# AD-A231 273

(12)

SECURI

## REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-[illegible]

| | |
|---|---|
| 1a REPORT SECURITY CLASSIFICATION: Unclassified | 1b RESTRICTIVE MARKINGS |
| 2a SECURITY CLASSIFICATION AUTHORITY | 3 DISTRIBUTION / AVAILABILITY OF REPORT: Approved for public release; distribution unlimited |
| 2b DECLASSIFICATION / DOWNGRADING SCHEDULE | |
| 4 PERFORMING ORGANIZATION REPORT NUMBER(S) | 5 MONITORING ORGANIZATION REPORT NUMBER(S) |
| 6a NAME OF PERFORMING ORGANIZATION: Carnegie-Mellon University; 6b OFFICE SYMBOL (If applicable) | 7a NAME OF MONITORING ORGANIZATION: Cognitive Science Program, Office of Naval Research (Code 1142CS), 800 North Quincy Street |
| 6c ADDRESS (City, State, and ZIP Code): Schenley Park, Pittsburgh, PA 15213-3890 | 7b ADDRESS (City, State, and ZIP Code): Arlington, VA 22217-5000 |
| 8a NAME OF FUNDING / SPONSORING ORGANIZATION; 8b OFFICE SYMBOL (If applicable): 227 | 9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: N00014-89-J-1964 |
| 8c. ADDRESS (City, State, and ZIP Code) | 10 SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO: 0602233N; PROJECT NO: RM33M20; TASK NO; WORK UNIT ACCESSION NO |

DTIC ELECTE JAN 28 1991 S E D

11 TITLE (Include Security Classification)
T[illegible]ing Artificial Intelligence to Aid in the Development of Causal Models
Unclassified

12 PERSONAL AUTHOR(S)
Peter Spirtes, Clark Glymour, Richard Scheines, Steve Sorensen

| 13a TYPE OF REPORT | 13b TIME COVERED | 14 DATE OF REPORT (Year, Month, Day) | 15 PAGE COUNT |
|---|---|---|---|
| Final | FROM 4/1/89 TO 3/30/90 | Dec. 6, 1990 | 42 |

16 SUPPLEMENTARY NOTATION
Supported by the Office of the Chief of Naval Research, Manpower, Personnell and Training R&D Program

| 17 COSATI CODES | | | 18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number) |
|---|---|---|---|
| FIELD | GROUP | SUB-GROUP | Linear modelling, causal modelling, TETRAD II, automated inference |
| 05 | 02 | | |

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

Data analysis that merely fits an empirical covariance matrix or that finds the best least squares linear estimator of a variable is not a reliable guide to judgements about policy, which inevitably involve causal conclusions. We have developed and tested a computer program, TETRAD II, that accepts as input background knowledge about a causal structure, a covariance matrix, and a sample size, and outputs a set of suggested models compatible with the background knowledge and that explain the data. In tests on simulated data, TETRAD II was able to suggest a set of models that included the correct one 94% of the time. We have also applied TETRAD II to several data sets supplied by the Naval Personnel Research and Development Center.

| 20 DISTRIBUTION / AVAILABILITY OF ABSTRACT | 21 ABSTRACT SECURITY CLASSIFICATION |
|---|---|
| ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS | Unclassified |
| 22a NAME OF RESPONSIBLE INDIVIDUAL: Dr. Susan Chipman | 22b TELEPHONE (Include Area Code): 703-696-4318; 22c OFFICE SYMBOL: ONR 1142CS |

DD Form 1473, JUN 86 — Previous editions are obsolete — SECURITY CLASSIFICATION OF THIS PAGE

S/N 0102-LF-014-6603

# Using Artificial Intelligence to Aid in the Development of Causal Models

Peter Spirtes, Clark Glymour, Richard Scheines, and Steve Sorensen

## Final Report



This research was supported by the Manpower, Personnel, and Training R&D Program, Office of the Chief of Naval Research, and the Naval Personnel Research and Development Center, under contract No. N00014-89-J-1964.

# Manpower, Personnel, and Training R&D Program, Office of the Chief of Naval Research

# and

# Naval Personnel Research and Development Center

Contract # N00014-89-J-1964

## Using Artificial Intelligence to Aid in the Development of Causal Models

Peter Spirtes, Clark Glymour, Richard Scheines, and Steve Sorensen

## Final Report

June 27, 1990

# 1. Introduction

TETRAD II is a computer program designed to aid researchers reliably infer causal relations from statistical data. In our contract with the Manpower, Personnel, and Training R&D Program, and the Naval Personnel Research and Development Center, we proposed to perform the following tasks:

1. Extend the domain of application and increase the reliability of TETRAD II.

2. Test the reliability of TETRAD II

3. Apply TETRAD II to data sets supplied by the Naval Personnel Research and Development Center.

This report describes our work on the first two tasks; our application of TETRAD II to analyzing the causes of satisfaction and success among Naval Recruiters is described in the paper **Causes of Success and Satisfaction Among Naval Recruiters**; our analysis of the causes of success among naval air traffic controller trainees is described in the paper **TETRAD Studies of Data for Naval Air Traffic Controller Trainees**.

## 1.1 The Problem

Data analysis that merely fits an empirical covariance matrix or that finds the best least squares linear estimator of a variable is not of itself a reliable guide to judgements about policy, which inevitably involve causal conclusions. The policy implications of empirical data can be completely reversed by alternative hypotheses about the causal relations of variables, and the estimates of a particular causal influence can be radically altered by changes in the assumptions made about other dependencies. For these reasons, one of the common aims of empirical research in the social sciences is to determine the causal relations among a set of variables, and to estimate the relative importance of various causal factors. Even where that aim is not acknowledged it its often tacit. A question of first importance about empirical social science is therefore: how are causal relations among variables to be discovered?

The difficulty of this question is apparent when one considers the number of possible causal models for a given set of variables. If the causal dependence of one variable on another is represented by a directed edge from a vertex representing the causal variable to a vertex representing the effect variable, then the number of possible causal structures on n variables is the number of directed graphs with n vertices, or $4^{\binom{n}{2}}$. If causal cycles are forbidden, then the number of possible causal structures on n variables is the number of acyclic directed graphs on n

variables. For 12 variables the number of directed graphs is approximately 5.4 X $10^{39}$ and the number of acyclic graphs is 521,939,651,343,829,405,020,504,063 (Harary 1973). Even when the time order of the variables is known, so that causal hypotheses in which later variables cause earlier variables can be eliminated, the number of alterative remaining is generally very large: for 12 variables it is 7.4 X $10^{19}$.

The social scientist who addresses a problem area where causal questions are of concern is therefore faced with an extremely difficult discovery problem, for which there are only three avenues of solution: (i) use experimental controls to eliminate most of the alternative causal structures; (ii) introduce prior knowledge to restrict the space of alternatives; and (iii) use features of the sample data to restrict the space of alternatives.

Experimental procedures for addressing social questions are much to be desired, but they are very expensive and often infeasible. Where quasi-experiments are used that control some variables but not others, the number of alternative causal structures possible *a priori* may remain very large. Generating the set of admissible causal structures from "substantive theory" is recommended routinely in methodology texts. In practice publications in the social science literature usually restrict the number of alternatives considered to a very few, and the restrictions are often justified by citing prior literature or by appealing to very broad theoretical frameworks. It is anybody's guess, however, whether such appeals constitute a reliable discovery procedure. It seems at least as likely that appeals to theory introduce bias and often exclude the true causal relations among the variables of interest. TETRAD II uses the third avenue.

### 1.2 TETRAD II

Our work for the Manpower, Personnel, and Training R&D Program, and the Naval Personnel Research and Development Center represents a significant step toward the goal of reliably inferring causal relations form statistical data. TETRAD II, the computer program improved under this contract, allows researchers who have already measured and screened data to conduct a systematic search for alternative causal models. The class of models searched is not yet exhaustive, but it represents an enormous increase over the space of models humans are capable of searching unaided. We have tested TETRAD II's reliability on 720 data sets produced by Monte Carlo techniques from known models. For data sets of sample size 2000, given a model for which elaborations were to be searched, TETRAD II was able to output a small set of models (usually between two and four) which included the correct model in more than 94% of the samples. (See Spirtes forthcoming)

To illustrate the size of the problem, consider researchers who wished to consider literally all the alternative causal arrangements among those 25 variables, they would have to look at $4^{300}$ different models. If they knew the time order of the variables, so that for each pair they could rule out the causal arrangement in which the later one caused the earlier one, the number would still be astronomical: $2^{300}$. Even if they were highly confident of the causal relations between most of the pairs among these 25 variables, the number of alternative models that would satisfy their constraints is still likely to be orders of magnitude greater than the number they could feasibly consider. The natural solution to problems that strain the combinatoric capacity of human researchers is make high speed computers artificially intelligent .

### 1.2 TETRAD II

Our work for the Office of Naval Research and the Naval Personnel Research and Development Center represents a significant step toward that goal. TETRAD II, the computer program improved under this contract, allows researchers who have already measured and screened data to conduct a systematic search for alternative causal models. The class of models searched is not yet exhaustive, but it represents an enormous increase over the class searched by our original program, TETRAD, which itself represents an enormous increase over the space of models humans are capable of searching unaided. We have tested TETRAD II's reliability on 720 data sets produced by Monte Carlo techniques from known models. For data sets of sample size 2000, given a model for which elaborations were to be searched, TETRAD II was able to output a small set of models (usually between two and four) which included the correct model in more that 94% of the samples.

In our previous contract with ONR, we built an automated causal inference system named TETRAD II. Our goal was to build a program in which a user need only input the covariance data, and whatever prior knowledge there is about the domain into TETRAD II. The program then automatically searches for the elaborations of this knowledge that best explain the data, and estimate and tests the models that result. The user receives a description of the best explanations of the data that are consistent with the knowledge provided to the program and automatically produced input files for statistical analysis programs.

To illustrate the size of the problem, consider researchers who wished to consider literally all the alternative causal arrangements among those 25 variables, they would have to look at $4^{300}$ different models. If they knew the time order of the variables, so that for each pair they could rule out the causal arrangement in which the later one caused the earlier one, the number would still be astronomical: $2^{300}$. Even if they were highly confident of the causal relations between most of the pairs among these 25 variables, the number of alternative models that would satisfy their constraints is still likely to be orders of magnitude greater than the number they could feasibly consider. The natural solution to problems that strain the combinatoric capacity of human researchers is make high speed computers artificially intelligent .

### 1.2 TETRAD II

Our work for the Office of Naval Research and the Naval Personnel Research and Development Center represents a significant step toward that goal. TETRAD II, the computer program improved under this contract, allows researchers who have already measured and screened data to conduct a systematic search for alternative causal models. The class of models searched is not yet exhaustive, but it represents an enormous increase over the class searched by our original program, TETRAD, which itself represents an enormous increase over the space of models humans are capable of searching unaided. We have tested TETRAD II's reliability on 720 data sets produced by Monte Carlo techniques from known models. For data sets of sample size 2000, given a model for which elaborations were to be searched, TETRAD II was able to output a small set of models (usually between two and four) which included the correct model in more that 94% of the samples.

In our previous contract with ONR, we built an automated causal inference system named TETRAD II. Our goal was to build a program in which a user need only input the covariance data, and whatever prior knowledge there is about the domain into TETRAD II. The program then automatically searches for the elaborations of this knowledge that best explain the data, and estimate and tests the models that result. The user receives a description of the best explanations of the data that are consistent with the knowledge provided to the program and automatically produced input files for statistical analysis programs.

The general approach that TETRAD II uses is to search for those causal structures that as closely as possible imply all and only those probabilistic constraints that are judged to hold in the population. (See the Technical Appendix for details.)

In more detail the original version of TETRAD II had the following elements:

**KNOWLEDGE FILE:** A file into which the user enters the data and whatever initial constraints there are on the causal relations among the variables.

