order independent. In the PC-algorithm,<sup>12</sup> the strategy is to begin by removing all edges that connect pairs that are 0-order independent, ascend to sets of size 1, size 2, etc.<sup>13</sup> The current strategy is to use PC on the latent variables, but instead of providing with facts about conditional independence, we provide facts about 0-order and  $n$ th-order *trek-separability*, which we can determine in pure latent variable models respectively from the correlations between indicators of distinct latent variables and from the tetrad equations that hold in 1-2-1 foursomes.

### 3.4.1 *Trek Separability.*

Let a **path** between vertices  $X$  and  $Y$  in an graph  $G$  be a sequence of edges from  $X$  to  $Y$ .

Let a **trek** between  $X$  and  $Y$  be either an acyclic path from  $X$  to  $Y$ , an acyclic path from  $Y$  to  $X$ , or a pair of acyclic paths from some other variable  $Z$  to  $X$  and to  $Y$  such that these paths intersect only at  $Z$ .

$V_1$  and  $V_2$  are  **$n$ th-order trek-separated** just in case there are  $n$  vertices  $V_j \dots V_n$  \*  $V_1, V_2$ , and no set smaller than  $V_j \dots V_n$  such that each trek between  $V_1$  and  $V_2$  includes some member of  $V_j \dots V_n$ .

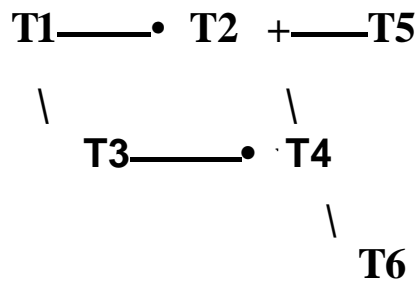
---

<sup>12</sup>See(Spirtes91).

<sup>13</sup>The number of  $n$ th-order independence facts PC needs to check is constrained not only by the edges removed in the  $n-1$ th order stage, but by the variables that are neighbors. For details, see (spirtes91).

**V1** and **V2** are **O-order trek-separated** just in case there is no trek connecting them.

**V1** and **V2** are **1st-order trek-separated** just in case there is a vertex **V3** such that all treks between **V1** and **V2** pass through **V3**.



**Figure 13**

For example, in the model in figure 13 the following trek-separability facts hold.

O-order trek-separated:

{ <T1,T5> <T3,T5> <T4,T5> <T6,T5> }

1st-order trek-separated:

{ <T1,T4 by T3> <T1,T6 by T4 or T3> <T2,T6 by T4> <T3,T6 by T4> }

2nd-Order trek-separated:

{ <T2,T3 by {T1 ,T4}> }

The SCALES program finds a pure latent variable model and with it determines the O-order and 1st-order trek separability facts. We use its output as input to a slight modification of the PC algorithm:

### **Latent-PC**

Begin with a complete undirected graph **G** among **T**.

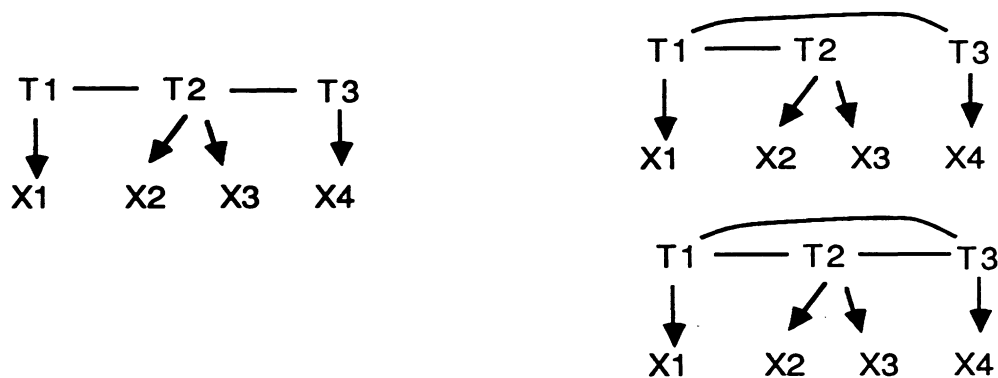
For each pair  $\langle T_i, T_j \rangle$  that are 0-order trek separated, remove the edge  $T_i-T_j$  in  $G$ .

For each pair  $T_i, T_j$  that still have an edge between them in  $G$ , remove  $T_i-T_j$  are 1st-order trek separated by some  $T_k$ ,  $k \neq i, j$ .

For each triple  $T_i, T_j, T_k$ , such that  $T_i-T_j$  and  $T_j-T_k$  and not  $T_i-T_k$ , direct  $T_i \rightarrow T_j$  and  $T_k \rightarrow T_j$  if  $T_i$  and  $T_k$  are 0-order trek separated.

The question now reduces to how we can determine facts about 0-order and 1st-order trek-separability relations. There is an obvious method for determining 0 order trek separability facts among latent variables. Two measured variables are uncorrelated in a pure latent variable model if and only if they are effects of distinct latent variables that are not trek connected. In figure 5,  $T_2$  is 0-order trek separated from  $T_4$  because  $\rho_{X_6 X_{13}} = \rho_{X_6 X_{14}} = \rho_{X_7 X_{13}} = \rho_{X_7 X_{14}} = 0$ . In practice, we consider all the correlation between an indicator from one latent and an indicator from another. If the majority of these correlations are insignificant, then we judge the pair of latents to be 0-order trek separated.

