RUNNING HEAD: Causal models

A philosophical investigation of causal interpretation in structural equation models

Chong Ho Yu, Ph.D.

Arizona State University

April, 2002

Paper presented at the Annual Meeting of American Education Researcher Association,

New Orleans, LO

and published in *Research Method Forum*

Chong Ho Yu, Ph.D., CCNA, MCSE, CNE

Arizona State University

331 West Musket Place

Chandler AZ 85248

Phone: 602-778-2722

Email: asumain@yahoo.com.hk

Abstract

This paper is a brief overview and evaluation of current mathematical/statistical causal models, including the structural equation model (SEM), TETRAD, and the graphical model. The efficacy of these approaches will be discussed in the philosophical context of the Duhem-Quine thesis, realism, simplicity, identifiability (testability), empirical adequacy, and probabilistic causality. The emphasis of this paper is on the philosophical aspect, not the mathematical or computational aspect of SEM, nonetheless, readers are not required to have a philosophical background to follow the arguments.

A philosophical investigation of causal interpretation in structural models

Chong Ho Yu

This paper is a brief overview and evaluation of current mathematical/statistical causal models, including the structural equation model (SEM), TETRAD, and the graphical model. The efficacy of these approaches will be discussed in the philosophical context of the Duhem-Quine thesis, realism, simplicity, identifiability (testability), empirical adequacy (it means fitness, but it has nothing to do with gymnastics), and probabilistic causality. It is argued that latent factors, which are the building blocks of causal models, could be legitimately interpreted as real entities and thus should be treated as seriously as observed items. Another component of the above causal models is the path model, which is a linear approximation. The path model is criticized as an over-simplification of the empirical world, which is said to be non-linear in nature. However, computational tractability, simplicity, and fitness together provide a strong justification for causal models. Although the untested assumptions in these causal models are challenged by critics, these models are good tools for causal analysis based upon partial knowledge. Given the consideration of realness, simplicity, and fitness, and the validity of probabilistic causation, both SEM and the graphical model are adequate to answer the Duhem-Quine question.

Weaknesses of theorization and experimentation

In everyday life, both scholars and non-scholars try to "theorize" things that happen around themselves. However, based upon empirical studies, psychologist Baron (2000) found that this theorization of causality is often flawed due to selection bias, prior belief, and the interaction of both. To be specific, people tend to pay attention to facts that confirm their prior belief regarding a particular issue.

Hoyle (1995) also asserted that use of theory is the most problematic approach to identify causal relationships, for usually there are competing theories that seem to be equally adequate in casual explanation. Hoyle considered research design the most powerful mean for generating casual

inferences, because a good research design could rule out rival hypotheses. This notion has been widely adopted by social scientists and statisticians. Classical books on experimental design (e.g. Campbell & Stanley, 1963, Cook & Campbell, 1979, Kerlinger, 1986) emphasized that in experimental settings researchers could exercise a high degree of control and manipulation of various factors. If threats against internal validity and external validity are controlled, and error variances are suppressed, then it is possible to rule out rival explanations. It is generally agreed that in experimental settings strong causal inferences could be made, whereas in quasi-experiments causal inferences are weaker but still possible. However, in non-experiments correlation or association should be reported as descriptive findings only.

Although experimental design could remediate some flaws of theorization, it is by no mean bulletproof. French physicist and philosopher Duhem (1954) said that usually a complex array of variables, hypotheses, and auxiliary assumptions may be involved in a study. Even if a complex set of theories is rejected, the theory remains inconclusive. For associationists such as Karl Pearson, this is a typical argument that relationships may be spurious and thus causal inferences cannot be affirmed. Following the thread of the Duhem's notion, Quine (1976) went even further to say that if some ad hoc assumptions are altered or added, any disputed theories could be accepted. The combination of Duhem's and Quine's notions was termed as the "Duhem-Quine thesis." This thesis accurately points out some potential problems of experimentation. Even though the experimenter could take as many variables into account as possible, reduce as many error variances as possible, and maximize the experimental error, the interaction of all variables and remaining noise together could still make the research question unsettled. Further, many issues are not subject to experimental manipulation. For example, it is unethical for the experimenter to assign a sample to a smoking group and another to a non-smoking group to study whether cancer and smoking is causally related. Consequently, the notion of the experimental school confines some issues into the domain of association only (non-causality).

In recent years, mathematical approaches were proposed as tools to strengthen causal inferences. The structural equation model, as well as TETRAD and the graphical model, which are extensions of

SEM, are noticeable "causationist" schools. In the following sections, the characteristics of these schools of thought will be introduced and the philosophical issues related to these schools will be discussed.

Structural equation modeling

Structural equation model has gained popularity among social scientists since 1970s. According to Pearl (2000), the causal elements of SEM are not paid much attention by researchers. Economists view structural models as convenient representations of density functions, and social scientists see them as summaries of covariance matrices. For over a decade, both Pearl (2000, in press) and Glymour & Cooper (1999) have devoted much effort to reinstate the causal interpretation of SEM.

In conventional experiments that involve many variables and relationships, researchers may perform several separate ANOVA and regression analyses. SEM is a different approach, in which variables are organized in a structural fashion. SEM is a synthesis of the latent factor model and the structural model, which will be introduced next.

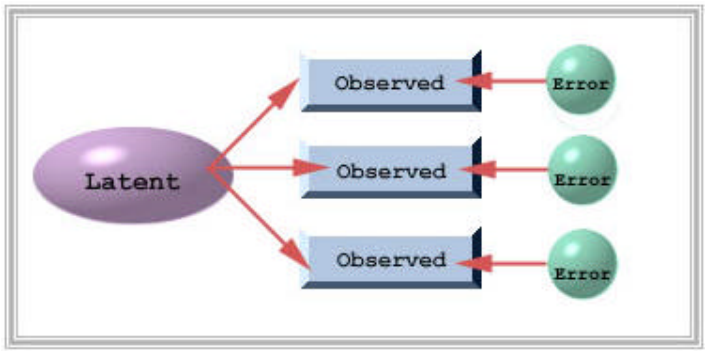
Factor model

The latent factor model is also known as the latent construct model, the latent variable model, or the measurement model. The term "latent variable model" may be misleading since "variables" are usually referred to as observed items while "factors" and "constructs" are referred to as theoretical entities. Thus, throughout this paper the term "latent factor " or "latent construct" is used instead.