**GENERATE PROGRAM:** The GENERATE program uses a body of heuristics to produce a collection of simple initial models. The initial models may include latent variables.

**EDIT INITIAL MODELS:** The user can stop the program after the initial models have been created and edit the initial models, either by eliminating initial models created by the program or by adding initial models other than those created by the program.

**AUTOMATIC TETRAD:** The initial models are elaborated using an algorithm that fully automates the procedures and heuristics now used by the TETRAD program.

**COMPARE MODELS:** The elaborated models are compared by TETRAD's heuristic fit criteria, which combine considerations of simplicity and ability to explain patterns in the data. The best of the elaborated models are retained.

**STATISTICAL ASSESSMENT:** The program automatically prepares input files for an existing commercial statistics package, the EQS program, and submits each of the best elaborated models to statistical estimation and testing. At the user's preference, the data used for estimation and testing may be either the data used in the search for models, or may be a new and distinct sample for the same variables.

**OUTPUT:** The output of the program includes a list of the best causal structures, in both human-readable form, and in the form of input files for statistical packages. The statistical packages can then be used to evaluate the models suggested by TETRAD II.

## 2. Results

We proposed to improve the existing TETRAD II by answering the following three fundamental questions:

1. Are there rigorous, fast algorithms for searching the enormous space of possible models compatible with given background knowledge, even when the number of observable variables is large?

2. Can we introduce new classes of constraints and more flexible representations of background knowledge that will allow the use to reduce the number of suggested models to a more manageable size?

3. Under what circumstances is TETRAD II reliable?

The following is a brief summary of the progress that we have made towards answering these questions. A more detailed description is given in the following sections, and theorems concerning the reliability and scope of the algorithms that we devised are stated in the Technical Appendix.

1. We have designed, implemented, tested, and employed an algorithm for building path models from covariance matrices.

2. We have designed, implemented, and employed an algorithm for selecting subsets of variables that form measurement models for a given latent variable.

3. We have speeded up the search TETRAD II conducts to elaborate initial causal models.

4. We have designed, implemented, and employed several tests for determining when latent variables should be introduced into a model.

5. We have designed, implemented, and employed two new algorithms for constructing multiple indicator models.

6. We have designed, implemented, and tested algorithms for reducing the number of models suggested TETRAD II without reducing the reliability of the program (by introducing new classes of constraints on the covariance matrix.)

7. We have designed, implemented, and employed an algorithm that turns the simple model specifications employed by TETRAD II into the more complex input files required by the EQS and LISREL statistical packages.

8. We have simplified and improved the user interface to TETRAD II.

9. We have analyzed statistical data and suggested causal models concerning the causes of recruiter success and satisfaction. We report our findings in an attached paper (**Causes of Success and Satisfaction Among Naval Recruiters**).

In the following section we will repeat the details of the modifications we proposed to make to TETRAD II (in italics), followed by a description of what modifications we actually made, and what evidence we have about the reliability of the algorithms.

**2.1 Input to LISREL:** *Specification of a model in TETRAD II is much simpler than specification of a model in LISREL. We will implement a module that translates the simple TETRAD II model specification into the more complex LISREL model specification. We will also allow the user have TETRAD II output its suggested models in files that can be used as input to LISREL.*

We have completely implemented this feature of TETRAD II in the "Lisrelinput" command. In addition, if a user creates a number of LISREL input files, he has the option of automatically creating a batch file (with the "Lisrelbatch" command) that will run all of the LISREL input files created with one command. The user has the option of specifying starting values for the free parameters of each model, or allowing TETRAD II to choose starting values for the free parameters. The user can choose how many of the models suggested by the "Suggest" command (which elaborates an initial model input by the user) will be automatically translated into LISREL input files; also the "Suggest" command will automatically create a batch file so that all of the LISREL input files can be run with one command. In addition, TETRAD II can automatically create input files for the EQS program in an analogous way (using "Eqsinput" and "Eqsbatch".)

**2.2 Addition of SELECT module:** *We believe that a judicious selection of the observable variables to include in a model can greatly simplify it construction. In many studies, such as in psychometric testing, hundreds of variables are measured. The vast majority of these variables are intended to be indicators of some latent variable, and the real interest is in the relations among the latent variables. By carefully selecting indicator variables, the task of creating plausible submodels can be greatly simplified. Many of the techniques we currently use to generate clusters of variables from covariance data can be modified to select only those observable variables that will make the construction of simple submodels easier and faster.*

We have completely implemented this feature of TETRAD II in the "Scales" command. The "Scales" command helps users construct measurement models of latent variables. We use the term "measurement model" of a latent variable L to mean a model with the following properties:

1. each indicator is directly caused by and is a linear function of L;

2. no pair of indicators has a common cause other than L;

3. no indicator causes any other indicator;

4. the value of each indicator is primarily determined by L (rather than by its error variable).

The first three conditions are restrictions on the causal structure; the fourth condition is a restriction on the free parameters of the model. A model satisfying the restrictions on the causal structure is depicted in the graph in Fig. 1.

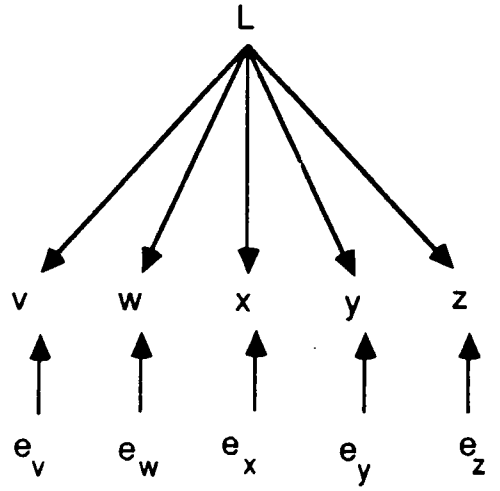



**Fig. 1: Measurement Model of L**

Models that violate condition (1) are depicted in Fig. 2(i) and 2(ii); a model that violates condition (2) is depicted in Fig. 2(iii); and models that violate condition (3) are depicted in Fig. 2(iv) and 2(v).

The input to the "Scales" command is a set of random variables which on substantive grounds are thought to form a measurement model of some latent variable L. For example, a user might guess on substantive grounds that the set {v, w, x, y, z} form a measurement model of L. If the actual causal relations among these variables is that depicted in Fig. 2(iv) however, these variables do not form a measurement model of L. But there is a subset of the variables {w, x, y, z}. that does form a measurement model of L. Hence, the "Scales" command would output the set {w, x, y, z}. Obviously, with just a few variables, the task is trivial; but the problem rapidly exceeds the ability of unaided humans at even quite small numbers of variables.

We employ the following steps to eliminate indicators that are not part of measurement models.

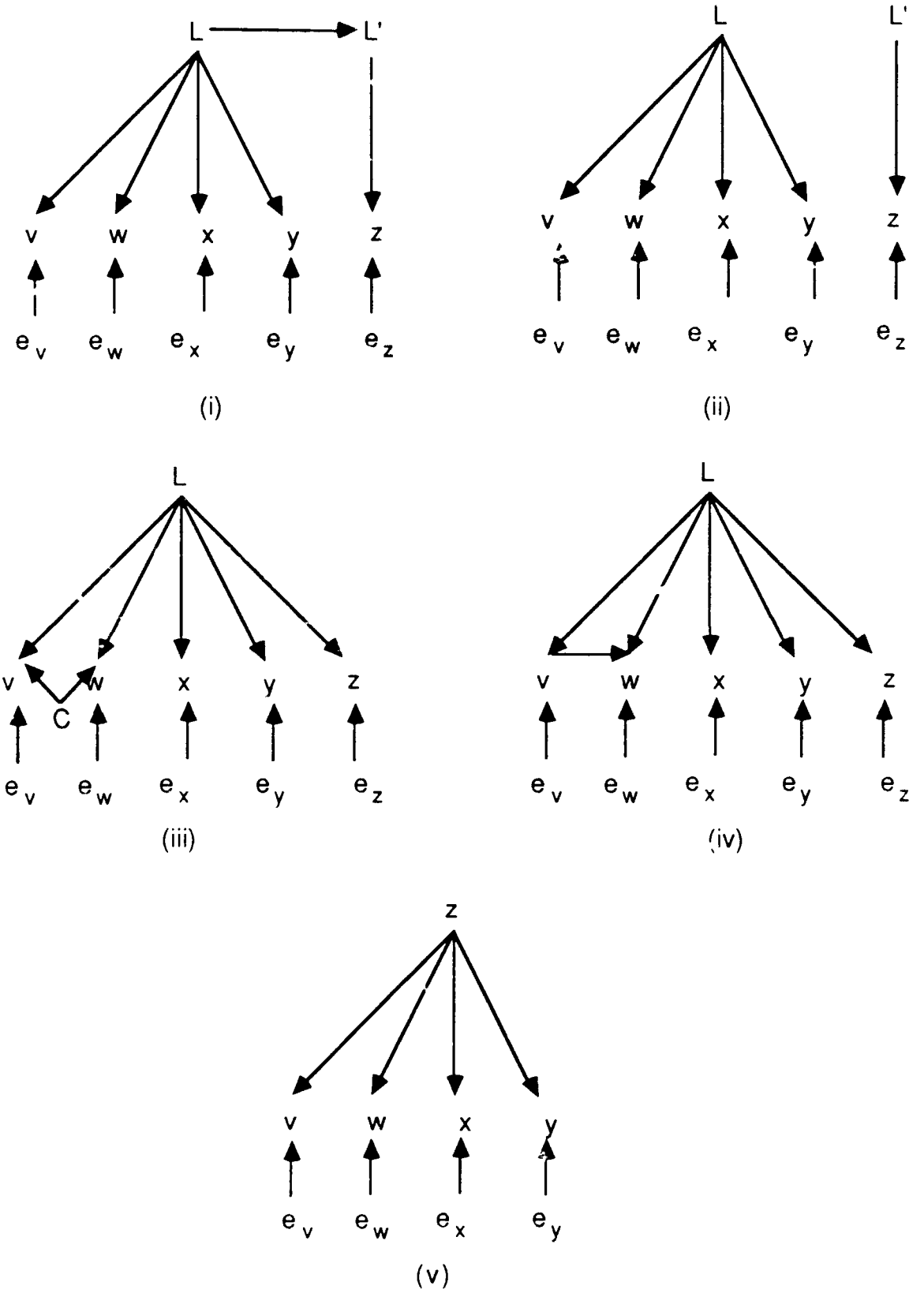



**Fig. 2: Non Measurement Models**

1. First we eliminate indicators that have a zero correlation with some other indicator in the original set. This eliminates models that have the type of causal structure that is depicted in Fig. 2(ii).

2. A tetrad difference among among variables x, y, z, w, is $\rho_{xy}\rho_{zw} - \rho_{xz}\rho_{yw}$. We eliminate those sets of five indicators which have a high proportion of tetrad differences that we judge not to vanish in the population.[1] This eliminates models with the type of causal structures depicted in Fig. 2(iii) and 2(iv). (If we cannot find any sets of five indicators having the properties we seek, then we search for sets of four indicators with these properties.)

3. We then eliminate models for which there is a high proportion of foursomes of indicators and subsets S of the indicators such that $\rho_{xy.S} = \rho_{zw.S} = \rho_{xz.S} = \rho_{yw.S} = 0$. This eliminates models with the type of causal structure depicted in Fig. 2(v).

4. TETRAD II lists all of its suggestions for good measurement models, and allows the user to select from among these suggestions a group of models for which TETRAD II automatically writes EQS input files. The user can then submit these input files to EQS, and eliminate models that fail to satisfy condition (4), or are non-linear.

Note that this procedure eliminates *all* alternatives to good measurement models *except* for models with the causal structure depicted in Fig. 2(i). This type of causal structure cannot be distinguished from the causal structure of Fig. 1 using only covariances among indicators for one latent variable. However, it can be detected at a later stage, when the various measurement models are assembled together.