Tetrad equations among 1-2-1 foursomes in a pure measurement model identify 1st order trek-separability relations. For example, on the left side of figure 14 we show a 1-2-1 foursome in which  $T_2$  1st-order trek separates  $T_1$  and  $T_3$ . This structure and only this structure implies all three tetrad equations among this foursome. Assuming that  $T_1, T_2$  and  $T_3$  are trek connected, which we have already determined from the 0-order trek separability procedure, any other causal arrangement among the three latents will imply only 1 tetrad equation, as we show on the right side of figure 14.



$$\begin{aligned}
 r_{X1,X3} * r_{X2,X4} &= r_{X1,X4} * r_{X2,X3} \\
 r_{X1,X4} * r_{X2,X3} &= r_{X1,X2} * r_{X3,X4} & r_{X1,X4} * r_{X2,X3} &= r_{X1,X2} * r_{X3,X4} \\
 r_{X1,X3} * r_{X2,X4} &= r_{X1,X2} * r_{X3,X4}
 \end{aligned}$$

**Figure 14**

In order to determine whether a latent variable T2 1st-order trek separates T1 and T3, we again exploit the redundancy inherent in a multiple-indicator model.

For each 1-2-1 foursome F in which T2 is the parent of the twosome and T1 and T3 are parents of the singeltons,

For both tetrad equations T in F implied only by a model in which T2 trek-separates T1 and T3,

If  $p(T) >$  user set significance level, then  $\text{Score} := \text{Score} + p(T)$   
 else  $\text{Score} := \text{Score} - (1 - p(T))$ .

If  $\text{Score} > 0$ , then T2 trek separates T1 and T3.

If, for a given triple of latent variables, we discover that the score for more than one 1st-order trek separation relation is greater than 0, we take only the relation with the largest score to be true, thus we avoid inconsistency.

#### **4. Reliability**

There are two issues that bear on the full algorithm's reliability. First we can consider PC-latents reliability, i.e. the full algorithm's reliability if the facts about 0-order and 1st-order trek-separability latent-PC takes as input are correct, and second we can consider its reliability in determining these facts, i.e. the reliability of SCALES.

#### *4.1 Latent-PC Reliability*

First suppose that latent-PC receives correct 0-order and 1st-order trek separability facts. Two variables  $V_i$  and  $V_j$  are adjacent if and only if  $V_i$  is a direct cause of  $V_j$  or vice versa. The full set of trek-separability facts determine the full set of adjacencies. A subset of the trek-separability facts determine a superset of the adjacencies. Because we can now only determine 0-order and 1st-order trek separability facts, the algorithm can at best identify all the correct adjacencies, but might also identify some that do not exist. It may well be that non-quadratic constraints on correlations determine further higher order trek separability relations among the unmeasured variables, but it is now an open question.

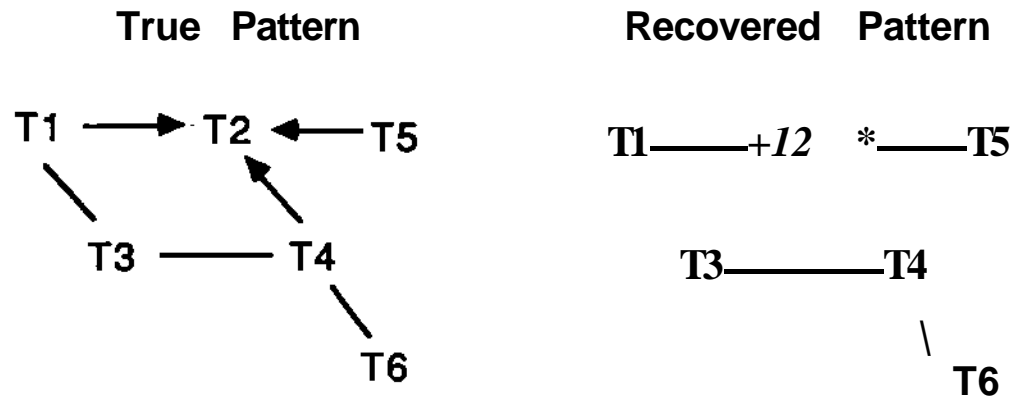
The test for ordering we perform in latent-PC is strictly weaker than the full PC test. Thus, latent-PC might fail to order an edge that the full PC algorithm could order, but it will never order an edge that the full PC algorithm would leave undirected.

For an example of what the latent-PC algorithm could recover with the correct 0-order and 1st-order trek separability facts, consider figure 15. Let a **pattern** be a graph in which some edges are directed and some are not. Both graphs in figure 15 are displayed as patterns. An edge is undirected in a pattern if the full PC algorithm for ordering cannot order it.<sup>14</sup>

---

<sup>14</sup>See (Spirtes 91)

E.g.,



**Figure 15**

Latent-PC would recover all the adjacencies, but add one that does not exist in the true pattern, T2-T3. It does so because  $\langle T2, T3 \rangle$  are 2nd-order trek-separated by  $\{T1, T4\}$ , but latent-PC cannot obtain such information and would thus leave in the edge.

