STRUCTURAL EQUATION MODELING, *1*(3), 253–267

Causation Issues in Structural Equation Modeling Research

Heather E. Bullock and Lisa L. Harlow
University of Rhode Island

Stanley A. Mulaik
Georgia Institute of Technology

As the use of structural equation modeling (SEM) has increased, confusion has grown concerning the correct use of and the conclusions that can be legitimately drawn from these methodologies. It appears that much of the controversy surrounding SEM is related to the degree of certainty with which causal statements can be drawn from these procedures. SEM is discussed in relation to the conditions necessary for providing causal evidence. Both the weaknesses and the strengths of SEM are examined. Although structural modeling cannot ensure that necessary causal conditions have been met, it is argued that SEM methods

searchers have been criticized for suspending this principle when employing SEM (e.g., Amdur, 1989; Games, 1988, 1990). Regardless of the increased sophistication and precision offered by SEM procedures, if the data are correlational in nature, no statistical method can change the design. Ultimately, it is the design, not the statistical method, that permits causal hypotheses to be adequately tested (e.g., Campbell & Stanley, 1963).

Although it is beyond the scope of this article to elaborate on all the conditions necessary to establish causality, three basic requirements have been cited: association between two variables (i.e., two variables must be correlated), isolation of the effect (i.e., ruling out extraneous variables), and temporal ordering, where a cause is shown to unambiguously precede an effect (for more discussion of these issues, see Bollen, 1989; Mulaik, 1986, 1987a, 1987b).

The purpose of this article is to examine the limitations as well as the strengths of SEM in relation to causality and to provide guidelines that may help researchers operationalize some of the pertinent issues. Our discussion is organized into the following sections: weaknesses and criticisms of SEM, strengths of SEM, and practical guidelines for concretizing some of the abstract prescriptions surrounding causality and SEM.

SEM: A CRITICAL LOOK

SEM methods have been applauded and attacked. Criticism ranges from relatively superficial to more substantive concerns about flawed logic (e.g., Baumrind, 1983; Cliff, 1983, 1987; de Leeuw, 1985; Freedman, 1987a, 1987b; Ling, 1982; Ragosa, 1987; Steiger, 1980). One must ask whether all

IGNORING BACKGROUND CONDITIONS

Background conditions such as mediating mechanisms, stability, and the form of the functional relation are often not adequately addressed in SEM applications (e.g., Mulaik, 1986, 1987b). Although it can never be determined whether background conditions have been fully met, researchers should make every attempt to ensure that such conditions have been considered.

A number of studies have examined various mediators to help explain a phenomenon. In the area of substance use, cigarette use and marijuana use have been cited as possible mediators between alcohol and harder substance use, offering support for a "gateway" hypothesis of substance use (Kandel, 1975; Welte & Barnes, 1985). In a different area, Guida, Ludlow, and Wilson (1985) found that "time on task" mediates anxiety and academic achievement, such that more time on task lessens the negative effects of anxiety on achievement. In still a different area, Harlow, Newcomb, and Bentler (1986) found that substance use and suicide ideation were mediated by a lack of purpose in life.

Stability refers to the need to measure the effect at the "correct" time interval. Making good estimates of the time interval can be difficult and often takes several studies to determine. This issue has been examined in a number of studies. For example, in a three-wave study of mediators of coronary heart disease, Fontana, Kerns, Rosenberg, and Colonese (1989) found that support was more stable 6 months later, whereas stress levels were more influential and stable after 12 months. In a different area, Francis, Fletcher, Maxwell, and Satz (1989) found that reading achievement was unstable across three waves of data from kindergarten to fifth grade, whereas cognitive skills remained fairly stable over time.

conditions have been met. Although Freedman's advice is sound, it is virtually impossible for researchers to meet such requirements, even with experimental designs. Pragmatically speaking, researchers could approach this task by carefully considering the most likely defeaters of a causal hypothesis and then attempting to control these conditions. This is not too different from the imperatives advocated by Campbell and his colleagues (Campbell & Stanley, 1963; Cook & Campbell, 1979) to attend to threats to internal and external validity.