A measurement model, as its name implies, is about measurement and data collection. A factor model identifies the relationship between observed items and latent factors. For example, when a psychologist wants to study the causal relationships between anxiety and job performance, first he/she has to define the constructs "anxiety" and "job performance." To accomplish this step, the psychologist employs Cronbach Alpha to evaluate the internal consistency of observed items (Yu, 2000), and also applies factor analysis to extract latent constructs from these consistent observed variables. If the factor structure indicates that observed items cluster around one eigenvector, which is the graphical representation of factors in subject space, the construct is said to be uni-dimensional.

The relationship between factors and observed variables are indicated in Figure 1. The ellipse represents a latent construct, and the rectangles represent observed variables, which are individual items in a scale. The circles denote measurement errors.

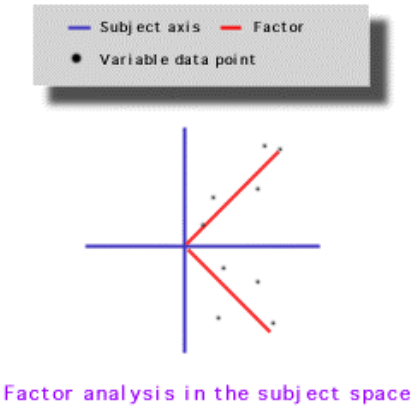
Figure 1. Factor model



It is important to note that every measurement includes some degree of measurement error. The purpose of Cronbach Alpha is to estimate the test score reliability (r) in terms of the measurement error (e). In other words, reliability and the measurement error is in a "see-saw" relationship (r = 1 – e). Moreover, factor analysis is a form of triangulation in attempt to minimize the measurement error. For example, if an instructor gives only one question in a test and judges whether the student is competent based on a single item response, even someone who does not have any statistical training could tell that this assessment is unfair. By giving more items in a test, the measurement errors would cancel out each other and thus the test score would become more reliable.

The relationship between factors and observed items are defined by factor loadings, which are computed based on item correlation or covariance matrices. The sum of squares of factor loading can be converted into eigenvectors in subject space. In subject space, the length of vectors indicates the value of factor loadings, and the angle between vectors indic ates inter-factor relationship (see Figure 2).

Figure 2. Vectors in subject space



Factors are the building blocks of SEM. Before proceeding to the path model, we have to examine why using latent factors is justified. There are two major claims involved in using the factor model. First, latent factors should be interpreted as real entities. Second, constructs and observed items are causally related.

Realism. Realism and anti-realism has been an endless debate among philosophers. Why is it important in the context of causality? It is because in order to establish the causal and effect relationship of entities, those entities must be perceived as real as observed variables. In an episode of "Star Trek," it is amusing to see how the photon fluctuation in a warp engine "causes" the starship to explode. However, the so-called cause and effect in science fiction is just a 3D animation effect generated by computers. There is no photon fluctuation. Neither the warp engine nor the starship is real. Thus, it is not amusing to see an engineer seriously talking about the causal relationship between photon fluctuation and warp engine. He needs psychotherapy! By the same token, if a psychologist talks about how depression causes poor job performance but he/she treats the constructs "depression" and "performance" as fictitious, his misuse of language is just like talking about how photon instability causes malfunctioning of a warp engine. If "depression" and "performance" are not treated as real entities, then at most the psychologist could issue a statement like "the factor called X that

summaries A to D behaviors and the one called Y that summaries E to F behaviors has a negative correlation." If correlation alone is adequate for theorizing, researchers could arbitrarily name factors as X, Y, Z, A, B, C, rather than assigning names that are conceptually sensible and isomorphic to the empirical world (e.g. anxiety, depression, obsession, performance). When sensible names are assigned to factors, the hidden assumption is that there exists some degree of mapping between the theoretical and empirical worlds.

Borsboom, Mellenbergh, & van Heerden (in press) also hold a realist position regarding latent factors. They compared and contrasted the operationalist and realist positions, and argued that the latter is more acceptable than the former. Operationalism is a form of anti-realism which maintains a sharp distinction between theory and observations, and theoretical constructs are nothing more than instruments for the sake of operational convenience only. Although operationalists view the latent construct as nothing more than a numeric trick to simplify the observations (collapsing many observed items into one factor), Borsboom et al assert that operationalism and the latent construct theory are fundamentally incompatible. If a latent construct is just for operational convenience, then there should be a distinct latent factor for every single test researchers construct. Upon the operationalist view, it is even impossible to formulate the requirement of uni-dimensionality. As a result, operationalists would have difficulties making sense of item response theory, which is a special case of factor analysis and assumes one single trait in the measurement. Borsboom et al argued that realism is typically associated with causality. If latent factors are real rather than operational, then latent factors are causally responsible for observed items.

The above argument seems to be unconvincing and open to counterattack. Empirically speaking, very often constructs do not demonstrate uni-dimensionality even though the theory said so. Reliability in terms of internal consistency, which is often measured by Cronbach Alpha, is a necessary but not sufficient condition for uni-dimensionality. However, meta-analyses of reliability across studies, also known as reliability generalization studies, indicated that reliability information could fluctuate from sample to sample (Vacha-Hasse, 1998). No wonder Thompson and Vacha-Haase

(2000) went even further to proclaim that "psychometrics is datametrics" (p.174). In other words, reliability attaches to the data rather than the psychological test. Moreover, Kelley (1940) also warned that constructs resulting from factor analysis are not timeless, spaceless, populationless truth.

At first glance, factors had better been interpreted in the context of operationalism. However, fluctuation in the measurement model does not necessarily deny the realness of constructs. In physical sciences, the measurement of tangible things also leads to inconsistent results. As indicated in Figure 1, the factor model takes measurement errors into consideration. In the meta-analyses mentioned above, although some inconsistency of reliability was found, those researchers were still able to make generalizations about reliability. If constructs are entirely operational, there should be no need to conduct meta-analyses at all, and studying generalization is a waste of time. Every researcher could write his/her own survey items and invent his/her own construct in each individual study. The hidden assumption of generalization study is that there are certain invariant elements in constructs in spite of measurement errors.

Causal relationships. The next issue to be addressed is the causal relationship. How could the causal relationship between factors and observed items be confirmed? One may argue that by mathematics alone, the causal relationship cannot be established. No matter how high the factor loadings are and how stable the factor structure appears to be, it seems to be a leap of faith to claim the clustering as a causal phenomenon.