We have applied this procedure to the latent variables in the Navy's Recruiter data. It has proved very successful in generating measurement models that perform well on statistical tests, as the following table shows. (A $P(\chi^2)$ above 0.05 is generally considered a good score for a model, especially at large sample sizes.)

---

[1] We chose to look for sets of five indicators because they are sets that are large enough to make it improbable that the statistical features that we are searching for occur by coincidence, but small enough to make the search feasible.

| Scale | $X^2$ | $P(X^2)$ |
|---|---|---|
| Adv | 1.342 | 0.5112 |
| Eval | 0.400 | 0.8188 |
| Fam | 0.296 | 0.8622 |
| Goals | 2.497 | 0.7769 |
| Mat | 0.220 | 0.8950 |
| Nimag | 0.110 | 0.9466 |
| Ojt | 3.300 | 0.6539 |
| Pjt | 0.832 | 0.6595 |
| Sat | 3.470 | 0.6280 |
| Sel | 4.898 | 0.0864 |
| Stress | 8.747 | 0.1196 |
| Super | 0.593 | 0.7434 |
| Support | 1.864 | 0.8677 |

**Table 1: Construction of Measurement Models For Naval Recruiter Data**

See our attached report **Causes of Success and Satisfaction Among Naval Recruiters** for more details.

**2.3 Addition of other types of models:** *Currently, all of the models that the GENERATE module of TETRAD II constructs automatically are latent variable models. It does not attempt to determine whether introduction of latent variables is appropriate. We will develop an algorithm for determining whether or not latent variable should be introduced into a model. We will also develop an algorithm for constructing path models when the introduction of latent variables is not appropriate. We have already found a class of constraints (vanishing and positive partial correlation constraints) that are implied by path models, and we have developed an algorithm for determining when these constraints are implied by a model. We will integrate this new class of constraints into our AUTOMATIC TETRAD and GENERATE modules.*

### 2.3.1 Construction of Path Models

We will restrict our discussion to those causal models among a given set of variables S that can be represented by an acyclic directed graph. We will say that a directed graph G is a causal graph of S

iff there is a directed arrow from A to B iff A causes B directly (i.e. without the mediation of any other variable in S.)

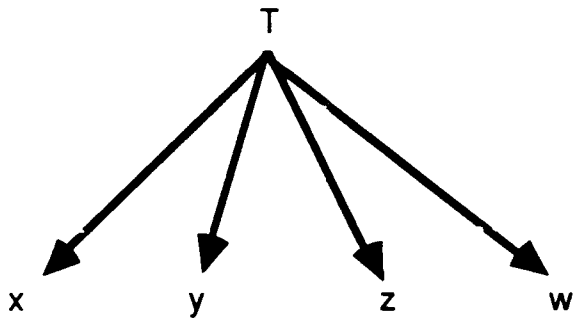



**Fig. 1: Graph G**

If every common cause of a pair of variables in S is also in S, then we say that S is **causally sufficient.** In Fig. 1, the set of variables S = {T,x,y,z,w} is causally sufficient; however the set of variables S' = {x,y,z,w} is not causally sufficient because the variable T, which is a common cause of x, y, z, and w, is not in S'. When we use a graph to represent the causal structure among a set of variables S, we assume that the set of variables is causally sufficient.

We have developed an algorithm for constructing a complete set of path models (models over causally sufficient sets of variables) compatible with a given set of independence constraints. The correctness of this algorithm follows from a theorem proved in Pearl[1990]. While the algorithm is exponential in the number of random variables, it is feasible for present-day computers for up to around 20 variables, which covers a large proportion of the variable sets studied in the social science literature (We have actually implemented the algorithm for 17 normally distributed variables on a DecStation 3100, and it runs in about 20 minutes. The problem with expanding the algorithm to larger sets of variables is largely one of space, not of time.) Use of certain kinds of background knowledge would allow the algorithm to be used on larger sets of variables.

The following **Path Model Construction Algorithm** that we devised constructs the set of all causal models that imply all and only a given set of independence relations over the set S of random variables (if such a causal structure exists):

1.) Add an undirected edge from A to B iff A and B are dependent given every subset of S not containing A and B.

2.) If there are undirected edges between A and B, and B and C, but not between A and C, then there is an edge from A to B and from C to B iff A and C are dependent given every subset of S containing B, but not A or C.

3.) The set of edges whose orientations are not fixed by 2) are given every possible orientation that does not create a collision (A -> B <- C).

The algorithm is applicable to a wide variety of probability distributions. (See the Appendix for more details.) However, we currently judge whether or not A and B are independent conditional on S by performing a statistical test to determine whether $\rho_{AB.S}$ vanishes; this test for independence is correct only for normal models. In order to make the algorithm reliable for other distributions, tests of conditional independence for these distributions need to be developed.

On samples of medium size the procedure that applies the Path Model Construction Algorithm has a tendency to underfit, that is it tends to omit undirected edges corresponding to causal dependencies in the structure from which the data were obtained.

Data for a sample of 2000 population units were generated by Monte Carlo methods from a linear model with the following causal structure (using normally distributed variates):

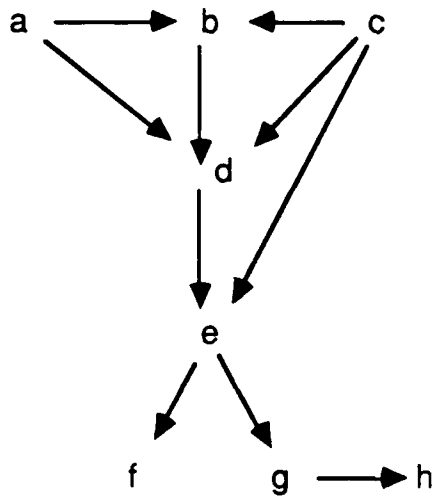



**Fig. 8**

The data were then given to the TETRAD II program, which produced the following undirected graph:

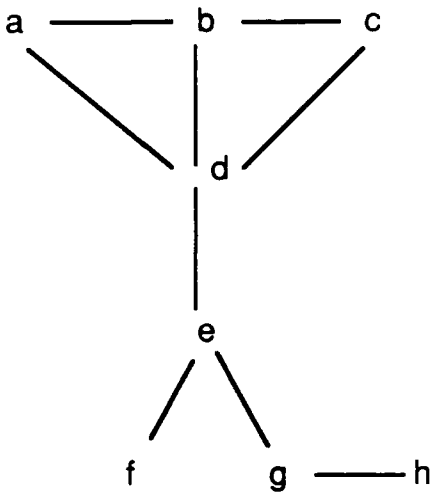



**Fig. 9**

Under the restriction that no unmeasured variables are to be introduced, Theorem 2 permits only two orientations of this graph from the sample data--exactly the orientation of the initial model used to generate the data, and the orientation that is otherwise the same but reverses the direction of the b - d connection.

An empirical example is provided by recent work of Rogers and Maranto (Rogers 1989). They studied a number of theoretical accounts of the determinants of publishing productivity in psychology, and compared these accounts with original survey data they obtained. After path analyses of six alternative models taken from the social science literature, they formed a combined model that includes all causal dependencies occurring in any of these models as well as two further dependencies. After estimating and testing the combined model, they eliminated the dependencies found not to be statistically significant. Their result is the following causal model (with coefficients estimated assuming linear dependencies)[2]:

---

[2]GPQ is a scale formed from indicators of the quality of graduate programs; QFJ is a measure of the quality of the subject's first job; PREPRO is a measure of publications while in graduate study; PUBS a measure of publications since leaving graduate school, and CITES indicates frequency of citation of the subject's scholarly works.

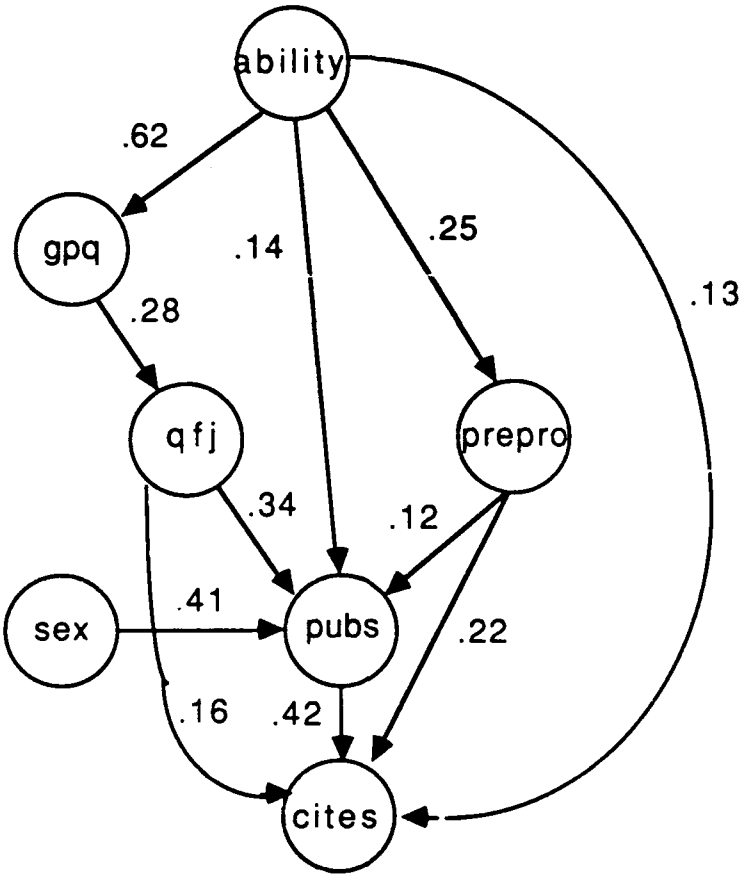



**Fig. 10**

When the correlations from Rogers and Maranto's survey data are given to the TETRAD II procedures (whose present implementation assumes normal variates), we automatically obtain the following undirected graph without imposing any prior substantive restrictions:

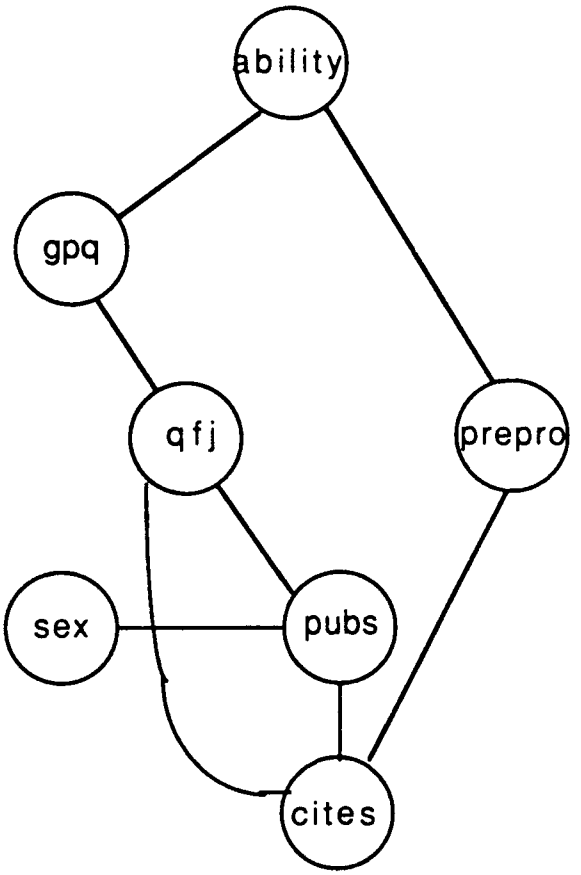



**Fig. 11**

The application of step 1 of the Path Model Construction Algorithm to the sample data yields eight of the eleven connections Rogers and Maranto postulate, and the three omitted edges are those which, in the linear model, have the smallest coefficients.