CROSS-SECTIONAL SEM STUDIES

To provide causal evidence, a minimal requirement is that a variable must be shown to temporally precede its effect. Temporal ordering may be difficult to establish in itself, particularly when working with structural equation models, because the latent variables cannot be observed. Unfortunately, cross-sectional applications predominate the literature, thus prohibiting evidence for directionality from being obtained. Longitudinal SEM procedures can offer some possibility of providing initial evidence for the direction of causation in a specific area of study. This preliminary evidence could be further verified with additional studies, preferably including experimental designs.

Anashensel and Huba (1983) presented a clear and understandable application of SEM with a longitudinal design examining depression, alcohol, and cigarette use across four waves. SEM methods for examining a mixture of cross-sectional and longitudinal data have been developed and may prove useful when data are not complete across all time spans of a study (McArdle, Hamagami, Elias, & Robbins, 1991). Farkas and Tetrick (1989) showed how instance, correlational data can only provide cues to causal relations (e.g., White, 1990). The use of correlational data may be particularly treacherous when the researcher cannot actively manipulate variables to simplify relations, even temporarily (Cliff, 1983). Because SEM is frequently used to analyze nonexperimental data for which manipulation is impossible, this criticism is a valid one. However, SEM procedures can certainly be used with experimental data. Bagozzi and Yi (1989) suggested that SEM procedures be applied to experimental data that have traditionally been analyzed using multivariate analysis of variance (MANOVA) and multivariate analysis of covariance. It is important to reiterate that no statistical routine (e.g., MANOVA, SEM)—by itself—can establish causation; causal potential is determined by the degree of control and validity built into the research design. An example of this issue is found in Amdur's (1989) critique of causal models of delinquency. Amdur reanalyzed six studies that made causal claims with cross-sectional correlational data, pointing out conceptual and methodological problems with each. He concluded that much less is known about the causes of delinquency than has been claimed.

CONFIRMING A MODEL

Another problem concerns the use of the term *confirmed*. Some researchers seem to believe that confirmation of a structural model implies proof or exclusive validation of the model (Biddle & Marlin, 1987). Games (1988) drew a striking parallel between claiming support for a model and the faulty logic used when affirming the consequent. Unfortunately, evidence cannot always validate a model because it is possible that many other models may be equally acceptable. Garrison (1986) referred to this dilemma as the "uninstead of using a one-shot structural model assessment that most likely provides little evidence about causality, in and of itself. For an example of how models can be compared, see Huba, Wingard, and Bentler (1981). An excellent example of using two methods (SEM for correlation design, ANOVA for experimental design) to examine computerized training systems can be found in Coovert, Salas, and Ramakrishna (1992). Lawton, Kleban, Dean, Rajagopal, and Parmelee (1992) examined a confirmatory factor model over five different age-level samples to investigate a model of positive and negative affect over the life span.

LATENT VARIABLES

Bentler (1980) attributed part of the controversy surrounding SEM to its use of latent constructs that typically cannot be observed or directly measured. Latent-variable models introduce additional ambiguity to causal inference that directly observed variables do not (Mulaik, 1987b). Clearly, some researchers may be uneasy trying to establish causality through unobservable constructs.

Some confusion also stems from the possibility that what serves as one researcher's measurement model may serve as another researcher's structural model (P. Cohen, J. Cohen, Teresi, Marchi, & Velez, 1990). For example, P. Cohen et al. (1990) asserted that, although some constructs are clearly emergent and some are clearly underlying causes, the nature of others can be controversial. Socioeconomic status (SES) raises such controversy because it is unclear whether high income, prestigious occupation, and high educational achievement cause an individual's SES or vice versa (P. Cohen et al., 1990). For a discussion of the usefulness and the operationalization of latent variables, see Huba and Bentler (1982), Huba, Wingard, and Bentler (1981), Taking a pragmatic stance, Martin (1987) asserted that most psychologists lack the resources to obtain the high-quality data on which to base latent constructs. However, uncertainty in defining latent variables may be reduced as the number of indicators and their individual validities increase (Cliff, 1983). Despite the desirability of using four or more measured variables per latent construct (Mulaik, 1987b), a review of 15 SEM applications revealed that 6 of the studies examined used only two indicators for at least one latent variable (P. Cohen et al., 1990). Cliff (1983) quickly pointed out that the status of a latent variable with three or four indicators, each with a correlation at .7, is still ambiguous. Even when four or more indicators are used, it is still highly possible that alternative sets of parameters would be equally consistent with the data and might lead to totally different conclusions concerning the nature of the latent variables. Passing the four-indicator test of a single common factor is a necessary, not a sufficient, condition that a single common factor has been identified.