Take planet clustering and motion as a metaphor. When astronomers observe that there are nine planets orbiting around the sun in a solar system, they could theorize that a hidden force causes the planets to behave in this manner. This causal claim is data-driven rather than a leap of faith. By the same token, the clustering of observed items around an eigenvector is as empirical as the clustering of planets around a solar system. Although we cannot see forces of orbits, multiple observations of planetary movements imply the existence of the gravitational forces. Thurstone (1947), an early psychometrican who co-developed factor analysis with other researchers, also used the analogy of forces in physics to support the use of factors: "A simple example is the concept force. No one has

ever seen a force. Only the movement of objects is seen. The faith of science is that some schematic representation is possible by which complexities of movement can be conceptually unified into order." (p.51)

History of psychometrics. Some scholars criticized the causal interpretation of factor analysis using a historical approach. For example, Abbott (1998) argued that early psychometricians viewed factor analysis as a mathematical convenience to reduce complex data to simple forms in order to reconcile quantitative data with intuitive categories, and thus it ignored causality altogether. This view seems to be concurred by Laudan (1977). Laudan classified psychometrics in the early 20th century as a "non-standard research tradition" because it does not have a strong ontology and metaphysics. Rather its assumption is "little more than the conviction that mental phenomena could be mathematically represented." (p.105)

There are several loopholes in this argument. Abbott argued that besides early psychometrics, biometrics, econometrics, and sociology are also a-causal. However, in the discussion of early psychometrics, Abbott did not use even one single citation to support his claim. It is true that Thurstone (1945) sometimes refer to theoretical entities as "convenient postulates," but he did not deny the possibility of some degree of correspondence between constructs and reality. While discussing the origin and development of factor analysis, Vincent (1953) asserted that factor analysis is an attempt to identify the causes that are operating to produce the variance and to evaluate the contribution due to each cause. In his view, the argument among early psychometricians was concerned with whether one common cause or multiple causes were appropriate.

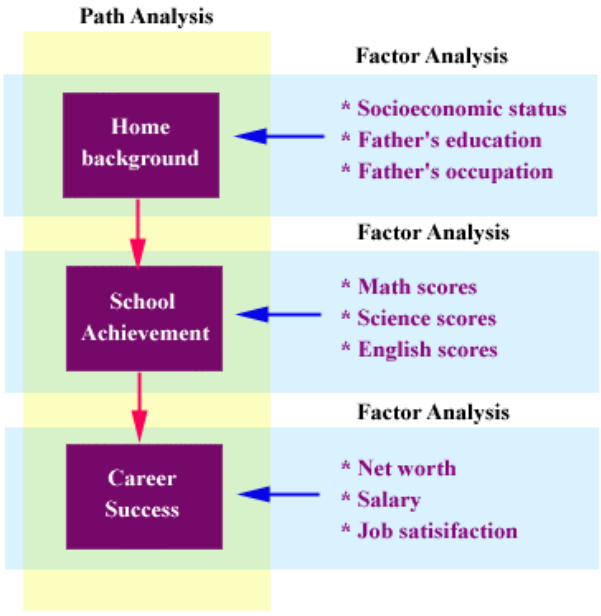
Abbott cited Yule's notion that correlation does not indicate a cause and effect to support his argument that early econometricians were associationsts rather than causationists. However, even if factor analysis and other statistical methods were a-causal at their early stage of development, it doesn't necessary imply that this idea should not be altered and factor analysis today should continue to be interpreted in a non-causal fashion. Indeed, modern scholars view factor analysis as an application of the principle of common cause (e.g. Glymour, 1982; Glymour, Scheines, Spirtes, &

Kelly, 1987). If someone denies their work just because their ideas depart from the founding fathers, it is like rejecting the design of a V-8 engine just because it violates the idea of Henry Ford's Model T.

Path model

Another component of SEM is the path model, which is also called the structural model. After latent constructs are identified, the relationships among these constructs are arranged to form "chains" or "paths." The example illustrated in Figure 3 is given by Lomax (1992). Based upon literature review, a researcher hypothesizes that "home background" could be a predictor to "school achievement," and "school achievement" could predict "career success", he defines such vague concepts as home background, school achievement, and career success by the factor model. Afterwards, a chain (path) of cause and effect is drawn among constructs. Then he/she employs SEM techniques to examine the fitness between the data and the model. Please keep in mind that this example is simplified. A real-life structural equation model could be more complicated. Because of the complexity of SEM, there are numerous possible ways to fit the data with the model. The fitness indices become the evidence of the causal inference in SEM. This point will be further explained in a later section.

Figure 3. Example of a simple SEM



A structural model is a linear model. Critics are skeptical whether a linear model could represent and causally explain complex phenomena in the empirical world (e.g. Ling, 1982; Freedman, 1987, 1997). Ling called the causal inference by path models "a form of statistical fantasy" (p.490), and Freedman called it "a faulty research paradigm." (p.102) As a matter of fact, many times relationships in the real world do not fall into a linear pattern. In many datasets, the residuals between the linear fit and the data points are manifested in scatterplots. It seems natural that the line should go through all data points in a non-linear fashion. It is understandable why people believe that the linear model is an over-simplification of the world.

Glymour, Scheines, Spirtes, and Kelly (1987) defended the sufficiency of linearity by using the fitness argument. According to Glymour et al, sciences have always proceeded by approximation and idealization. Linear approximation is not literally true, of course. Nevertheless, the principal justification for a linear model is that it explains the correlation data very well and no alternative linear model is readily available which provides a comparably good explanation of the correlations. In addition, linear models are conceptually simple, computationally tractable, and often empirically adequate.