The time order of ability, graduate school quality, quality of first job, and publication in graduate school and other variables uniquely determine the directions that can be given to the edges in this undirected graph. Step 1 of the Path Model Construction Algorithm, plus partial time order information, yields a unique directed graph that contains all but the three smallest of the dependencies Roger's and Maranto postulate.

In the Rogers and Maranto data the application of step 2 of the Path Model Construction Algorithm to the undirected graph in figure 6 fails to determine the correct order of the directed edges. It is interesting to see why step 1 of the Path Model Construction Algorithm seems to succeed so well in this case and step 2 of the Path Model Construction Algorithm fails to provide the goods. The essential reason is that any errors in the undirected graph obtained with Path Model Construction

Algorithm that arise from sampling errors are localized. With step 2 of the Path Model Construction Algorithm the consequences of a sampling error may not be localized at all; the same sorts of dependencies that enable us to orient the graph from very little information, also enable us to misorient it badly from a very few errors.

The Rogers and Maranto sample size is small, and some of the variables are certainly not even approximately normally distributed. The result is that tests performed on the data result in more vanishing partial correlations than we would expect to hold in the population were their model correct and the variables normally distributed. So the Path Model Construction Algorithm work with incorrect conclusions about conditional independence. In applying step 1 of the Path Model Construction Algorithm, we infer A - B if the partial correlation of A,B on X is non-zero for every set X not contain A or B. In applying step 2 of the Path Model Construction Algorithm, we infer A -> B <- C if the undirected graph is A - B - C and the partial correlation of A, B on X is non-zero for every set X containing C but not A or B. Thus a sampling error which results in the conclusion that A and B are conditionally independent on some set X containing C may cause the procedure to omit an A - C connection that obtains in the true causal structure. This error will have no effect on the reliability of inferences about other edges in the *undirected* graph. If the result of this mistaken inference is the graph A - B - C, the same sampling error will lead by Theorem 2 to the erroneous conclusion that the correct structure is *not* A -> B <- C. The information that the edges from A and C are into B may, however, be essential for determining the direction of other edges in the graph. So an error in estimation of conditional independence that results in localized errors in the undirected graph may have global effects on the directed graph.

This appears to be what happens in the Rogers and Maranto data where B = cities, A = pubs and C = preprod. Without either the Ability -> cities connection (not found using Theorem I) or the preprod -> cities orientation, the QFJ - cities direction is indeterminate according to step 2 of the Path Model Construction Algorithm . The result is that rather than giving the set of two orientations that includes the correct one, the TETRAD II program outputs a large number of orientations that do not include the correct one. step 2 of the Path Model Construction Algorithm , while sound with data that nearly perfectly represent the population, nonetheless suffers from a fundamental instability in real inference problems. We are currently working on an alternative algorithm to step 2 which will not be unstable.

### 2.3.2 Introduction of Latent Variables

We have taken two approaches to the problem of when and how to introduce latent variables. In the first approach we have devised tests for determining when latent variables are needed to explain the statistical features of a population; in the second approach we have determined what causal conclusions can be reliably drawn from the statistical features of a population whether the set of measured variables is causally sufficient or not.

We have developed two tests for determining when latent variables are required by a causal model. The first test can be used in the situations described below:

> **Latent Variable Test 1:** Whenever an unmeasured cause affects two variables, X and Y, such that X is the direct effect of some unmeasured variable and Y is the direct effect of the same unmeasured variable, the presence of the unmeasured variable can be identified from statistical dependencies. If the Graph Construction Algorithm orients a given edge in *both* directions, then a latent variables must be present.

To illustrate the point, the statistical dependencies among the measured variables (where T is unmeasured) enable us to distinguish each of the following causal structures from the others.

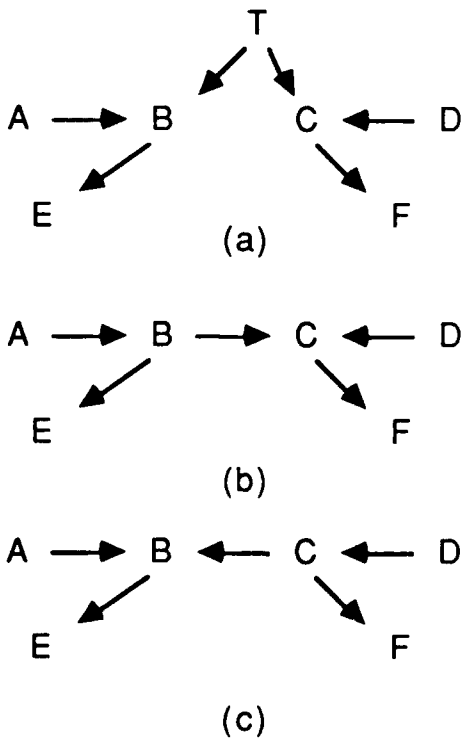



**Fig. 7**

We are currently implementing this test into TETRAD II.

The second test for latent variables is applicable to linear normal models:

> **Latent Variable Test 2:** Suppose a multinormal probability distribution P implies a vanishing tetrad difference $\rho_{ij}\rho_{kl} - \rho_{il}\rho_{jk} = 0$. Then introduce a latent variable if either $\rho_{ij}$ and $\rho_{kl}$, or $\rho_{ik}$ or $\rho_{jl} \neq 0$, or there does not exist a set S of vertices of G such that the variables i and j are independent conditional on S, and so are the pairs of variables k, l and i, k and j, l.

This test has been incorporated into TETRAD II, and warns users when latent variables should be introduced. It is also used as one step in the "Scales" command where it has been quite useful in constructing measurement models. See Table 1.

The second approach to the problem of latent variables that we have adopted is to find statistical features of populations that imply the presence or absence of some kinds of causal relations whether the measured variables form a causally sufficient set or not. If S' is *not* causally sufficient, the distribution P' over S' is the marginal of some distribution P over a causally sufficient set S, where S' is properly included in S. If we could determine what features of the path models constructed by the Path Model Construction Algorithm from P' are necessarily also features of the causal structure that generated P, it wouldn't matter whether S' was causally sufficient or not. Using a recent theorem of Verma and Pearl about the equivalence of marginal distributions, we have proved the following results:

> **Theorem I:** If there is no edge between A and B in any causal structure that implies just the conditional independence relations true of P', then there is no edge between A and B in any causal structure that implies just the conditional independence relations true of P.

> **Theorem II:** If there is no directed path from A to B in any causal structure that implies just the conditional independence relations true of P', then there is no directed path from A to B in any causal structure that implies just the conditional independence relations true of P.

Theorem II is crucial because it allows one to reliably infer that A is not a cause (even indirectly) of B. We also have a sufficient condition for the conditions under which one can reliably conclude that A is a cause of B, but it cannot be stated in this limited space.

**2.4 Improvement of GENERATE module:** *Once subsets of variables have been selected, they will be passed to an improved GENERATE module that will generate submodels out of the various subsets of variables. Currently, the output of GENERATE is a number of different clusterings of the observable variables into groups. Each variable in a group of observables is an indicator of the same latent variable. It does not suggest how the latent variables are connected together; this is done by the AUTOMATIC TETRAD module. We believe that we can revise the GENERATE module to do more of the work of constructing initial models without substantially increasing the amount of time that it takes to operate. This additional information will greatly speed up the operation of the AUTOMATIC TETRAD module in generating initial models. We have already succeeded to some extent in getting the GENERATE module to generate more information than the simple clustering information it currently outputs. Some of this additional information concerns direction connections between latent variables, and others are simply constraints on how the latent variables may be connected.*

We have abandoned our original algorithm in the GENERATE module (at least temporarily). It suggested too many alternatives, and was much too sensitive to sampling error to be reliable. For path models we have replaced it with the Graph Construction Algorithm described in the previous section; for latent variables models we have replaced it with the two algorithms described in the following section.

**2.5 Addition of COMBINE module:** *Once the GENERATE module has generated a set of submodels, the submodels will be passed to the AUTOMATIC TETRAD, COMPARE, and STATISTICAL ASSESSMENT modules to be elaborated and then assessed for fitness. Once the best elaborated submodels have been selected, they will be sent to the COMBINE module, which will have the task of combining the various submodels into a single model. A rudimentary algorithm for combining small submodels into larger ones has already been implemented in the GENERATE module, but it needs to be extended and speeded up in order to work on larger submodels.*

In a multiple indicator model, each measured variable has exactly one immediate causal ancestor, and that ancestor is a latent variable. Fig. 3 depicts an example of a multiple indicator model. Note

that it consists of a collection of causal structures of measurement models, with possible additional causal connections among the latent variables.

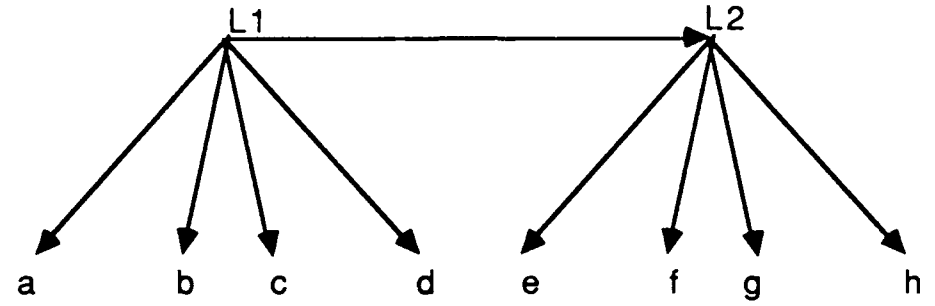



**Fig. 3: Multiple Indicator Model**

Fig. 4 depicts several different ways in which a collection of measurement models can fail to form a multiple indicator model. In Fig. 4(i), there is a variable that is directly caused by two latent variables and in Fig. 4(ii) there is a measured variable which is caused by another measured variable.

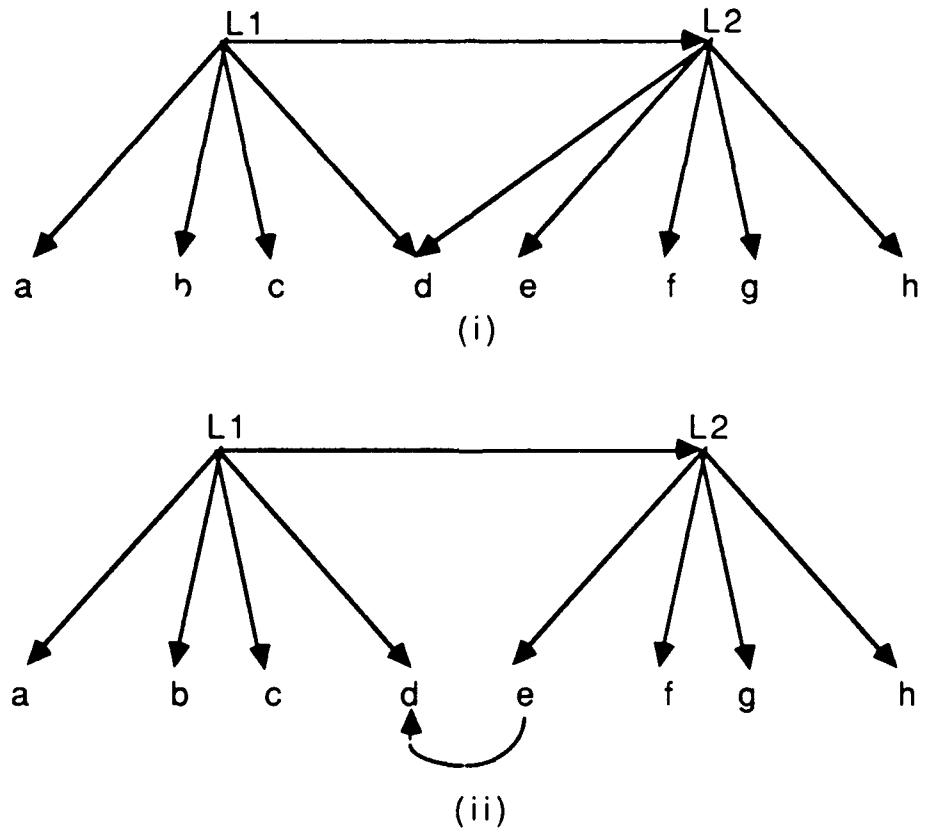



**Fig. 4: Non Multiple Indicator Models**

We have implemented two algorithms for constructing multiple indicator models; each has advantages and disadvantages. (A more complete description of the algorithms is presented in the attached report **Causes of Success and Satisfaction Among Naval Recruiters**.) The first step in each algorithm is to use the "Scales" command to select sets of indicators that form measurement models. Each of the algorithms then searches for additional causal connections between the latent variables. (Ideally, these algorithms should eliminate any indicators that have more than one parent, such as d in Fig. 4; however, the current versions fail to check for this possibility.)