Thus, two distinct issues arise when considering measured and latent variables: the difficulty of adequately assessing a latent construct, particularly when one measured variable is employed, and the difficulty of adequately naming a construct, regardless of the number of indicators that are used. It should be noted that both of these are validity issues that plague not only SEM but other statistical techniques as well. The effects of incorrectly specifying a latent variable can lead to the same sort of interpretational errors that are associated with an incorrect experimental manipulation (e.g., Tanaka, Panter, Winborne, & Huba, 1990).

POST HOC ADJUSTMENTS

representation. Thus, it would seem that potentially stronger causal statements might be drawn from a model that has not been adjusted than from one that has. If objectivity is a primary concern, this consideration should be kept in mind before adjustments are made.

FAILURE TO REPLICATE OR CROSS-VALIDATE

Many, if not most, SEM applications involve a single, one-shot model that may have had post hoc adjustments. This provides little or no opportunity to assemble causal evidence. If post hoc adjustments are made, however, researchers should replicate and/or cross-validate their findings on other data. Replication involves testing a model in a different sample under different conditions. Cross-validation is more stringent because it requires that the same parameter estimates from an initial sample be used in a second, independent sample. The extent to which the model fits the data with the restricted estimates provides some evidence of the generality of the model. Only portions of data not used in determining solutions for parameter estimates should be used to validate hypotheses because only these data will be independent in content from the hypotheses (Mulaik, 1990). Breckler (1990) suggested that researchers divide the original sample into two parts: a derivation sample and a cross-validation sample. The derivation sample could be used to fit the initial model and to develop modifications, whereas the cross-validation data could be used to assess the fit of the adjusted model (Breckler, 1990). Such cross-validation is essential when highly efficient computerized procedures—such as are used with SEM—may increase the likelihood that chance associations are processed as if they were real. Analyzing new data may clarify whether overfitting or fitting the model to

One benefit of SEM is that it allows latent constructs to be represented by multiple measures (e.g., Martin, 1987). This may be very advantageous for psychologists because it is unlikely that single measures can represent most major psychological constructs. Martin claimed that this is SEM's greatest asset. The use of multiple indicators may provide us with more valid and more reliable measurement of latent constructs. Latent variables can be considered a tradeoff. They raise issues surrounding the "nominalistic fallacy," but multiple indicators may provide accurately defined latent constructs and may reduce the severity of missing variables by providing more richly defined latent constructs.

Furthermore, using latent variables may allow researchers to use a limited number of exploratory constructs to explain phenomena. Philosophical justification for this use of latent variables is found in the works of both Thurstone and Kant. Thurstone advocated using a limited number of explanatory constructs common to a broad range of phenomena as a way of achieving objective constructs that are applicable to all phenomena, not just specific phenomena on an ad hoc basis. For a more complete analysis of the philosophical contributions of Thurstone and Kant to SEM, refer to Mulaik (1994, in press).

Second, Bentler (1980) asserted that SEM's great potential rests in its ability to handle very complex, multivariate models—particularly with quasi-experimental or nonexperimental research, in which methods for testing are not well developed. This may improve our ability to draw causal inferences because testing more sophisticated models and theories with good data and cross-validation may allow us to get a richer, more complete understanding of the phenomenon studied.

Third, even Cliff (1983) acknowledged that SEM is a powerful tool when substance abuse, whereas physical abuse tends to show both direct and indirect effects on substance use.

If causal evidence can be accrued through replication and cross-validation with rigorous designs and analyses, a procedure that allows multiple indicators and complex relations is necessary. Such a description aptly fits SEM.

SEM AND CAUSALITY: AN ASSESSMENT

In conclusion, does SEM offer anything "new" at the level of causal inference? Based on the previous discussion, the best answer appears to be both yes and no. In one sense, the answer is clearly negative because SEM cannot ensure that the necessary conditions of isolation, association, and direction of influence have been met. However, it should be noted that no other statistical procedure can ensure that these conditions have been met either. Although some could argue that SEM may not offer anything new at the level of causal inference, it does not offer anything less than other analyses.