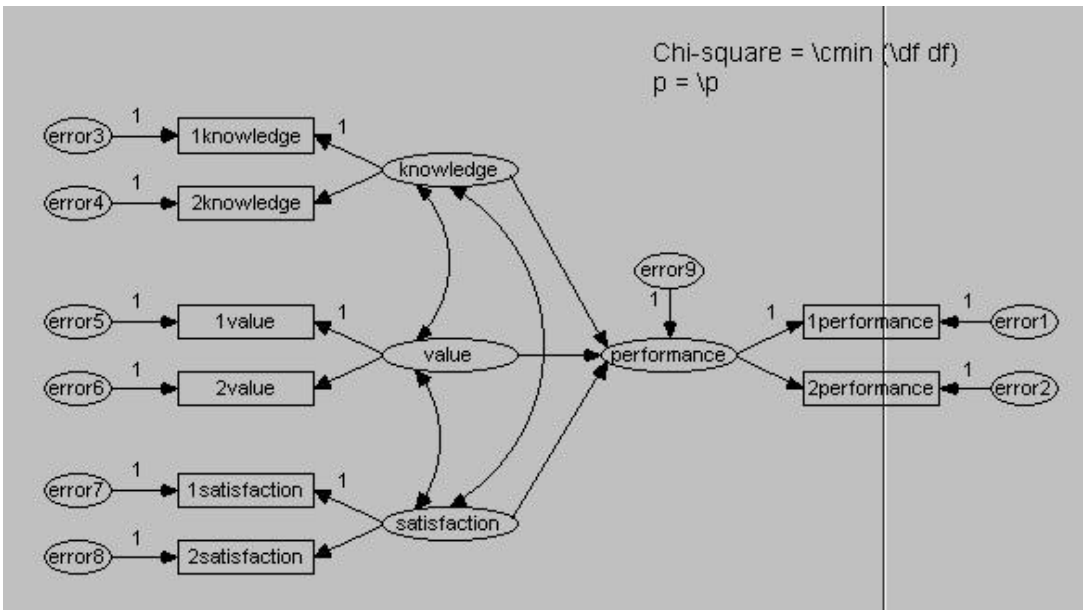
It is important to point out that all three criteria must be presented together. Computational tractability is a manifestation of the testability and repeatability. When data could be computed and the procedure can be replicated by the same algorithm, the model is said to be testable. Testability is a pre-requisite of empirical adequacy (fitness). If a model cannot be verified or falsified, no one could tell whether the data fits the model or not.

Simplicity alone is not a good criterion of judging the validity of the model. First, simplicity does not warrant whether the model is true. Simplicity is relevant to the pragmatic issue of research methodology, but is irrelevant to the epistemic aspect (van Fraassen, 1980). It is a common practice that when researchers face two equally adequate models in terms of explanatory power, they tend to choose the simpler model. However, perhaps the complicated one is closer to the truth. Thus, this theory choice is pragmatic rather than epistemic.

Second, simplicity is a relative concept. On some occasions the balance between fit and parsimony can be objectified by mathematics. For example, when comparing regression models with different sets of predictors, model comparison and variable selection procedures can be employed to determine whether the increase of R^2 can justify the increased complexity of the model. However, linear models in SEM are not necessarily simpler than non-linear models in regression analysis, and there is no objective way to tell whether one is simpler than another. Figure 4 shows a typical SEM. Even though the entire model is composed of linear models, through intuition it is by no means simple.

Further, even within the same research methodology, simplicity is still a relative concept. For example, in the regression context, how many variables should be retained to formulate a simple model is tied to the fitness criterion. In other words, researchers attempt to achieve the balance of fit and parsimony. The issue of simplicity and fitness will be further discussed in the section concerning identification. In short, simplicity alone might be open to attack, but combining simplicity, tractability, and fitness provide a strong justification of using linear models.

Figure 4. An example of SEM

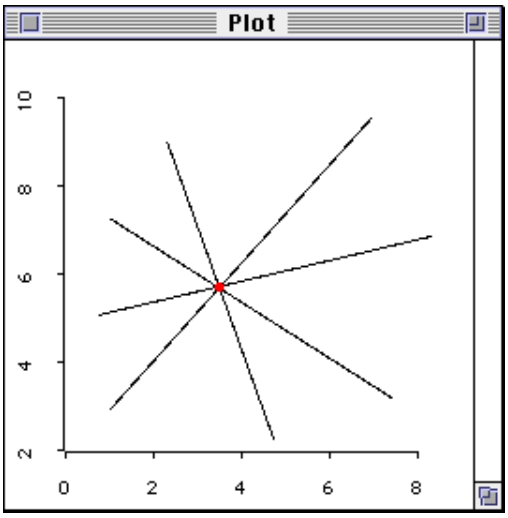


Causal explanation power of SEM

As mentioned before, testability is one of the criteria of establishing a valid model. Identification is a special case of testability in SEM. Pearl (2000) asserted that in a structural equation such as Y = BX+E, the causal connection between X and Y must have no other value except B. He used a circuit board as a metaphor to SEM. In a circuit board, in which different components are joined by different paths, it shows not only how the circuit behaves under normal conditions, but also shows how the circuit would mis-behave under millions of abnormal conditions. While there are many ways for a signal to go through the circuit, only one correct way allows the signal to reach the destination so that the electronic device could perform the proper function. By the same token, a structural model formed by a web of complex relationships can have a million ways of model mis-specification. Assume that there is only one way that the model can be properly specified. If one unique solution is found out of many possible combinations, then a cause and effect relationship can be claimed. The uniqueness of the solution is tied to the issue of identification.

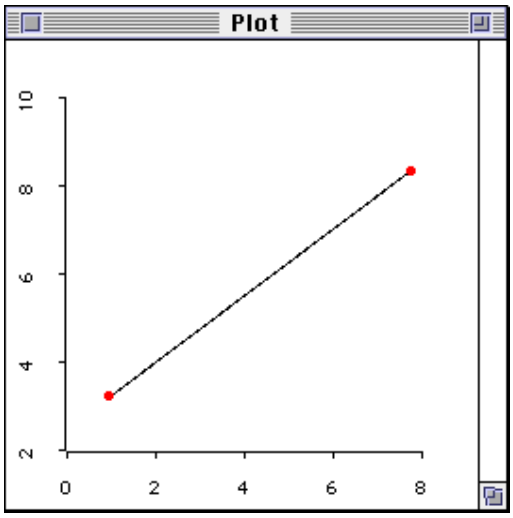
When there are more unknown parameters than the number of equations, this situation is called under-identification. For example, given the equation X+Y=2, this equation may yield infinite sets of solutions i.e. (X=1, Y=1), (X=3, Y=-1), (X=2, Y=0)…etc. For example, in Figure 5, the equation can be written as Y = a + bX. However, a line could fit the datum in any direction. In other words, X and Y could take any value. The influence of the Popperian principle of falsifability can be found in this case. When the resulting equations fail to specify a unique solution, the model is said to be unfalsifiable, because it is capable of perfectly fitting any data. To be specific, if a model is "always right" and there is no way to disprove it, this model is useless. Thus, in SEM testability could be viewed as falsifability.