In the first algorithm that we use to construct multiple indicator models, we pass the measurement models formed by the "Scales" command to the "Suggest" command, which performs a heuristic search for causal connections among the latent variables. (See the Technical Appendix for a detailed description of the search.) The advantage of this technique is that it is not restricted to latent variables that are normally distributed, or linear functions of each other. The disadvantage is that as it is currently constructed it can work on only relatively small models (5 or 6 latent variables) due to time limitations.

In the second algorithm that we use to construct multiple indicator models, we input the collection of measurement models suggested by the "Scales" command into EQS, and ask EQS to estimate the covariances between the latent variables. We then use the estimated covariance matrix among the latents as input to our Path Model Construction Algorithm, and form a path model among the latent variables. The advantage of this method is that it can work on models containing up to 17 latent variables in a relatively short period of time (approximately 20 minutes.) The disadvantages are that the estimation technique for the covariance matrix among the latents assumes that the variables are linearly related, and that the need to estimate the covariance matrix among the latent variables introduces an extra element of uncertainty.

This algorithm has succeeded on Monte Carlo simulation tests that we have performed. Fig. 5 depicts a model that we used to generate data.

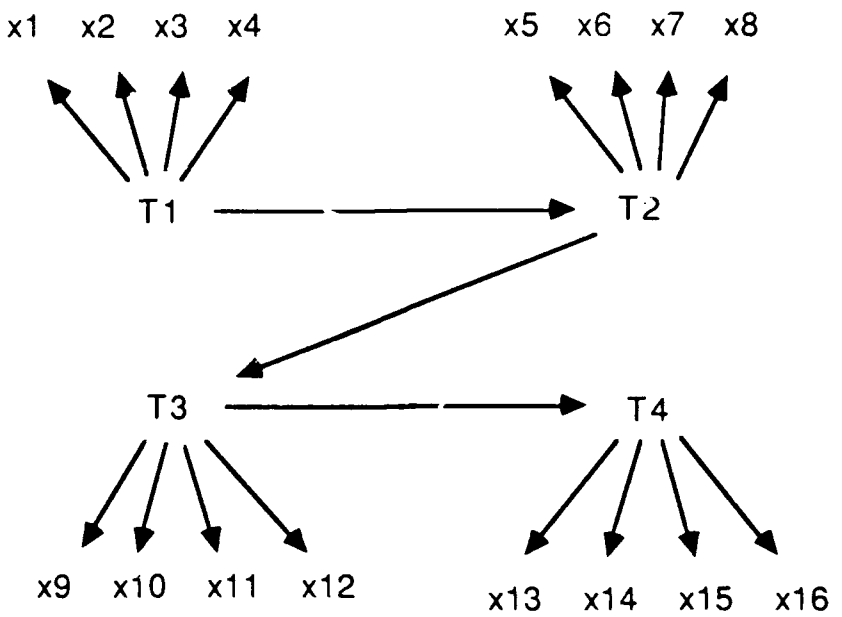



**Fig. 5: Model that Generated Data**

We assumed that the clusters of variables were given. When we estimated the covariances among T1, T2, T3, and T4 using EQS, and used this covariance matrix as input to the Path Model Generating Algorithm, TETRAD II successfully suggested a set of model that included the correct one.

Unfortunately, when we attempted to use this procedure on 15 latent variables in the Naval Recruiter data, the algorithm did not produce reasonable results; it suggested that there were no causal connections among the latent variables. On smaller subsets of the latent variables, the procedure produced more reasonable suggestions, but whether or not TETRAD II suggested a causal link between a given pair of variables varied to a certain extent depending upon which other variables were examined. We believe that the major problem in applying this technique to the Naval Recruiter data set was that the variables were highly non-normal, invalidating the statistical tests for independence that we performed.

**2.6 Extension and Integration of KNOWLEDGE FILE:** *The KNOWLEDGE FILE will be used by the GENERATE module as well as the AUTOMATIC TETRAD module, in order to make use of the users knowledge as early as possible in the generation process. The pilot version of the KNOWLEDGE FILE allows users to require or forbid specific edges, paths, or treks in the graphs that represent the models suggested by TETRAD. More work needs to be done in allowing users to enter other types of knowledge that they may have.*

We have not yet changed the format of the Knowledge File. We are in the process of incorporating the constraints in the Knowledge File into the Graph Construction Algorithm, and allowing the user to specify the time order of variables.

**2.7 Addition of new constraints:** *TETRAD II works by comparing various constraints on the covariance matrix that are implied by a model with the actual data. We have recently discovered a new class of constraints that is capable of distinguishing between models that have the same T-scores. Use of these constraints requires knowledge of some of the signs of the causal connections between variables. This knowledge is not always available, but it is available in many cases of psychometric testing, where observable variables are often deliberately chosen to be positive indicators of some latent variable. We still need to prove that our algorithm for calculating these constraints is correct, to incorporate it into our current modules, and to find out how much of a reduction in the number of suggested models including these constraints will make. Preliminary tests have shown promising results.*

We have incorporated these constraints into the "Suggest" command of TETRAD II, which elaborates models input by the user. In Monte Carlo simulations, the reduction of the number of models suggested by TETRAD II varied greatly according to the causal structure. We ran an extensive Monte Carlo simulation of the reliability of the "Suggested" command. (For details see "Simulation Studies of the Reliability of Computer Aided Model Specification Using the TETRAD II, EQS and LISREL VI Programs", **Sociological Methodology and Research**, forthcoming.) We generated data sets from 9 different causal models. In each case we gave as input to TETRAD II part of the causal structure that generated the data, and asked TETRAD II to recover the part of the causal structure that we had omitted. TETRAD II was able to suggest a set of models that included the correct model 95% of the time. Table 2 shows how many models were suggested by TETRAD II when the signs of the edges input to the "Suggested" command were known, and how many were suggested when the signs of the edges were not known.[3]

---

[3]For convenience, we have calculated the number of models suggested by TETRAD II only in those samples where TETRAD II was correct. Since TETRAD II was correct in 95% of the cases, this is a good approximation to the average number suggested in all the samples.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Signs Known | 4 | 2 | 3 | 1 | 1 | 3 | 19 | 16 | 3 |
| Signs Unknown | 4 | 2 | 3 | 1 | 3 | 4 | 29 | 20 | 3 |

**Table 2: Number of Models Suggested by TETRAD II**

## 3. Work In Progress

We are in the process of making the following modifications to TETRAD II, which will extend the domain of applicability of TETRAD II and make it easier to use:

1. We are implementing statistical tests for vanishing tetrad differences and vanishing partial correlation that are asymptotically distribution free.

2. We are making changes to the Path Construction Algorithm to make it less sensitive to sampling error, and clarifying the output of the Algorithm.

3. We are adding a feature that allows TETRAD II to output all of the models equivalent to a given model.

4. We are designing a menu-driven interface, similar to that used by programs on the Macintosh.

5. We will allow the user to specify what subset of variables in the covariance matrix should be employed in the elaboration and model building procedures.

6. We are improving the estimates that TETRAD II gives of how long the elaboration procedure is likely to take.

# TECHNICAL APPENDIX

## A.1 The "Suggested" Command

### A.1.1 Structural Equation Models and Graphs

The "Suggested" command of TETRAD II is intended to correct mis-specifications in the class of structural equation models. A *structural equation model*, or *linear causal model* consists of four parts:

A set of random variables with a joint distribution.

A set of linear equations among the random variables.

Distributional assumptions about the random variables.

A set of causal relations among the random variables.

The following is an example of a structural equation model.

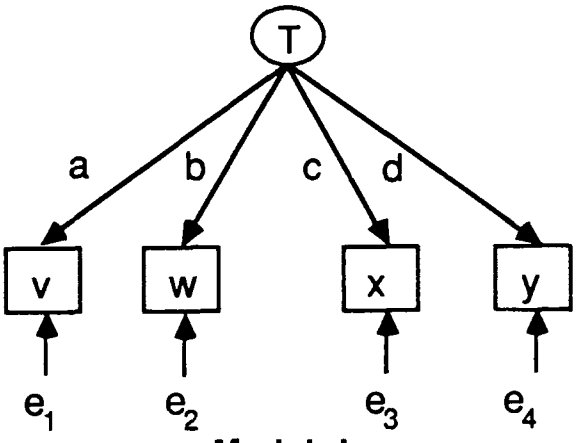
The set of random variables is $\{v,w,x,y,T,e_1,e_2,e_3,e_4\}$. In this case, the variables v,w, x, y, and z are measured variables, T is a latent variable, and $e_1,e_2,e_3,e_4$ are "error" or "disturbance" terms.

The set of linear equations is:

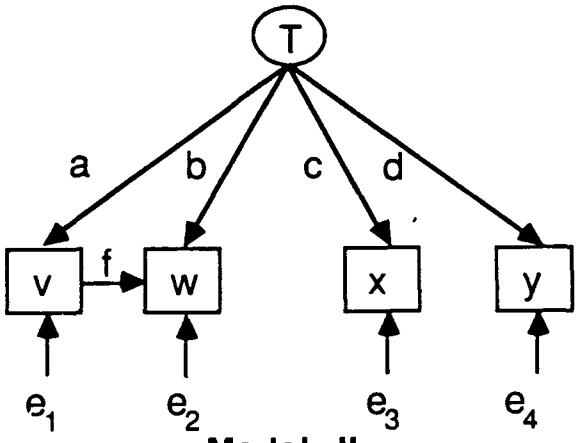
$$v = a\,T + e_1$$
$$w = b\,T + e_2$$
$$x = c\,T + e_3$$
$$y = d\,T + e_4$$

The collection of all variables is multi-normally distributed. T and the $e_i$ are uncorrelated and have unit variance and zero mean.

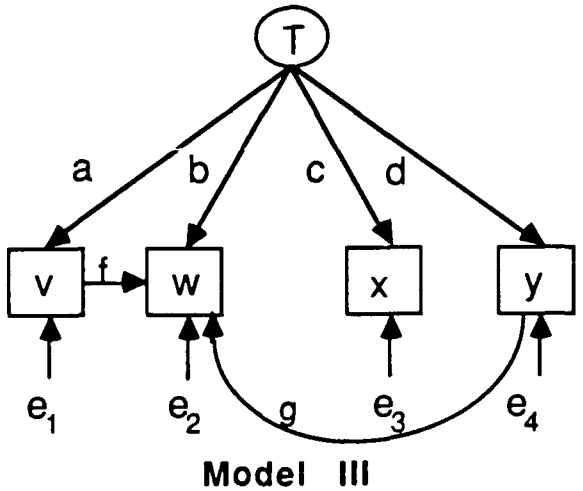
The set of causal relation is: $\{<T,v>,<T,w>,<T,x>,<T,y>,<e_1,v>,<e_2,w>,<e_3,x>,<e_4,y>\}$, where <r,s> is in the set of causal relations if and only if r is an immediate cause of s.