On the other hand, the answer is affirmative because SEM does offer what can be referred to as compelling potential. The potential lies in SEM's ability to analyze direct and indirect effects, assess both measurement and prediction error, allow multiple measures to represent latent variables, and provide simultaneous estimation of measured and structural relations in a complex, integrated mathematical model. Although using latent variables may increase ambiguity, making causal inferences difficult, they also allow complex theories to be tested. Our world is a complex place, and, if causal evidence is ever to be effectively acquired, it will only be through designs and statistical procedures that can take meet all necessary conditions; however, we cannot passively accept their inaccessibility. We advocate an active scientific approach that employs a number of rigorous procedures in an effort to begin to assess causality. In the absence of causal guarantees, Garrison (1986) encouraged researchers to practice the pragmatic virtues of epistemological conservatism and good sense. Researchers must actively rely and insist upon randomization, repetition, and replication, regardless of whether SEM or other analyses are used. Conclusions drawn from SEM must be drawn carefully and tenuously, just as with conclusions from other analyses. No analyses, alone, can directly prove causal relations. Researchers must remain critical of findings and not accept or draw causal conclusions simply because a sophisticated latent variable analysis was conducted (e.g., Martin, 1987). As SEM becomes more familiar, it is hoped that the confusion surrounding these procedures will diminish and that more emphasis will be placed on improving designs and the quality of data, thereby increasing the potential for building accurate causal evidence with SEM.

FINAL PRECAUTIONS AND GUIDELINES

By way of summary, several practical suggestions are offered to help researchers concretize some of the abstract prescriptions surrounding causality and SEM:

1. Regardless of the level of sophistication, no statistical procedure can ensure that the necessary conditions for establishing causality have been met. Be skeptical of any research that makes absolute causal claims. As with

a. Assess the relevance of and control for as many background conditions as possible.

b. Strive for longitudinal SEM designs to help assess the direction of causality.

c. Carefully operationalize latent variables.

d. Use four or more high-quality indicators per latent variable when it is appropriate.

e. Compare alternative models for a set of data.

f. Keep post hoc adjustments to a minimum.

g. Replicate and cross-validate all findings.

h. View each SEM study as just one part of a larger program of research to help understand a phenomenon. Other studies could include additional SEM and preferably one or more carefully planned experimental designs to further validate the findings. Certainly, results that hold over several studies using different designs provide stronger evidence than can be obtained from a single study. Several excellent studies that have incorporated a number of the aforementioned suggestions include Reynolds (1989) and Reynolds and Walberg (1991, 1992).

Following these general guidelines cannot guarantee that causal relations have been established with certainty. They do, however, increase the confidence with which tentative causal evidence can be accrued in a rigorous, ongoing scientific inquiry.

ACKNOWLEDGMENTS

This research was supported in part by National Institute of Mental Health

Bentler, P. M. (1978). The interdependence of theory, methodology, and empirical data: Causal modeling as approach to construct validation. In D. B. Kandel (Ed.), *Longitudinal research on drug use* (pp. 267–302). New York: Wiley.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. In M. R. Rosenweig & L. W. Porter (Eds.), *Annual review of psychology* (Vol. 31, pp. 419-456). Stanford, CA: Annual Review, Inc.

Bentler, P. M., & Mooijaart, A. (1989). Choice of structural model via parsimony: A rationale based on precision. *Psychological Bulletin, 106,* 315–317.

Biddle, B. J., & Marlin, M. M. (1987). Causality, confirmation, credulity, and structural equation modeling. *Child Development, 58,* 4–17.

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

Breckler, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin, 107,* 260–273.

Brook, J. S., Whiteman, M., Cohen, P., & Tanaka, J. S. (1992). Childhood precursors of adolescent drug use: A longitudinal analysis. *Genetic, Social, and General Psychology Monographs, 118,* 195–213.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24,* 445–455.

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

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

Cliff, N. (1987). Comments on Professor Freedman's paper. *Journal of Educational Statistics, 12,* 158–160.

Cohen, P., Cohen, J., Teresi, J., Marchi, M., & Velez, N. C. (1990). Problems in the measurement of latent variables in structural equations causal models. *Applied Psychological Measurement, 14,* 183–196.

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

Coovert, M. D., Salas, E., & Ramakrishna, K. (1992). The role of individual and system characteristics in computerized training systems. *Computers in Human Behavior, 8,* 335–352.