Figure 5. Insufficient data for falsification



If there are two equations such as X+Y=2 and 2X+Y=4, then the problem is more solvable. But the condition is still less than desirable. When there are two data points in the graph (see Figure 6), the statistician could draw a perfectly fitted line to connect two data points. Anyone could obtain two data points in any study and always come up with a "perfect solution." Thus, this model is also not falsifiable.

Figure 6. Perfectly fit data that cannot be falsified



When there is just enough information to get a value for every parameter, the model is said to be just-identified. However, when there are more equations than unknown parameters, the model is considered over-identified. With over-identification, there is also no exact solution. This condition at first seems unfortunate, but it is actually a blessing in disguise. Although we may not obtain an exact solution, we may define a criterion and obtain the most adequate solution (Chou & Bentler, 1995). Following this thread, Pearl endorsed the use of over-identified models for data/model fit.

Petrovic (2000) further explained the importance of identification in causal interpretation by relating simplicity to fitness. If SEM would not have causal meaning, we should be able to compare observationally equivalent models solely on the basis of their parsimony rather than considering their respective choice of parameters. Why does the researcher burden himself/herself with an over-identified model when a simpler just-identified model could provide a satisfactory answer and save some CPU cycles? It is because an over-identified model could provide a better fit in terms of uniqueness.

At a quick glance, the identification approach looks similar to the one used by experimenters as mentioned in the beginning. The goal of careful experimental design and hypothesis testing is to rule out rival hypotheses. If SEM that seeks for data/model fit is nothing more than finding the best explanation out of rival models, then why is it considered an improvement over conventional methods? Nevertheless, there are major differences between the experimental school and the SEM school. In the experimental school, non-experimental and observational studies are not qualified to generate strong causal claims. However, Glymour et al asserted that using SEM causal inferences are still possible with non-experimental data. In attempt to support this claim, Glymour et al developed a program named TETRAD as a supplement for SEM software programs such as LISREL and EQS. Since TETRAD is co-developed by Spirtes, Glymour, and Scheines, this module is also known as SGS. Given the input as the covariance structure (joint distributions of variables), TETRAD is capable of generating paths among factors/variables. It is important to note that TETRAD does not output a unique path model and affirm the causal relationships. Instead, the output from TETRAD is a family

of path models, which could be compatible with the covariance matrix. The automated path generation is an aid to, but not a replacement of, subsequent testing by human researchers.

Second, in hypothesis testing such as a t-test or ANOVA, the researcher looks for the significance of a few effects. However, a structural model is holistic. The fitness judgment is based upon not just one omnibus test statistics. Besides the global fit information, SEM provides local fit information for the researcher to debug the model when misfit occurs.

Graphical model

Local fitness testing

While Pearl (2000) rigorously defended SEM, he also developed the graphical model as an enhancement to SEM. According to Pearl, Data/model fitness and over-identification are necessary, but not sufficient conditions to infer cause and effect relationships. Pearl pointed out two potential problems of fitness tests:

1. If some parameters are not identifiable, the first phase of the test may fail to reach stable estimates for the parameters and the investigator must simply abandon the test.
2. If the model does not fit the data adequately, the researcher receives little information about which modeling assumptions are wrong.

In this case, the global fitness test is not helpful. As Duheim pointed out, when a complex set of variables and hypotheses goes wrong, a global answer at most could tell the researcher that something is wrong somewhere. To rectify this situation, Pearl suggested that local fitness testing is a better alternative. Local fitness testing examines the restrictions implied by the model one by one. It is considered more reliable than the global testing because it involves fewer degrees of freedom and is not affected by irrelevant measurement error. As discussed before, every measurement model carries certain errors. By using local fitness test the errors are localized.

Pearl's approach is directly opposed to Laudan's idea. In answering the Duheim challenge mentioned earlier, Laudan (1997) argued that when a complex set of theories is tested and generates an anomaly, the researcher should not try to localize blame or distribute credit to specific portions of

the model. Rather, the rational strategy is to select a better set because the anomaly affects each element within the complex. Pearl's approach is considerably superior to Laudan's as a solution to the Duheim's problem. Laudan's idea resembles the global test thinking, which could hinder scientific research from accumulation of learning and progress. In contrast, Pearl's idea of local fitness testing can provide information for further model refinement. Unless the entire model is seriously mis-specified and global fitness testing correctly rejects the entire model, local fitness testing is no doubt a better choice.

Covariance equivalent models

When relationships are expressed in terms of functions and equations, the association could not be interpreted as causation without further justification. For example, $Y = A + BX$ can be rewritten as $X = (Y - A) / B$. Thus, X could not be viewed as a cause of Y because the positions of X and Y could be swapped around the equation even if B is the only value that could solve the equation. In other words, the relationship between X and Y is not directional.

Using a circuit analogy, Pearl argued that a circuit diagram captures the very essence of causation because a circuit diagram could predict outcomes but equations cannot. In a circuit the layout of paths is directional instead of functional. By drawing causal diagrams in a graphical model, one could go beyond testing equations to testing possible directions of equivalent models. Two models are considered equivalent if their reproduced covariance matrices are identical, regardless of the direction of the arrows. For example, X ✍ Y is equivalent to X ✍ Y if the covariance structure between X and Y in the first model is the same as that of the second one. When there are many variables, different combinations of paths form numerous equivalent alternative models. Pearl relates the significance of equivalent models to the falsifiability criterion. In this way, the researcher does not test a single model but a whole class of observationally equivalent models. This class of equivalent models can be constructed and inspected graphically. If one unique solution comes up from many alternative models, a firm causal inference can be made.

The logic of Glymouir's TETRAD is very similar to Pearl's approach. TETRAD also output a

pattern (class) of possible path models which are covariance equivalent. However, it may be the case that there are SEMs which are not covariance equivalent, but nonetheless fit the data almost equally well. This problem is addressed by outputting multiple patterns of possible SEMs (Scheines et al, in press).

Testing all/many possible models can be further explained by using the metaphors of best subset regression and the counterfactual model. In regression when the researcher has many predictors, variable selection procedures such as maximum R^2 can be used to try out all possible combinations of variables in order to obtain the best model. Testing covariance equivalent models in SEM is analogous to variable selection in regression. Both are considered research approaches for achieving the balance between simplicity and fitness.