#### 4.2 SCALES Reliability

The reliability of the SCALES procedure is less well understood.

If

- 1) the statistical assumptions are satisfied, and
- 2) the clustering information is correct, and
- 3) there is a measurement model of the true model such that there are at least two pure indicators connected to each latent, and
- 4) there is no sampling error,
- 5) no tetrad equations hold in the population that are not implied by the causal structure,

then,

there exist cutoffs such that SCALES will output the correct 0-order and 1st-order trek separability facts.



We do not know exactly what these cutoffs are, and we are currently performing monte carlo simulation tests to optimize them.

## 5. Complexity.

The algorithm's complexity is determined by the number of tetrad differences it must check, which is determined by how many foursomes of variables there are. If there are  $n$  measured variables the total number of foursomes is  $O(n^4)$ . We don't check each possible foursome, however, and the actual complexity depends on the number of latent variables and how many variables measure each latent. If there are  $m$  latent variables and  $s$  measured variables for each, then the number of foursomes is  $O(m * s^4)$ . Under these assumptions, the number of intra-construct foursomes is exactly  $m*s^4$ . The number of 2x2 foursomes is  $s-1 * \binom{s}{2} * m$ . The number of 3x1 foursomes is  $m * \binom{s}{1} + s*m*\binom{s-1}{1}$ . Thus the order of the calculation is  $m*s^4$ . Since  $m*s = n$ , this is  $O(n * s^3)$ , which is much lower than  $O(n^4)$  if  $s \ll n$ .

We have implemented the algorithm in Pascal on a Decstation 3100 and tested it on graphs with up to six latent variables and 50 measured variables. It runs in under 20 seconds on graphs of such size, and we are thus confident that it is feasible to construct models with well over 100 measured variables.

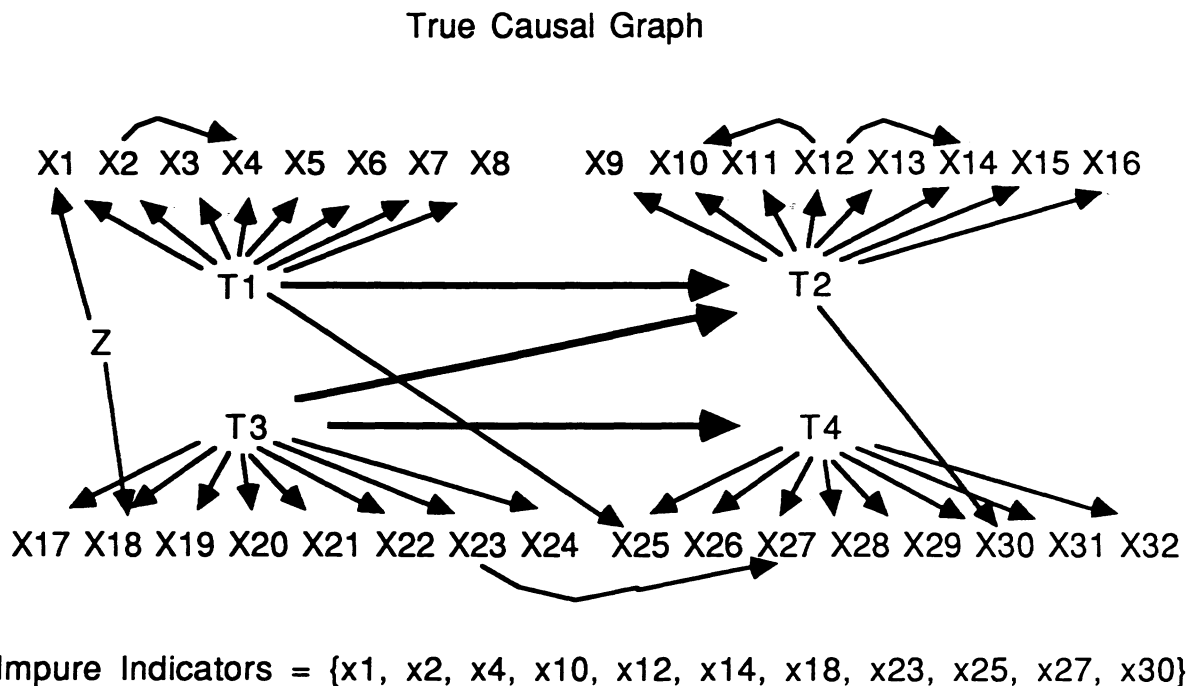
## 6. Simulation Studies

The algorithm has already been used to extract causal information from real data bases. Naval manpower researchers used an earlier version of the algorithm to build latent variable models from a data base that includes more than 200 variables and 7,000 observations. They were able to find several plausible pure latent variable models that passed a statistical goodness of fit test, even on a sample size powerful enough to detect very small failures in any of the assumptions underlying the model.

Pure latent variable models that pass a chi-square test on samples of this size are almost unheard of.