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

Freedman, D. A. (1987b). A rejoinder on models, metaphors, and fables. *Journal of Educational Statistics, 12,* 206–223.

Games, P. A. (1988, April). Invited commentary: "Correlation and causation: An alternative view." *The Score, XI,* pp. 9–10.

Games, P. A. (1990). Correlation and causation: A logical snafu. *Journal of Experimental Education, 58,* 239–246.

Garrison, J. W. (1986). Some principles of postpositivistic philosophy of science. *Educational Researcher, 15,* 12–18.

Guida, F. V., Ludlow, L. H., & Wilson, M. (1985). The mediating effect of time-on-task on the academic anxiety/achievement interaction: A structural model. *Journal of Research and Development in Education, 19,* 21–26.

Guttman, L. (1976). What is not what in statistics. *The Statistician, 26,* 81–107.

Harlow, L. L., Newcomb, M. D., & Bentler, P. M. (1986). Depression, self-derogation, substance use, and suicide ideation: Lack of purpose in life as a mediational factor. *Journal of Clinical Psychology, 42,* 5–21.

Huba, G. J., Wingard, J. A., & Bentler, P. M. (1981). A comparison of two latent variable causal models for adolescent drug use. *Journal of Personality and Social Psychology, 40,* 180–193.

Huba, G. J., & Bentler, P. M. (1982). On the usefulness of latent variable causal modeling testing theories of naturally occurring events (including adolescent drug use): A rejoinder to Martin. *Journal of Personality and Social Psychology, 43,* 604–611.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data.* Beverly Hills, CA: Sage.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika, 57,* 239–251.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 85–112). New York: Academic.

Kandel, D. B. (1975). Stages in adolescent involvement in drug use. *Science, 190,* 912–914.

Kaplan, D. (1989). A study of the sampling variability and z-values of parameter estimates from

Mulaik, S. A. (1994). Kant, Wittgenstein, objectivity, and structural equations modelling. In C. R. Reynolds (Ed.), *Advances in cognitive assessment: An interdisciplinary approach* (pp. 209–236). New York: Plenum.

Mulaik, S. A. (in press). The critique of pure statistics: Artifact and objectivity in multivariate statistics. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 3). Greenwich, CT: JAI.

Mumford, M. D., Weeks, J. L., Harding, F. D., & Fleishman, E. A. (1988). Relations between student characteristics, course content, and training outcomes: An integrative modeling effort. *Journal of Applied Psychology, 73,* 443–456.

Newcomb, M. D., & Bentler, P. M. (1987). The impact of late adolescent substance use on young adult health status and utilization of health services: A structural-equation model over four years. *Social Science and Medicine, 24,* 71–82.

Newcomb, M. D., McCarthy, W. J., & Bentler, P. M. (1989). Cigarette smoking, academic lifestyle, and social impact efficacy: An eight-year study from early adolescence to young adulthood. *Journal of Applied Social Psychology, 19,* 251–281.

Ragosa, D. R. (1987). Causal models do not support scientific conclusions: A comment in support of Freedman. *Journal of Educational Statistics, 12,* 206–223.

Reynolds, A. J. (1989). A structural model of first-grade outcomes for an urban, low socioeconomic status, minority population. *Journal of Educational Psychology, 81,* 594–603.

Reynolds, A. J., & Walberg, H. J. (1991). A structural model of science achievement. *Journal of Educational Psychology, 83,* 97–107.

Reynolds, A. J., & Walberg, H. J. (1992). A structural model of science achievement and attitude: An extension to high school. *Journal of Educational Psychology, 84,* 371–382.

Steiger, J. H. (1980). [Review of *Correlation and causality*]. *Psychometrika, 45,* 403–405.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25,* 173–180.

Stein, J. A., Newcomb, M. D., & Bentler, P. M. (1987). Personality and drug use: Reciprocal effects across four years. *Personality and Individual Differences, 8,* 419–430.

Tanaka, J. S., Panter, A. T., Winborne, W. C., & Huba, G. J. (1990). Theory testing in personality and social psychology with structural equation models: A primer in 20 questions. *Review of Personality and Social Psychology, 11,* 217–242.

Velicer, W. F., Huckel, L. H., & Hansen, C. E. (1989). A measurement model for measuring attitudes toward violence. *Personality and Social Psychology Bulletin, 15,* 349–364.