Testing covariance equivalent and other possible fit models could also be viewed as an expansion of the counterfactual model. Counterfactual questions, as the name implies, are "what-if" questions. When X occurs and Y follows, the researcher could not jump to the conclusion that X causes Y. The relationship between X and Y could be "because of," "in spite of," or ' regardless of." A responsible researcher would ask, "What would have happened to Y if X were not present?" In other words, the researcher does not base his/her judgment solely on the existing outcome, but also other potential outcomes. Thus, this model is also known as the potential outcome model.

Controlled experiments often have a counterfactual aspect. To be specific, the control group gives the information about how Y behaves when X is absent while the treatment group tells the experimenter about how Y reacts when X is present. However, the counterfactual approach taken by experiments is limited in two senses. First, causal inferences can not be made to non-experimental data. Second, the experimenter can manipulate just a few scenarios.

Similarly, testing covariance equivalent and other fit models is also asking "what-if" questions. However, the researcher who employs the graphical model exhausts all possible scenarios by manipulating the model graphically (reversing arrows). For example, he/she may consider, "what would happens if we assume that Y causes X and Z causes Y, instead of assuming X causes Y, and Y

causes Z?" Moreover, manipulating possible models enables the researcher to draw causal interpretations from non-experimental data. Some researchers mis-perceived SEM as a competitor to randomized experiments (Pearl, 1995). Indeed, besides exploring more "what-if" scenarios, SEM extends the counterfactual model, in that the actual outcome from a given function may serve as an input to subsequent potential outcome functions (Greenland, 2000).

Criticisms

Probabilistic vs. deterministic causality

Although the aforementioned causal models are philosophically sound and mathematically sophisticated, objections against these models are still visible in the academic arena. For example, while discussing causality in sociology, Abbott (1998) deliberately removed the mathematical models from his writing since he doubted the validity of probabilistic causation. In the same vein of prominent sociologist Durkheim, Abbott contended that causality means determination, which is necessary and sufficient.

Probabilistic causation and deterministic causation is an old topic in philosophy. Determinists asserted that scientific laws could only be founded on certainty and on an absolute determinism, not on a probability (Hacking, 1992). In scientific determinism, every outcome is a necessity. i.e., given the cause, the effect must occur. This idea originated from French mathematician Laplace. Based on the Newtonian physics, Laplace claimed that everything is determined by physical laws. If a powerful intellect (called Laplace's demon) fully comprehends the Newtonian law, and knows the position and momentum of every particle in the universe, no doubt he could predict every event in the history of the universe. Laplace's determinism was applied to the realm of extended, spatial, material substance. Later determinism was expanded to the realm of psychological and sociological events.

Philosophers have been puzzled by how one could implement causal relations in a non-deterministic context. Mulaik and James (1995), who are vocal endorsers of SEM, use the item response theory (IRT) to argue for the probabilistic causal model. In IRT, item difficulty and subject ability (theta) jointly determine a specific probability distribution on the response variable. Varying

ability and varying item difficulty varies the probability distribution of outcomes on the response variable. In IRT, estimation of subject theta is aided by the Bayesian approach, which updates the probability based upon new information (Mislevy, 1993). The probability that the examinee could answer a question correctly is contingent upon his/her ability. And his estimated ability is contingent upon his/her ongoing performance, especially in an adaptive test. In this scenario, it is more appropriate to interpret causation in a probabilistic fashion. By the same token, the probabilistic property of SEM should not be viewed as a sign of invalidity.

Pearl (2000) is well-aware of the issue of probabilistic causality. In the graphical model, Bayesian Networks (BN) are employed to encode causal relations. In contrast to the determinist view of Laplace, causal relationships defined in BN are assumed to be probabilistic. Pearl (in press) argues that conventional statistics has difficulties in expressing causal concepts because statistics deals with static conditions. However, causal analysis involves a web of interacting variables and changing conditions, and thus BN is more applicable to causal analysis. In the graphical framework, BN performs three roles:

1. to represent the causal assumptions about the environment;
2. to facilitate economical representation of joint probability functions;
3. to facilitate efficient inferences from observations.

Pearl argued that owing to the wide acceptance of quantum mechanics, natural laws are said to be probabilistic and determinism is just a convenient approximation. In this view, BN appeals to the modern concept of physics. Quantum mechanics may be very remote to ordinate people. Nonetheless, Salmon (1984) pointed out that probabilistic causality, rather than deterministic causality, is more aligned to our common sense in everyday life. For example, heavy smokers do not necessarily get lung cancer. It is only probable that a heavy smoker could become a cancer patient. In advocating probabilistic causality, Salmon did not deny the existence of sufficient causes. However, sufficient causes constitute a limiting case of probabilistic cases, which seems to be restrictive.

Untested assumptions

The most vocal critic against SEM is Freedman. Freedman (1997) denounced Glymour's TETRAD program and criticized that "causation has to do with empirical reality, not with mathematical proofs based on axioms. The issue is not one of theorems, but of the connection between theorems and reality." (p.76) In another paper that also refuted TETRAD, Freedman and Humphreys (1998) repeated the same notion, "There is no coherent ground—just based on the mathematics—for thinking that the graphs represent causation … The mathematics in SGS will not be of much interest to philosophers seeking to clarify the meaning of causality." (p.3)

Interestingly enough, Freedman's criticisms against Glymour and Pearl can also be framed in the Duhem question: "If assumptions A, B, C … hold, then H can be tested against the data. However, if A, B, C … remain in doubt, so must inferences about H." (p.102). When facing an expected outcome, Duhem might say the theory remains inconclusive. In contrast, Freedman simply rejected the whole theory altogether. In Freedman's eyes, untested assumptions are just "maintained hypotheses." Freedman argued that the causal model suggested by Glymour carried many untested assumptions and the only empirical data are the covariance structure. Freedman gave the same challenge to Pearl: "To make real progress, those assumptions have to be tested." (p.693)

With regard to the validity of the assumptions for the path model, Freedman (1987) pointed out three possible threats:

1. Measurement error in the exogenous (independent) variables.
2. Nonlinear relationship between the endogenous (dependent) and exogenous variables.
3. Omitted variables.

Social scientists use latent factor models to address the first problem. However, Freedman said that this solution involves another set of assumptions. For example, it is assumed that there are repeated measurements linearly related to the latent factors. Pertaining to the second problem, Freedman said that when the variables are related in a non-linear fashion, the estimated coefficient would be biased. About the last problem, Freedman asserted that missing important variables could

lead to a mis-specified model. He mocked that "this problem too is well known to workers in the field, and their solution is to expand the system by adding more variables...Current social science theory cannot deliver that sort of specification with any degree of reliability, and current statistical theory needs this information to get started." (p.109). Besides the preceding assumptions, Freedman (1997) questioned other assumptions embedded in SEM and TETRAD such as faithfulness and causal Markov conditions, but discussion of those assumptions is beyond the scope of this paper.