Decisions throughout the SCALES procedure rest on simultaneously testing individual tetrad equations that are not independent. Simultaneous inference is a well known statistical problem, but solutions to it that we might apply are still forthcoming. Our best evidence about its reliability on realistic data can at present only come from Monte Carlo studies.

We conducted simulation studies on the causal graph in figure 9, which has 10 impure indicators out of a possible 32.



**Figure 16**

We set the distribution for the exogenous variables to be standard multivariate normal (all variables have mean 0 and variance 1), and chose the linear coefficients randomly to be inbetween .5 and 1.5. We generated data from this model for each unit in the sample by pseudo-randomly sampling to produce values for the exogenous variables, and then by using

the linear equations to fill in the values for the rest of the variables. We recorded only data for the measured variables, i.e. those that begin with X.

We conducted 20 trials at sample sizes of 100, 500, and 2000. We counted errors of commission and errors of omission for 0 order and 1st order trek separability. In each case we counted how many errors the procedure could have made and how many it actually made. We also give the number of samples in which the algorithm identified the trek separability facts perfectly. At sample size 2,000, the algorithm was literally perfect.

Sample Size	0 order		1st Order		Number Perfect
	Commission	Omission	Commission	Omission	
100	2/80	0/40	7/220	1/20	13/20
500	1/80	0/40	2/220	0/20	19/20
2000	0	0	0	0	20/20

In a more substantial study, we used the base model we show in figure 17, and added causal connections to it randomly that make indicators impure.

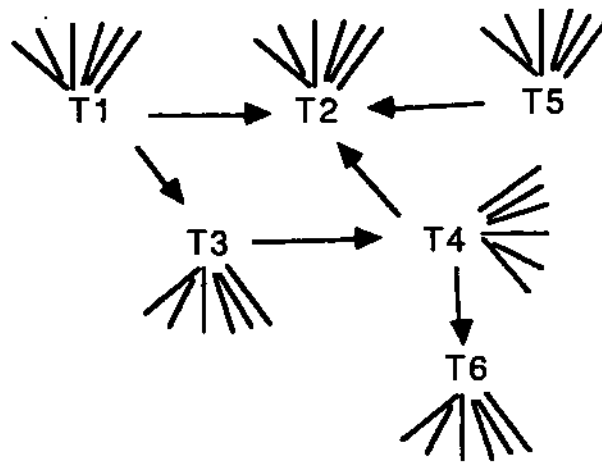
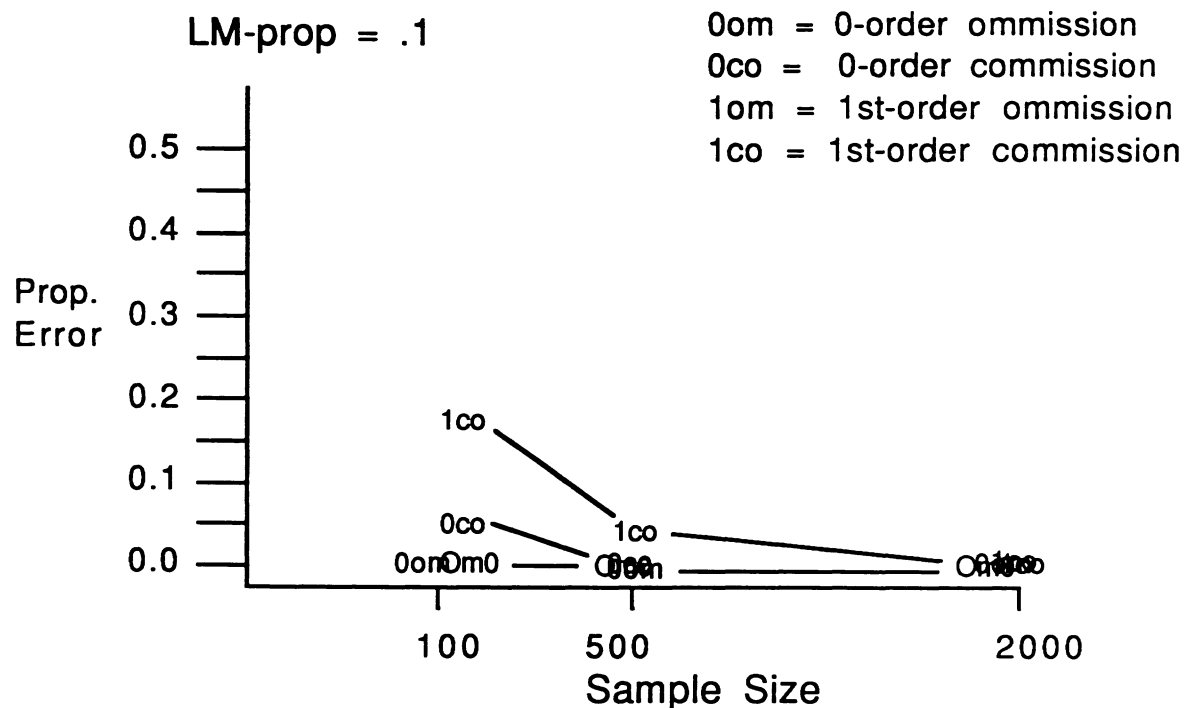