**Model I**




**Model II**




**Model III**

**Fig. A.1: 3 Models**

The set of causal relations can also be represented in a directed graph, in which there is an edge from r to s if and only if r is an immediate cause of s. The directed graph for this example is depicted in Model I of Fig. A.1.

By convention, the linear equations are expressed in a canonical form in which r appears in the equation for s iff s is a direct cause of r. Thus v is expressed as a function of $e_1$, while $e_1$ is not expressed as a function of v. This convention allows important parts of the statistical model to be recovered from the graph alone. The graph encodes the form of the linear equations, and it encodes 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., multinormal).

### A.1.2 Input and Output of Suggested

The "Suggested" command is designed to aid researchers in correcting misspecifications of causal models. The "Suggested" command accepts as input:

a sample size,

a correlation or covariance matrix, and

domain knowledge in the form of required edges, forbidden edges, whether the graph can be cyclic, and whether or not the graph can contain direct cycles (i.e. an edge from A to B and B to A).

The required or forbidden edges are given to the program simply by specifying a list of paired causes and effects.

"Suggested" assigns to each model a score that we will call its Tetrad-score. Its output is a list of models, ranked according to their Tetrad-scores. TETRAD II does not perform parameter estimation or compute a $\chi^2$ statistic. Typically, we take the models suggested by the "Suggested" command and submit them as input to parameter estimation programs such as LISREL VI or EQS. The algorithm that the "Suggested"

command uses can be divided into two parts, a scoring algorithm and a search algorithm.

### A.1.3 Overidentifying Constraints

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 *overidentifying constraints* that a structural equation model may imply. Several such classes of constraints concern *tetrad differences*. A tetrad difference is just the determinant of a 2 X 2 submatrix of the covariance matrix: $\gamma_{ij}\gamma_{kl} - \gamma_{ik}\gamma_{jl}$, where $\gamma_{ij}$ is the covariance between i and j.

Consider the three graphs of structural equation models depicted in Fig. 1. (We have labelled each edge in the graph by the corresponding coefficient in the set of linear equations.)

In Model I the tetrad difference $\gamma_{vw}\gamma_{xy} - \gamma_{vx}\gamma_{wy} = abcd\,\sigma^4_T - abcd\,\sigma^4_T = 0$. Note that in this case, the tetrad difference vanishes regardless of the values of the linear coefficients and the distributions of the independent variables. When a structural equation model robustly specifies a vanishing tetrad difference, as with Model 1 above, we say the model *strongly implies* the vanishing tetrad difference (or that the vanishing tetrad difference is *implied by the causal structure*). (For details on how tetrad differences can be calculated for a model, and how the graph determines whether a vanishing tetrad difference is implied see the following sections.

In Model II, the tetrad difference $\gamma_{vw}\gamma_{xy} - \gamma_{vx}\gamma_{wy} = cdf\sigma^2_T\sigma^2_{e1}$. In this case, the tetrad difference is positive regardless of the distributions of the independent variables, as long as the product of coefficients cdf is positive. In this case we say that given the sign of cdf, the tetrad difference is *strongly implied to be positive.*

In Model III, the tetrad difference $\gamma_{vw}\gamma_{xy} - \gamma_{v}\gamma_{wy} = cdf\sigma^2_T\sigma^2_{e1} + acg\sigma^2_T\sigma^2_{e4}$. This tetrad difference may be zero for particular values of the coefficients and variances, such as $a = d = f = g = \sigma^2_{e1} = \sigma^2_{e4} = 1$. But this constraint is not *robust* in Model III, because the tetrad difference does not vanish if the non-zero coefficients are varied in that model. And even if the signs of the coefficients were given, Model II does not imply that the tetrad difference is positive or negative.

TETRAD II is based upon the following fundamental methodological principles.

**Falsification Principle:** Other things being equal, prefer models that do not strongly imply constraints that are judged not to hold in the population.

**Explanatory Principle:** Other things being equal, prefer models that strongly imply constraints that are judged to hold in the population.

**Simplicity Principle:** Other things being equal, prefer simpler models (i.e. models with higher degrees of freedom).

As we have already seen, a model may fail to *strongly* imply a vanishing tetrad constraint, but still imply that constraint for particular values of its coefficients and variances. The intuition behind the Explanatory Principle is that an explanation of a constraint based on the causal structure of a model is superior to an explanation that depends upon the free parameters of a model coincidentally taking on values that happen to imply the constraint. This intuition has been widely shared in the natural sciences; it was used to argue for the Copernican theory of the solar system, the General Theory of Relativity, and the atomic hypothesis. One justification for the Explanatory Principle is provided in the following theorem. It states that under very plausible assumptions the probability of a vanishing tetrad difference being implied by a model, but not strongly implied by its causal structure, is zero.

**Theorem 3**: Let M be a linear model with n free linear coefficients $a_1,..., a_n$. and k variances $v_1,..., v_k$. Let M(U) be the class of models obtained by specifying values $U = \langle u_1,...,u_n \rangle$ for the parameters $a_1,..., a_n$. Let **P** be the set of probability measures P on the space $R^n$ of values of the parameters of model M such that for every subset S of $R^{n+k}$ having Lebesque measure zero, P(S) = 0. Let Q be the set of vectors of coefficient values such that for all U in Q every multinormal probability distribution consistent with M(U) has at least one vanishing tetrad constraint $\rho_{ij}\rho_{kl} - \rho_{il}\rho_{jk} = 0$ that is not implied by the causal structure of M. Then P(Q) = 0.

Analogous theorems hold for vanishing partial correlations.

Unfortunately, the principles can conflict with each other. Suppose, for example, that model M' is a modification of model M, formed by adding an extra edge to M. Suppose further that M' implies fewer constraints that are judged to hold in the population, but also implies fewer constraints that are judged not to hold in the population. Then M' is superior to M with respect to the Falsification Principle, but inferior to M with respect to the Simplicity and Explanatory Principles. TETRAD II introduces a scoring function that balances out the relative merits and demerits of models. The scoring functions is explained below in the section entitled Scoring.

We will now describe the two main sections of the "Suggested" command, the scoring function and the search.

### A.1.4 Constraints Judged to Hold In the Population

There are five types of constraints that TETRAD II uses. For each one, we will explain how TETRAD II judges whether or not it holds in the population.

#### A.1.4.1 Tetrad Constraints

HT0 (*H*old *T*etrad at *0*) is the set of tetrad differences judged to vanish in the population, HT+ is the set of tetrad differences judged to be positive in the population, and HT- is the set of tetrad differences judged to be negative in the population. We sort the tetrad differences into these classes in the following way.

First, we calculate the associated probability p(t) of a vanishing tetrad difference t. The *associated probability p(t)* of a tetrad difference t is the probability of obtaining a tetrad difference as large or larger than the one actually observed in the sample, under the assumption that the tetrad difference is zero in the population, and that the sampling distribution of tetrad differences is normal.[4] Wishart showed that the variance of the sampling distribution of the vanishing tetrad difference $\rho_{ij}\rho_{kl} - \rho_{il}\rho_{jk}$ is equal to

$$\frac{D_{12}D_{34}(N+1)}{(N - 1)(N -2)-D}$$

[4]The assumption that the sampling distribution of tetrad differences of normal covariates are normally distributed is approximately true in large samples.

where D is the determinant of the population correlation matrix of the four variables i, j, k, and l, $D_{12}$ is the determinant of the two-dimensional upper left-corner submatrix, and $D_{34}$ is the determinant of the lower right-corner submatrix, and i, j, k, and l have a joint normal distribution. In calculating p(t) we substitute the sample covariances for the corresponding population covariances in the formula. Given the variance of the tetrad difference, and the assumption that the sampling distribution of the tetrad differences are distributed normally, p(t) is determined by look up in a chart for the standard normal distribution.

Note that among any four distinct variables i, j, k, and l, we compute three tetrad differences:

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

If any two of these vanish in the population, the third is implied to vanish also. If p(t) is larger than the given significance level, we place t into HT0. If p(t) is smaller than the significance level, but the other two tetrad differences have associated probabiliies higher than the significance level we have conflicting evidence about whether t vanishes or not; in that case we do not put t into any class. If p(t) is smaller than the significance level, and at least one of the other two tetrad differences has an associated probability smaller than the significance level, we place t into HT+ or HT-, depending on the sign of the measured tetrad difference.

(Kenneth Bollen has recently discovered an asymptotically distribution-free sampling distribution for tetrad differences. We plan to replace the test we currently use and which assumes a multi-variate normal distribution with his new test.)

### A.1.4.2 Partial Correlation Constraints

We divide partial correlations into three classes. HP0 (*H*old *P*artial Correlation at *0*) is the set of partial correlations judged to vanish in the population, HP+ is the set of partial correlations judged to be positive in the population, and HP- is the set of partial correlations judged to be negative in the population.

Suppose K is a set of variables, and |K| is the cardinality of K. Again, we calculate the associated probability of a partial correlation $\rho_{ij.K}$ on the assumption that it vanishes in the population.

Fisher has shown that for a given partial correlation $\rho_{ij.K}$,

$$z = \frac{1}{2}\sqrt{s - 4}\ \ln\left(\frac{1 + |\rho_{ij.K}|}{1 - |\rho_{ij.K}|}\right)$$

has a standard normal distribution (where s is the sample size minus the cardinality of K). We use this transformation to calculate the associated probability of a given partial correlation x vanishing in the population. If the associated probability p(x) is greater than the significance level then x is placed in HP0; otherwise it is placed in HP- or HP+ depending upon the sign of the measured partial correlation. A vanishing correlation constraint is a special case of a vanishing partial correlation, with |K| = 0.

(We plan to replace the test we currently use and which assumes a multi-variate normal distribution with an asymptotically distribution-free test.)

### A.1.5 Calculating the Implied Constraints

In order to explain how the various kinds of constraint implications are calculated, we will introduce the following terminology.

**Definition**

Given an ordered n-tuple $N = \langle c_1, \ldots, c_n \rangle$, an object **o is in N** iff $o = c_i$ for some i between 1 and n inclusive. We shall also write that $o \in N$.

This notation is somewhat ambiguous since $\varepsilon$ is also used to mean set membership, but the context will always make it clear which use of $\varepsilon$ is intended.

**Definition**

A **digraph** is an ordered pair <R,E>, where R is a set of **vertices** and E is a set of **edges**. Each edge is an ordered pair of elements of R. The first element in an edge is called the **tail**, and the second element is called the **head**. An edge with a tail $v_i$ and a head $v_j$ is an edge **from** $v_i$ to $v_j$; it is also said that the edge is **out of** $v_i$ and **into** $v_j$. $v_i$ is **adjacent** to $v_j$ iff there is an edge from $v_i$ to $v_j$. **Adj(i)** is the set of all variables adjacent to i. The **indegree** of a vertex v is equal to the number of distinct edges into v; the **outdegree** of a vertex v is equal to the number of distinct edges out of v.

**Definition**
A **path of length n** in a digraph <R,E> is an ordered n+1-tuple of vertices $<v_1, \ldots, v_{n+1}>$ where for $1 \leq i \leq n$, $<v_i, v_{i+1}>$ is an edge in E. The path is said to **contain** edge $<v_i, v_{i+1}>$. The first vertex in the path is called the **source** of the path; the last vertex in the path is called the **sink** of the path. The path is said to **connect** the source to the sink. Two paths **intersect** iff they have a a vertex in common; any such intersection (of paths) common vertex is a **point of intersection**. A **cycle** is a path of at least length 1 in which the source equals the sink. A path **contains a cycle** iff it has a subpath which is a cycle. An **open path** is a path with no cyclic subpaths. A digraph is **acyclic** if and only if every path in the graph is open. A path with one vertex is an **empty path**. If path p is equal to $<v_1, \ldots, v_n>$ and path q is equal to $<v_n, \ldots, v_{n+m}>$, then **the concatentation of p and q** is equal to $<v_1, \ldots, v_n, \ldots, v_{n+m}>$ and is denoted by p&q. Note that empty paths are the only paths that contains no edges. Also the concatenation of p with an empty path is p, and the concatenation of an empty path with p is p. The single vertex in an empty path is both its source and its sink.