Freedman (1997) dismissed all popular reasons of accepting SEM assumptions:

> In the social sciences, however, statistical assumptions are rarely made explicit, let alone validated. Questions provoke reactions that cover the gamut from indignation to obscurantism. *We know all that. Nothing is perfect. Linearity has to be a good first approximation. The assumptions are reasonable. The assumptions do not matter. The assumptions are conservative. You cannot prove the assumptions are wrong. The bias will cancel. We can model the bias. We are only doing what everybody else does. Now we use more sophisticated techniques. What would you do? The decision-maker has to be better off with us than without us. We all have mental models; not using a model is still a model.* (p.103) (Italic appears in the original text.)

Use of latent factors and linear models have been discussed in previous sections. Freedman identified measurement error as one of the problems of the factor model. If error-free measurement is required in research, I am afraid that most research studies would be "mission impossible." This section will concentrate on the problem of missing variables. It is true that certain variables that are crucial causes to the outcome may be overlooked by the researcher. However, It is totally acceptable to miss some variables and then expand the system by adding more later. It is curious that Freedman denied adding more variables as a viable solution because model specification with a high degree of certainty is difficult. Demanding the researcher to identify all relevant causal variables with certainty is like expecting the researcher to be a Laplace demon, who has the full knowledge of the whole world. We conduct research exactly because we don't know the cause and effect, not because we

know everything. Our knowledge of the world is incomplete and it is perfectly fine to admit that any model or theory is fallible.

Nevertheless, even if the researcher possesses the intelligence of the Laplace demon, is it necessary for him/her to include all relevant variables into the model? Like linearity, simplicity is also a reason that the researcher may omit certain "important" variables. All models are mis-specified in the sense that some variables are always excluded from the model. For example, a student asked me what variables cause school performance. I told him/her about my fifty-variable model: Study long hours, earn more money, marry a good wife, buy a reliable car, watch less TV, browse more often on the Web, exercise more often, attend church more often, pray more often, go to fewer movies, play fewer video games, cut your hair more often, drink more milk and coffee...etc. Needless to say, this over-specified model is not useful at all. It is understood that Freedman was concerned with "important variables," not trivial variables. In this example, Freedman might worry that the most important variable "study long hours" could be left out while others such as "drink more milk and coffee" are retained. Nonetheless, methodologically speaking it is not a bad thing to leave out important variables, because the model will be simple enough to falsify.

One question implied by the Duhem thesis is: Could theories be refuted? Quine (1976) argued that by adding or adjusting ad hoc hypotheses, any disputed theory could be accepted. If the model fails to fit the data, the researcher may say, "perhaps some important variables are missing." In this manner, the same theory could be tested over and over by adding more variables endlessly. On the other hand, Freedman used the same argument to refuse the validity of causal models: A model is invalid if some important variables are missing. Neither Quine nor Freedman could answer the Duhem question adequately. No model could fit the data perfectly. Again, this type of question could also be endless no matter how fit or unfit the model is. Therefore, advocates of causal models have explicitly spelled out the criterion of identification to set the parameters of testability.

In response to Freedman, Bentler (1987), one of the developers of EQS (a software program for SEM), defended the value of SEM in terms of simplicity and fitness test. Bentler stated that the

central question of SEM is whether ? =? (?), where ? is a vector of population parameters and ? is a vector of smaller dimension than ? . For example, say there are 1000+ elements in ? and the researcher can find 100 parameters to characterize ? , the researcher will have obtained a tremendous simplification in representing the data. Further, goodness of fit of any model might be judged by the size of residuals (the difference between the predicted and the actual) or by fit indices. Bentler argued that to consider SEM as worthless, Freedman rejected a valuable idea and his action was discarding "the baby with the bathwater."

At first glance Freedman's strong demand for empirical support and skepticism of untested assumptions are reasonable. However, science does not progress based upon empiricism and a high degree of certainty. When Copernicus and Galileo developed the heliocentric model, there were not sufficient empirical data to support their claim. Before the introduction of the high-powered microscope, subatomic entities were considered untested assumptions. In the modern era, many sciences also proceed with untested assumptions and theoretical constructs. For example, the mental entities and processes proposed by cognitive psychologists are derived from a web of tested and untested assumptions. Asking for empirical substantiation and denouncing untested assumptions would inevitably reverse psychology to behaviorism. Inferences, by nature, are actions that take the researcher from one point to another. A typical example is that we usually draw an inference from a sample to a population, in which the size is infinite and the distribution is unknown. This is the type of uncertainty that researchers must live with unless the research goal is simply description, in which no inference is made.

In reply to Freedman's challenge, Spirtes and Scheines (1997) admitted that the TETRAD method is incomplete and there may be many other kinds of assumptions that should be investigated. Nevertheless, this is a systematic examination using partial knowledge. Scheines et al (in press) further articulated the benefits of using assumptions in the context of causal model building. The bolder assumptions the researcher makes, the more knowledge he/she can learn about the causal structure:

> The result … do not free one from having to make assumptions; instead, they make rigorous and explicit what can and cannot be learned about the world if one is willing to assume that causal relations are approximately linear and additive, that there is no feedback, that error terms are i.i.d and uncorrelated, and that the Causal Independence and Faithfulness assumptions are satisfied, then quite a lot can be learned about the causal structure underlying the data. If one is only willing to make weaker assumptions, then less can be learned. (p.2)

Last but not least, Freedman devoted tremendous effort to argue against causal models, but didn't spend a page to argue for a-causal models or suggest any better alternative. In the conclusion of Freedman's paper (1987), he said, "This kind of negative article may seem incomplete. Path analysts will ask, not unreasonably, 'Well, what would you do?' To this question, I have no general answer." (p.125) Even though he did not give an answer, other researchers would take no answer as an answer. Based on Freedman's a-causal attitude, Abbott (1998) asserted that sociology should depart from causal accounting and spend more effort on descriptive work. In educational research, qualitative researchers independently adopt the a-causal notion and promote descriptive/narrative research. Studying phenomena without knowing why is just like operating a "black-box." This attitude sets the clock backward. Take computing as a metaphor. It is not good enough for a computer programmer to correctly describe the signs of a system crash. A competent programmer should know what causes the system crash and is able to diagnose the problem.