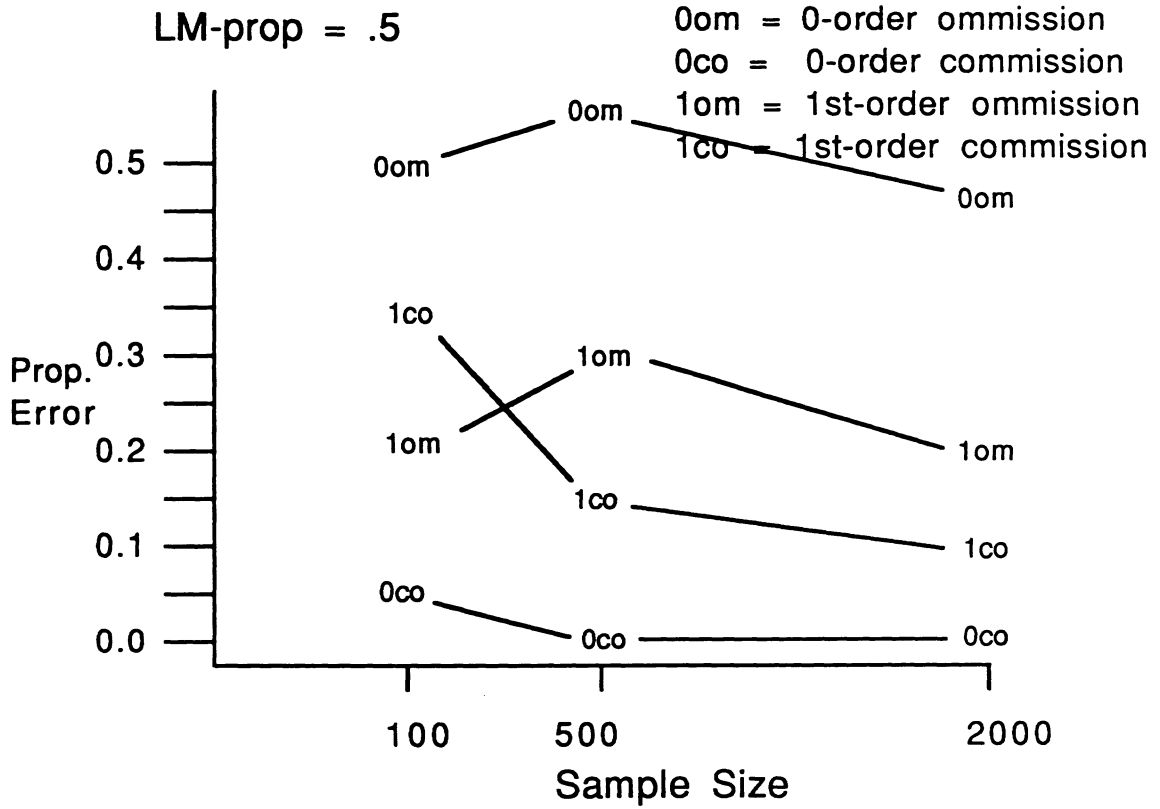
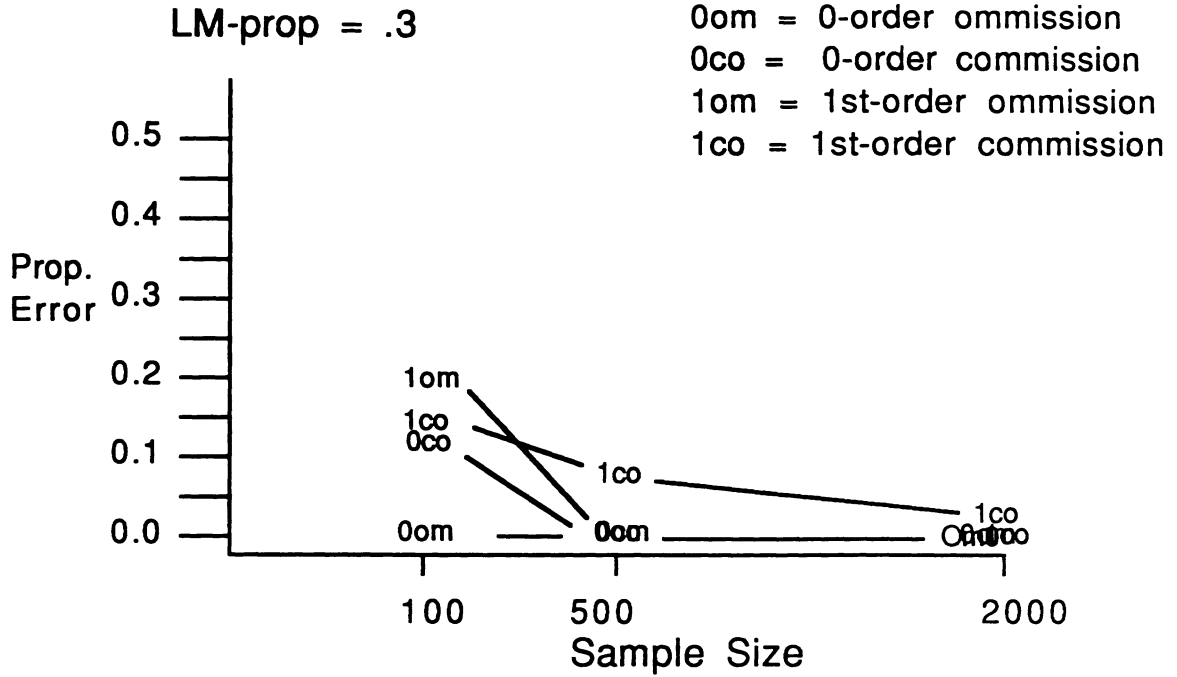
Figure 17