**Definition**
A trek t between two distinct vertices $v_i$ and $v_j$ is a pair of open paths from some vertex u to $v_j$ and $v_i$ respectively that intersect only at u. The source of the paths in the trek is called the **source** of the trek. $v_i$ and $v_j$ are called the **termini of the trek**. Given a trek $t_{ij}$ between i and j, $\mathbf{i(t_{ij})}$ will denote the path in t from the source of t to i and $\mathbf{j(t_{ij})}$ will denote the path in t from the source of t to j. $\mathbf{P_{ij}}$ is the set of all paths from i to j. $\mathbf{T_{ij}}$ is the set of all treks between i and j.

One of the paths in a trek may be an empty path. However, since the termini of a trek are distinct, only one path in a trek can be empty.

**Definition**
Suppose there are two treks $t_{ij}$ and $t_{kl}$ such that $i(t_{ij}) \cap k(t_{kl}) = 0$, and $j(t_{ij}) \cap l(t_{kl}) = 0$, and either $i(t_{ij}) \cap l(t_{kl}) = 0$, or $j(t_{ij}) \cap k(t_{kl}) = 0$. If $i(t_{ij}) \cap l(t_{kl}) \neq 0$, let $\mathbf{Overlap(t_{ij},t_{kl})}$ equal the product of the label of edges in either $i(t_{ij})$ or $l(t_{kl})$ between the first and last points of intersection of $i(t_{ij})$ and $l(t_{kl})$, and $\mathbf{Non\text{-}overlap(t_{ij},t_{kl})}$ equal the product of the labels of all the other edges in $t_{ij}$ or $t_{kl}$. If $i(t_{ij})$ intersects $l(t_{kl})$ in only one point, or both $i(t_{ij}) \cap l(t_{kl}) = 0$ and $j(t_{ij}) \cap k(t_{kl}) = 0$, then $Overlap(t_{ij},t_{kl}) = 1$, and $Non\text{-}overlap(t_{ij},t_{kl}) = L(t_{ij})L(t_{kl})$. If $j(t_{ij}) \cap k(t_{kl}) \neq 0$, then Overlap and Non-overlap can be defined in an analogous way.

The implications of a model are calculated in the following way.

### A.1.5.1 Vanishing Tetrad Constraints

The calculation of vanishing tetrad constraints implied by a model S is based on the following theorem.

**Theorem.**
An acyclic model S strongly implies that $\gamma_{ij}\gamma_{kl} - \gamma_{il}\gamma_{jk}$ vanishes iff for every $t_{ij} \varepsilon T_{ij}$ and $t_{kl} \varepsilon T_{kl}$ either

$i(t_{ij}) \cap k(t_{kl}) \neq \{\}$, or

$l(kl) \cap j(t_{ij}) \neq \{\}$

and for every $t_{il} \varepsilon T_{il}$ and $t_{jk} \varepsilon T_{jk}$ either

$i(t_{il}) \cap k(t_{jk}) \neq \{\}$, or

$l(t_{il}) \cap j(t_{jk}) \neq \{\}$.

We conjecture that the theorem is also true for cyclic models.

### A1.5.2 Vanishing Correlation Constraints

The calculation of vanishing correlation constraints implied by a model S is based on the following theorem.

**Theorem.**
S strongly implies that $\rho_{ij}$ vanishes iff there is no trek between i and j.

### A.1.5.3 Positive Tetrad Constraints

The calculation of vanishing correlation constraints implied by a model S is based on the following conjecture.

**Conjecture**
Given an assignment of signs to a model S, S does not strongly imply that $\gamma_{ij}\gamma_{kl} - \gamma_{i-j}\gamma_{jk} > 0$ relative to that sign assignment iff there exists a $t_{ij} \varepsilon T_{ij}$ and $t_{kl} \varepsilon T_{kl}$ such that

$i(t_{ij}) \cap k(t_{kl}) = \{\}$, and

$j(t_{ij}) \cap l(t_{kl}) = \{\}$, and

either $i(t_{ij}) \cap l(t_{kl}) = \{\}$, or $j(t_{ij}) \cap k(t_{kl}) = \{\}$, and

either Non-Overlap($t_{ij}$,$t_{kl}$) is negative, or there exists a pair of paths $p_1$ and $p_2$ from some independent variable I to the source of one of the treks $t_{ij}$ or $t_{kl}$ (call it $t_1$) such that $p_1$ intersects exactly one branch of the other trek (call it $t_2$) at a

point y, $p_2$ does not intersect that branch of $t_2$, and sign($p_{1xz}$) <> sign($p_{2xz}$) (where x is the last point of intersection of p1 and p2 before y, and z is the first point of intersection of p1 and p2 after y),

or there exists a $t_{il} \sim T_{il}$ and $t_{jk} \varepsilon T_{jk}$ such that

$i(t_{il}) \cap k(t_{jk}) = \{\}$, and

$i(t_{jk}) \cap l(t_{il}) = \{\}$, and

either $i(t_{il}) \cap j(t_{jk}) = \{\}$, or $l(t_{il}) \cap k(t_{jk}) = \{\}$, and

either Non-Overlap($t_{il}$,$t_{jk}$) is positive, or there exists a pair of paths $p_1$ and $p_2$ from some independent variable I to the source of one of the treks $t_{il}$ or $t_{jk}$ (call it $t_1$) such that $p_1$ intersects exactly one branch of the other trek (call it $t_2$) at a point y, $p_2$ does not intersect that branch of $t_2$, and sign($p_{1xz}$) <> sign($p_{2xz}$) (where x is the last point of intersection of $p_1$ and $p_2$ before y, and z is the first point of intersection of $p_1$ and $p_2$ after y).

### A.1.6 The Tetrad-Score

Using the above theorems we classify each tetrad difference as IT0 (*I*mplied *T*etrad *0*), IT+, IT- (or none of these), and each correlation as IC0 (*I*mplied *C*orrelation *0*), IC+, IC- (or none of these). (The partial correlation constraints are not incorporated into the TETRAD-score. They are used, however, by the "Partial" command, which implements the Path Model Construction Algorithm. See the section entitle "Partial".) However, the user can ask for a list of all of the partial correlations which indicates whether they hold at 0, +, or -, and whether or not they are implied to be 0, +, or -).

Then we define

$$\textit{Tetrad-score} =$$

$$\sum_{c \varepsilon IC0 \cap HC0} p(c) + \sum_{t \varepsilon IT0 \cap HT0} p(t) -$$

$$\textit{weight} * \left( \sum_{c \varepsilon IC0 - HC0} p(c) + \sum_{t \varepsilon IT0 - HT0} p(t) + \sum_{t \varepsilon IT+ \cap HT-} p(t) + \sum_{t \varepsilon IT- \cap HT+} p(t) \right)$$

The first two terms implement the Explanatory Principle since they gives credit for explaining vanishing residuals that are judged to hold in the population. The rest of the

terms implement the Falsification Principle since they penalizes a model for making predictions about residuals that are judged not to hold in the population. The Simplicity Principle is implemented by preferring, among models with identical Tetrad-scores, those that have more degrees of freedom.

The weight determines how conflicts between the Explanatory and Falsification Principles are resolved by determining the relative importance of explanation versus residual reduction. The higher the weight, the less important explanation is relative to residual reduction.

### A.1.7 Search

The "Suggested" command employs a relatively fast search algorithm that examines many more plausible models than the searches employed by either LISREL VI or EQS. The search is fast for three main reasons.

First, there are well known, fast algorithms for analyzing directed graphs, algorithms that we have modified to determine the set of all vanishing tetrad differences implied by a model.

Second, most of the computational work required to evaluate a model M can be stored and re-used to evaluate elaborations of M.

Finally, the scoring function is such that if a model M can be conclusively eliminated from consideration because of a poor score, so can any elaboration of M.

The search procedure is a recursive procedure that is difficult to describe in non-technical English. The following rough outline of the procedure gives the basic structure of the search. At each point of the search, Top_score represents the best TETRAD-score of any model examined up to that time; Local_Top_score represents the best TETRAD-score of any elaboration of Initial Model; percentage represents a user-settable parameter that controls the width of the search; Good represents the list of models that are among the best found so far; and L represents the list of all elaborations of Initial Model. The procedure is first called on the initial model supplied by the user.

```
Procedure  Elaborate_Model(Initial_Model);
begin
    Generate all one edge elaborations of Initial Model;
    Place all one edge elaborations of Initial Model in list L in
            order of decreasing TETRAD-score;
    Set Local_top_score to highest TETRAD-score in L;
    for each model M  in L do
        begin
            if (TETRAD-score(M) > Percentage * Local_top_score)) and
               (TETRAD-score(M)> Percentage * TETRAD-score(Initial Model)) then
                  Elaborate_Model (M);
            if TETRAD-score(M) > Percentage * Top_score then place M in list Good;
            if TETRAD-score(M) > Top_score then
                  set Top_Score to TETRAD-score(M);
        end;
```

This procedure is much faster than previous search algorithms that we have employed, but it is also theoretically less reliable than those algorithms. In a series of Monte Carlo simulations, however, it has proved to be equally reliable in practice.

The user can further constrain the search by use of the knowledge file. If an edge is required in the knowledge file, TETRAD II will add it to every initial model, before it begins its search. If an edge is forbidden in the knowledge file, TETRAD II will never add it to any model in the search. If the user specified that the graph must be acyclic, TETRAD II will never add any edges to the initial model that create cycles that are not in the initial model. If the user specifies that no direct cycles (edges from A to B and B to A) can occur in the model, then TETRAD II will never add any edges that create direct cycles that were not in the initial model. The user can also specify a maximum depth for the search, and a maximum number of times TETRAD II will add edges that fail to improve the score.

Up to this point, the scores for all models have been calculated on the assumption that the signs of the additional edges suggested by TETRAD II are not known. Models that are indistinguishable on the basis of the vanishing tetrad differences that they strongly imply are sometimes distinguishable on the basis of the positive or negative tetrad

differences that they strongly imply. If the signs of the edges in the initial model are not known, we simply suggest among all of the models that are tied for the highest Tetrad-score, those that have the fewest edges. But, if the signs of the edges in the initial model are known, after we have generated a list of suggested models with our search, we calculate scores for every possible combination of sign assignments for the suggested edges. This may reduce the scores of some models, since they could imply positive or negative tetrad differences that are judged not to hold in the population. The models that no longer are tied for the highest score are eliminated from the list, and those remaining are suggested by TETRAD II.

## A.2 The "Partial Command"

The Path Model Construction Algorithm is implemented in the "Partial" command (so-called because we judge conditional independence by performing a statistical test upon vanishing partial correlations.) In order to justify the algorithm, we need to introduce the following definitions.

**Definition:** An **undirected path** in a directed graph is a is an ordered n+1-tuple of vertices $<v_1, \ldots, v_{n+1}>$ where for $1 \leq i \leq n$, either $<v_{i+1}, v_i>$ or $<v_i, v_{i+1}>$is an edge in the graph.

If an undirected path p contains edges from $v_{k-1}$ to $v_k$ and $v_{k+1}$ to $v_k$, then $v_k$ is a **p-collider**.

If there is an undirected path p such that $v_k$ is a p-collider, then $v_k$ is a **collider**.

There is an **edge between i and j** iff there is an edge from i to j or an edge from j to i.

A **directed path of length n** in a directed graph <R,E> is an ordered n+1-tuple of vertices $<v_1, \ldots, v_{n+1}>$ where for $1 \leq i \leq n$, $<v_i, v_{i+1}>$ is an edge in the graph.

A directed graph G is **acyclic** (or a **DAG**) iff on every directed path in G, no vertex occurs more than once.