Conclusion

The Duhem question is central to this discussion: When multiple variables, hypotheses, auxiliary assumptions exist, how could a researcher reach a conclusion or infer a causal and effect interpretation? To answer the Duhem question, the mathematical approach, such as Structural equation models, TETRAD, and graphical models, attempts to exhaust almost all possible combinations of paths. Other components of these causal models are also rigorously defended. Use of latent factors is justified by the argument of realism and the threat of measurement errors are addressed by Cronbach Alpha and

triangulation in factor analysis. Linearity of path models is justified in the context of achieving simplicity and fitness. By employing Bayesian Networks, probabilistic causality is considered legitimate and even better than deterministic causality. Although there are certain untested assumptions in these causal models, theories are fallible and scientific inquiry essentially carries some degree of uncertainty, yet SEM is a good tool to make causal inferences based upon incomplete knowledge.

Acknowledgement

Special thanks to Dr. Brad Armendt and Mr. Shawn Stockford for reviewing this paper

References

Abbott, A. (1998). The causal devolution. Sociological Methods & Research, 27, 148-180.

Baron, J. (2000). Thinking and deciding (3rd ed.). Cambridge: Cambridge University Press.

Bentler, P. (1987). Structural modeling and the scientific method: Comments on Freedman's critique. Journal of Educational Statistics, 12, 151-157.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (in press). Philosophy of science and psychometrics: Reflections on the theoretical status of the latent variable.

Campbell, D. & Stanley, J. (1963). Experimental and quasi-experimental designs for research. Chicago, IL: Rand-McNally.

Chou, C. H., & Bentler, P. M. (1995). Estimates tests in structural equation modeling. In R. H. Hoyle (Eds.), Structural equation modeling: Concepts, issues, and applications (pp. 37-55). Thousand Oaks: Sage Publications.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston, MA: Houghton Mifflin Company.

Duhem, P. M. M. (1954). The aim and structure of physical theory. Princeton: Princeton University Press.

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

Freedman, D. (1995). Discussion of causal diagrams for empirical research by J. Pearl. Biometrika, 82, 692-693.

Freedman, D. A. (1997). From Association to Causation via Regression. Advances in Applied Mathematics, 18, 59-110.

Freedman, D. A. & Humphreys, P. (1998) Are there algorithms that discover causal structure? Technical Report 514. Berkeley, CA: University of California, Berkeley.

Glymour, C. (1982). Casual inference and causal explanation. In R. McLaughlin (Ed), What? Where? When? Why? Essays on induction, space, and time, explanation (pp. 179-191). Boston, MA: D.

Reidel Publishing Company.

Glymour, C. (1986). Comments: Statistics and metaphysics. Journal of the American Statistical Association, 81, 964-966.

Glymour, C., Scheines, R., Spirtes, P, & Kelly, K. (1987). Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling. Orlando, FL: Academic Press, Inc.

Glymour, C. & Cooper, G. F. (eds.). (1999).Computation, causation, and discovery. Cambridge, Mass.: MIT Press.

Greenland, S. (2000). Causal analysis in the health sciences. Journal of the American Statistical Association, 95, 286-289.

Hacking, I. (1992). The taming of chance. Cambridge, UK: Cambridge University Press.

Hoyle, R. H.. (1995). The structural equation modeling approach: Basic concepts and fundamental issues. In R. H. Hoyle (Eds.), Structural equation modeling: Concepts, issues, and applications (pp. 1-15). Thousand Oaks: Sage Publications.

Kelley, T. L. (1940). Comment on Wilson and Worcester's Note on Factor Analysis. Psychometrika, 5, 117-120.

Kerlinger, F. N. (1986). Foundations of behavioral research (3rd ed.). Forth Worth, TX: Holt, Rinehart and Winston.

Laudan, L. (1977). Progress and its problems: Toward a theory of scientific growth. Berkeley, CA : University of California Press.

Ling, R. (1982). Review of "Correlation and causation" by David Kenny. Journal of American Statistical Association, 77, 481-491.

Lomax, R. G. (1992). Statistical concepts: A second course for education and the behavioral sciences. White Plains, NY: Longman.

Mislevy, R. (1993). Some formulas for use with Bayesian ability estimates. Educational & Psychological Measurement, 53, 315-329.

Mulaik, S. A., & James, L. R. (1995). Objectivity and reasoning in science and structural equation

modeling. In R. H. Hoyle (Eds.), Structural equation modeling: Concepts, issues, and applications (pp. 118-127). Thousand Oaks: Sage Publications.

Pearl, J. (1995). Rejoinder to discussions of causal diagrams for empirical research. Biometrika, 82, 702-710.

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

Pearl, J. (in press). Causal inference in the health science: A conceptual introduction.

Petrovic, M. (2000). Probabilistic and structural causality. [On-line] Available URL: http://www.soc .washington.edu/courses/soc582/misha3.html

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

Spirtes, P., and Scheines, R. (1997). Reply to Freedman. In S. Turner and V. McKim (Eds.), Causality in Crisis: Statistical Methods and the Search for Causal Knowledge in the Social Sciences. (pp.163-176). University of Notre Dame Press.

Scheines, R., Spirtes, P., Glymour, C., Meek, C., and Richardson, T. (in press). The TETRAD Project: Constraint Based Aids to Causal Model Specification, Multivariate Behavioral Research.

Quine, W. V. O., (1976). Two Dogmas of Empiricism" In S. G. Harding (Ed.), Can theories be refuted?: essays on the Duhem-Quine thesis (pp. 41-64). Boston : D. Reidel Pub. Co.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. Educational and Psychological Measurement, 60, 174-195.

Thurstone, L. L. (1947). Multiple-factor analysis: a development and expansion of the vectors of mind. Chicago: The University of Chicago press.

Vacha-Hasse, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. Educational and Psychological Measurement, 58, 6-20.

Van Fraassen, B. C. (1980). The scientific image. Oxford: Clarendon Press.

Vincent, D. F. (1953). The origin and development of factor analysis. Applied statistics, 2, 107-117.

Yu, C. H. (2001). An Introduction to computing and interpreting Cronbach Coefficient Alpha in SAS.

Proceedings of 26th SAS User Group International Conference. [On-line] Available: URL:

http://seamonkey.ed.asu.edu/~alex/pub/cronbach.html