We varied three parameters, the proportion of indicators are latent-measured impure (lm-prop), the proportion of indicators that are measured-measured impure (mm-prop), and the sample size. For this study we kept the cutoff fixed arbitrarily at .25. In the first study we kept mm-prop at 0 and varied only the lm-prop and the sample size. We show the results graphically below. The behavior of the algorithm is very reliable at sample sizes above 500 and at levels of impurity below .5.

### Study 1

- 1) LM-prop = (.1 .3 .5)
- 2) MM-prop = 0
- 3) Sample size = (100 500 2000)
- 4) Cutoff criterion = .25

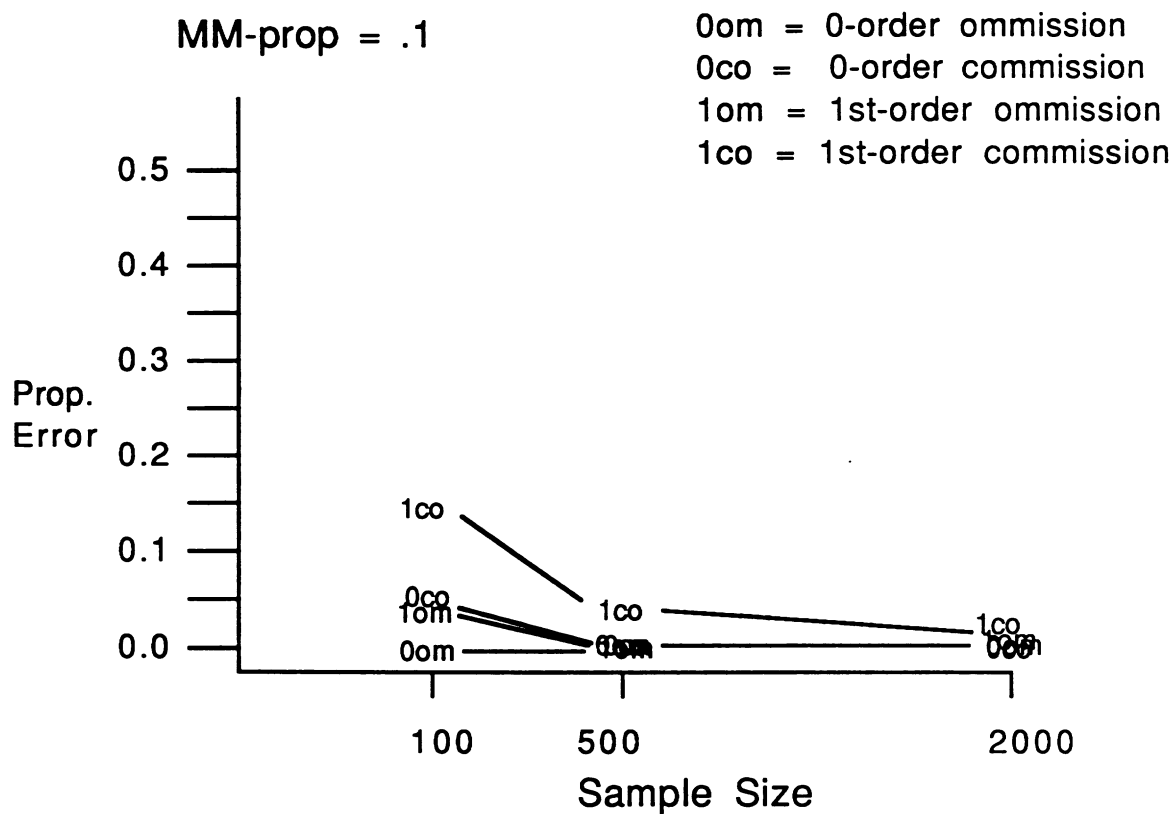


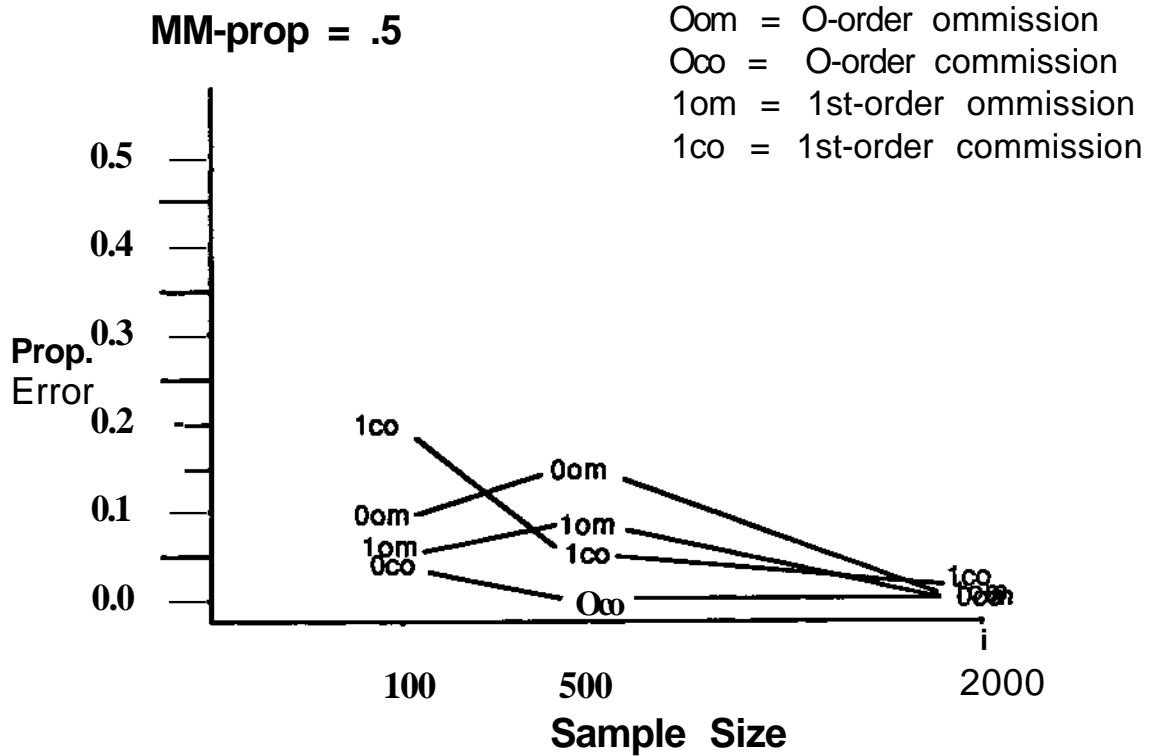
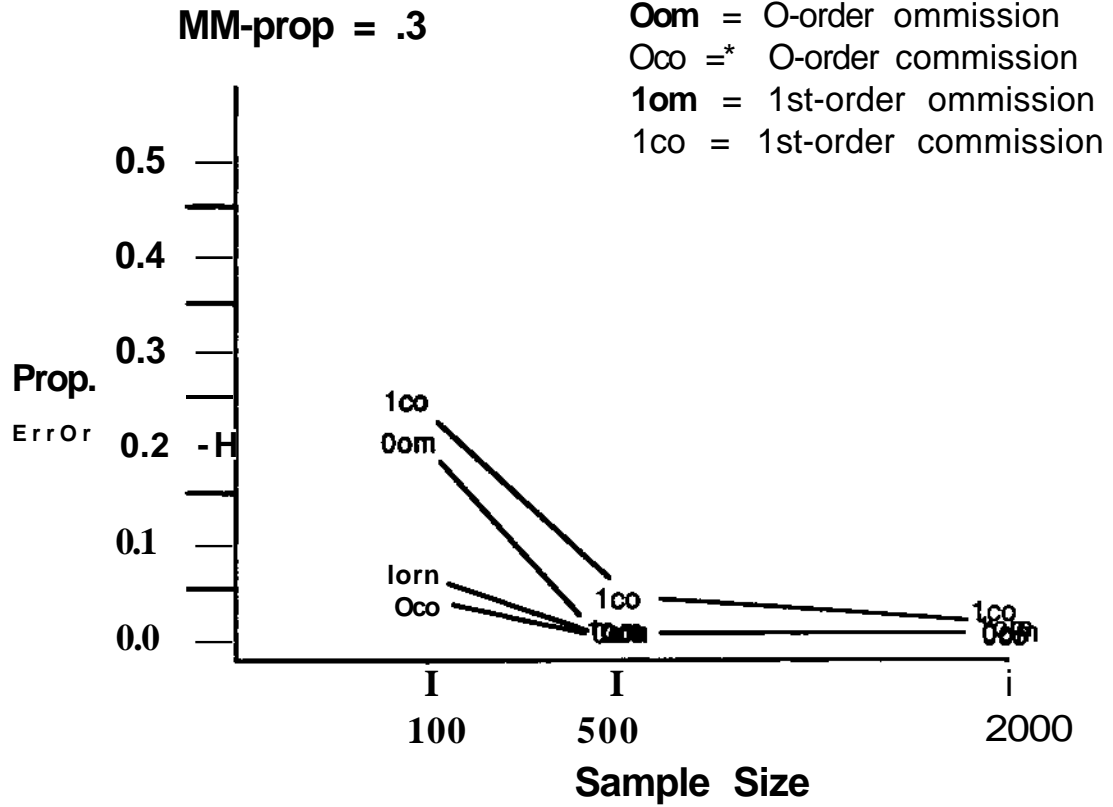