In a DAG G, if an undirected path Y between i and j is such that every head-on vertex in Y has a descendant in a set of vertices s, and no non-head-on vertex in V is in s, then Y is a **dependency-making path between i and j relative to s**.

In a DAG G, **I(x,S,y)** iff x and y are d-separated by the set of vertices S. In a probability distribution P, **I(x,S,y)$_P$** iff in P x and y are independent conditional on S.

A vertex y is a **descendant** of a vertex x if and only if there is a directed path from x to y. (Since the empty path from j consists simply of j, every vertex is a descendant of itself.)

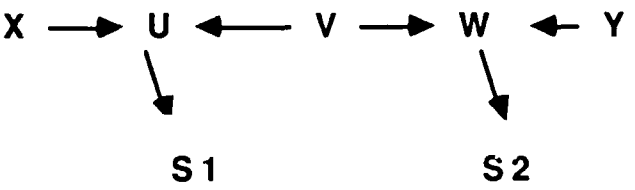
Following Pearl, we say that variables x, y are **d-separated** by set S if and only if there exists no undirected path U between x and y, such that (i) every collider on U has a descendent in S and (ii) no other vertex on U is in S. We say that x, y are **d-connected** with respect to S if and only if they are not d-separated with respect to S. We say that two sets X, Y of variables are d-separated by S if and only if every pair <x,y> in the cartesian product of X and Y is separated by S.

With these definitions we can state Pearl's idea, which can be given as a definition:

**Pearl's Representation Definition:** Let G be a directed acyclic graph and let P be a joint probability distribution on the vertices of G. G **perfectly represents** P if and only if I(X,Z,Y)$_P$ if and only if D(X,Z,Y)$_G$.

An investigation of the properties of this notion of representation may be found in Pearl(1988). For every directed acyclic graph there exists a probability distribution that it perfectly represents.

An illustration of d-connectedness is given in the following graph:




X and Y are not d-connected with respect to the empty set
X and Y are d-connected with respect to the set {S1, S2}
X and Y are not d-connected with respect to the set {S1, S2, V}

X and Y are independent
X and Y are not independent conditional on {S1, S2}
X and Y are independent conditional on {S1, S2, V}

Think of a mechanical device or electrical circuit arranged so that the variables in the graph have the causal relations illustrated. If you know the value of X and nothing else, it provides you with no information about the value of Y. If you know values of $S_1$ and $S_2$, then information about the value of X will give you additional information about the value of Y. If you know the value of $S_1$, $S_2$ and V--or even just the value of V, then information as to the value of X tells you nothing further about the value of Y.

Theorem 4 states that for linear normal theories, under a wide variety of plausible probability distributions over the free parameters of the theory, the probability that two variables are conditionally independent, but not implied to be conditionally independent by their causal structure, is zero.

**Theorem 4**: Let M be a linear model with n free linear coefficients $a_1,..., a_n$. and k variances $v_1,..., v_k$. Let M(U) be the model obtained by specifying values $U = <u_1,...,u_n, u_{n+1},...,u_{n+k}>$ for $a_1,..., a_n$ and $v_1,...v_k$. Let **P** be the set of probability measures P on the space $R^{n+k}$ of values of the parameters of model M such that for every subset S of $R^{n+k}$ having Lebesque measure zero, P(S) = 0. Let Q be the set of vectors of coefficient and variance values such that for all U in Q every multinormal probability distribution consistent with M(U) has at least one statistical independence relation not represented in the directed acyclic graph of M according to d-separability. Then P(Q) = 0.

The following two theorems prove that the Path Model Construction Algorithm constructs a perfect representation of the data, if there is one.

**Theorem 5:** If probability distribution P over a set of variables V is perfectly representable, then G perfectly represents P if and only if:

1. there is an edge between vertices A and B in G if and only if $\sim I(A,S,B)_P$ for all $S \subseteq V$ containing neither A nor B;
2. in G there is an edge between vertices A and B, and an edge between vertices B and C, but no edge between vertices A and C, there is an edge from A to B and from C to B if and only if $\sim I(A,S,C)_P$ for all $S \subseteq V$ containing B, but containing neither A nor C.

The justification of our second test for the existence of latent variables is provided by the following Theorem.

**Theorem 6:** The causal structure G of a linear model M implies a vanishing tetrad difference $\rho_{ij}\rho_{kl} - \rho_{il}\rho_{jk} = 0$ only if the causal structure of G implies either $\rho_{ij}$ or $\rho_{kl} = 0$ , and $\rho_{il}$ or $\rho_{jk} = 0$, or that there is a non-empty set q of random variables in G such that $\rho_{ij.q} = \rho_{kl.q} = \rho_{il.q} = \rho_{jk.q} = 0$.

## References


Glymour, C., Scheines, R., Spirtes, P., and Kelly, K., *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling*, Academic Press, San Diego, 1987.

Harary, F., and Palmer, E., *Graphical Enumeration*, Academic Press, New York, 1973.

Joreskog, K., and Sorbom, D., *LISREL VI: User's Guide*, Scientific Software, Mooresville, IN, 1984

Pearl, J., *Probabilisitic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers Inc., San Mateo, CA, 1988

Pearl J. and Dechter, R. Learning Structure from Data: A Survey" in *Proceedings Colt '89*, pp. 230-244.

Pearl J. and Tarsi, M. "Structuring Causal Trees" Journal of Complexity , 2, (1986), pp. 60-77.

Rodgers, R.C., and Maranto, C.L., Causal Models of Publishing Productivity in Psychology, *Journal of Applied Psychology*, Vol. 74, No. 4, pp.636-649., 1989.

Spirtes, P., Scheines, R., and Glymour, C., "Simulation Studies of the Reliability of Computer-Aided Model Specification Using the TETRAD II, EQS, and LISREL Programs", forthcoming in *Sociological Methods and Research*, Sage, 1989.

Spirtes, P., Glymour, C., and Scheines, R., "Causality from Probability," in *Evolving Knowledge in Natural and Artificial Intelligence*, ed. by J. Tiles, Pittman, UK, forthcoming (1990).

Spirtes, P., A Necessary and Sufficient Condition for Conditional Independencies to Imply a Vanishing Tetrad Difference, Technical Report Number CMU-LCL-89-3, Dept. of Philosophy, Carnegie Mellon University, Pittsburgh, PA, 15213.

Whitney, H. "Elementary Structures of Real Algebraic Varieties" *Annals of Mathematics*, 66, 1957, 545-556.

Technical Report Distribution List - Manpower, Personnel, and Training R&D Program

As of 11/1/90
One copy to each addressee except as noted:

Director, Contract Research Department
Office of Naval Research (Code 11)
Arlington, VA 22217-5000

Chairman, MPT R&D Committee
Office of the Chief of Naval Research
Code 222
Arlington, VA 22217-5000

Program Manager, Operations
Research (Code 1111MA)
Office of Naval Research
Arlington, VA 22217-5000

Director, Life Sciences (Code 114)
Office of Naval Research
Arlington, VA 22217-5000

Director, Cognitive & Neural Sciences
(Code 1142)
Office of Naval Research
Arlington, VA 22217-5000

Cognitive Science (Code 1142CS)
Office of Naval Research
Arlington, VA 22217-5000

Perceptual Science (Code 1142PS)
Office of Naval Research
Arlington, VA 22217-5000

Defense Technical Information Center*
DTIC/DDA-2
Cameron Station, Building 5
Alexandria, VA 22314

CDR J. S. Hanna, Office of the Deputy
Asst. Secretary of the Navy (Manpower)
5D800, The Pentagon
Washington, DC 20350-1000

Head, Manpower, Personnel, and
Training Branch
Office of the CNO (Op-813)
4A478, The Pentagon
Washington, DC 20350-1000

Assistant for Manpower and Training
Office of the CNO (Op-911H)
5D772, The Pentagon
Washington, DC 20350-2000

Assistant for Planning and Technology
Development
Office of the DCNO(MPT) (Op-01B2)
Department of the Navy, AA-1822
Washington, DC 20350-2000

Deputy Director Total Force Training
and Education Division
Office of the DCNO(MPT) (Op-11B)
Department of the Navy
Washington, DC 20370-2000

R&D Coordinator, Attn: Jan Hart
Office of the DCNO(MPT) (Op-11K1)
Department of the Navy, AA-G817
Washington, DC 20370-2000

Deputy Director Military Personnel
Policy Division
Office of the DCNO(MPT) (Op-13B)
Department of the Navy, AA-1825
Washington, DC 20370-2000

Head, Military Compensation Policy
Branch
Office of the DCNO(MPT) (Op-134)
Department of the Navy, AA-2837
Washington, DC 20370-2000

*Note: 12 copies go to DTIC

Headquarters U.S. Marine Corps
Code MA
Washington, DC 20380-0001

Head, Leadership Branch
Naval Military Personnel Command
Attn: LT Gary Kent, NMPC-621
Department of the Navy, AA-1603
Washington, DC 20370-5620

Director, Research & Analysis Division
Navy Recruiting Command (Code 223)
4015 Wilson Boulevard, Room 215
Arlington, VA 22203-1991

Technical Director Silva
Attn: Dr. Kobus
Naval Health Research Center
P.O. Box 85122
San Diego, CA 92138-9174

Head, Human Factors Division
Naval Training Systems Ctr. (Code 26)
12350 Research Parkway
Orlando, FL 32826-3224

Naval Training Systems Center
ATTN: Dr. Eduardo Salas, (Code 262)
12350 Research Parkway
Orlando, FL 32826-3224

Naval Training Systems Center
ATTN: Dr. Robert Hays, (Code 262)
12350 Research Parkway
Orlando, FL 32826-3224

Technical Director
Navy Personnel R&D Center
Code 01
San Diego, CA 92152-6800

Director, Manpower Systems Dept.
Code 11
Navy Personnel R&D Center
San Diego, CA 92152-6800

Director, Personnel Systems Dept.
Code 12
Navy Personnel R&D Center
San Diego, CA 92152-6800

Director, Testing Systems Department
Navy Personnel R&D Center
Code 13
San Diego, CA 92152-6800

Director, Training Systems Dept.
Code 14
Navy Personnel R&D Center
San Diego, CA 92152-6800

Director, Training Technology Dept.
Code 15
Navy Personnel R&D Center
San Diego, CA 92152-6800

Director, Organizational Systems Dept.
Code 16
Navy Personnel R&D Center
San Diego, CA 92152-6800

Naval Ocean Systems Center
Command Support Technology Division
Attn: Mr. Jeffrey Grossman, Code 4402
San Diego, CA 92152-5000

Chairman, Dept. of Admin. Sciences
(Code AS)
Naval Postgraduate School
Monterey, CA 93943-5100

Chairman, Department of Operations
Research (Code OR)
Naval Postgraduate School
Monterey, CA 93943-5100

Director, Instructional Development and
Educational Program Support Dept.
Naval Education and Training Program
Management Support Activity
(NETPMSA)
Pensacola, FL 32509-5100

Academic Programs and Research
Branch
Naval Technical Training Command
Code N6, NAS Memphis, Bldg. C-1
Millington, TN 38054-5056

Director, Defense Personnel Security
Research and Education Center
Suite E, Building 455
99 Pacific Street
Monterey, CA 93940-2481

Technical Director
U.S. Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Chief Scientist
Air Force (AFHRL)
Brooks AFB, TX 78235

Director, Manpower & Training Program
Center for Naval Analyses
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Library- Code 231
Navy Personnel R&D Center
San Diego, CA 92152-6800

Library
Naval War College
Newport, RI 02940

Chief, Survey and Market
Analysis Division
Defense Manpower Data Center
1600 Wilson Boulevard, #400
Arlington, VA 22209-2593

Program Director
Manpower Research & Advisory Services
Smithsonian Institution
801 North Pitt Street, Suite 120
Alexandria, VA 22314-1713

Dr. Steven Sorensen
Navy Personnel Research and Development Center
217 Catalina Blvd.
San Diego, CA 92152-6800