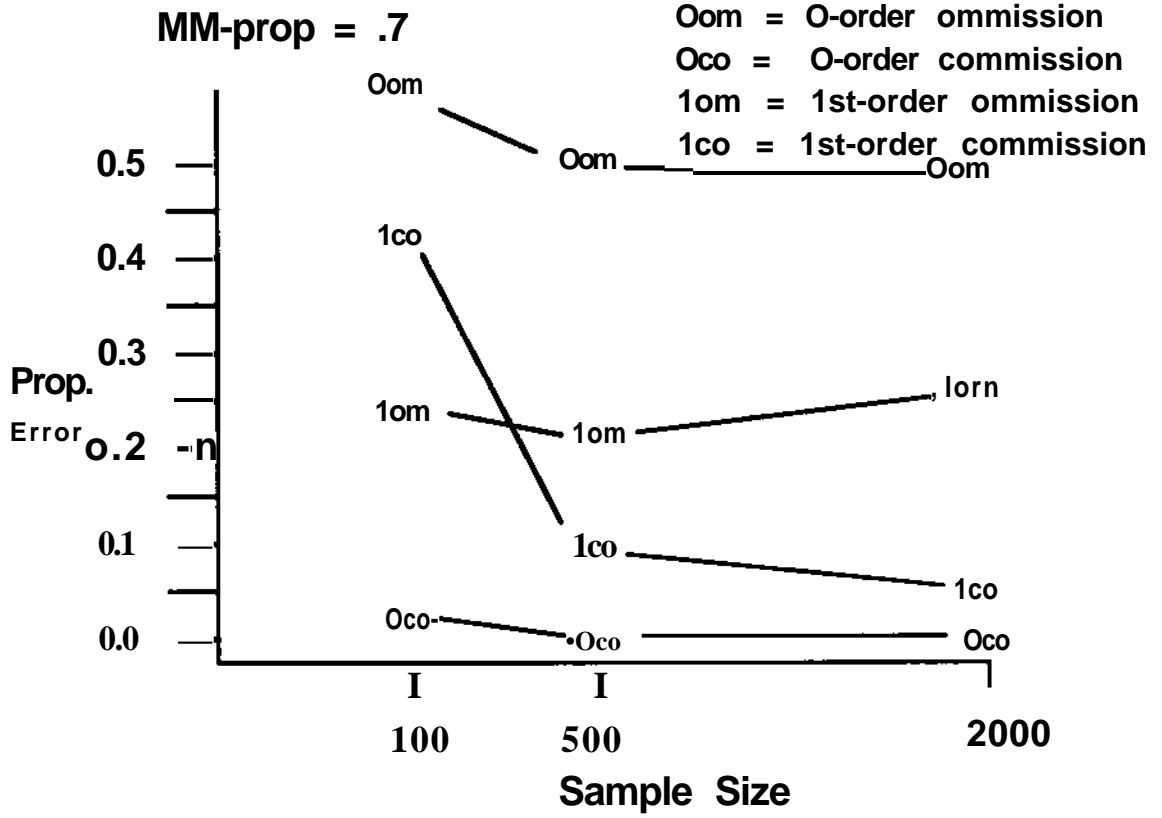
## Study 2

In the second study we kept  $lm\text{-prop}$  at 0 and varied  $mm\text{-prop}$  and the sample size. Again, the performance is quite good at sample sizes over 500 and at levels of  $mm\text{-prop}$  impurity below .5.

- 1)  $LM\text{-prop} = 0$
- 2)  $MM\text{-prop} = (.1 .3 .5 .7)$
- 3) Sample size = (100 500 2000)
- 4) Cutoff criterion = .25









## References

Bollen, K. (1989). *Structural Equations with Latent Variables*, Wiley.

Glymour, C, Scheines, R., Spirtes, P., and Kelly, K. (1987). *Discovering Causal Structure*, Academic Press, San Diego, California,.

Glymour, C, Scheines, R., Spirtes, P. (1989). "Why Aviators Leave the Navy: Applications of Artificial Intelligence Procedures in Manpower Research," *Report to the Naval Personnel Research Development Center*, January.

(Scheines, R., (1991), "Building Latent Variable Models," LCL-Technical/ Report 91-2, Carnegie Mellon University, Pittsburgh, PA.)

Scheines, R., Spirtes, P., Glymour, G., and Sorensen, S. (1990). "Causes of Success and Satisfaction Among Naval Recruiters," *Report to the Naval Personnel Research Development Center*, July.

Spirtes, P. (1989). "Fast Geometrical Calculations of Overidentifying Constraints", Carnegie-Mellon University Laboratory for Computational Linguistics Technical Report CMU-LCL-89-3.

Spirtes, P., Glymour, C, and Scheines, R., and Sorensen, S. (1990). TETRAD Studies of Data for Naval Air Traffic Controller Trainees, *Report to the Naval Personnel Research Development Center*, January.

Spirtes, P., Glymour, C, and Scheines, R., (1991). "Causality, Statistics, and Search." Manuscript submitted for publication.

Spirtes, P., Scheines, R., and Glymour, C. (1990a). "Simulation Studies of the Reliability of Computer Aided Specification Using the TETRAD II, EQS, and LISREL Programs", *Sociological Methods and Research*, pp. 3-60.