10-31-2003

Risky Predictions, Damn Strange Coincidences, and Theory Appraisal: A Multivariate Corroboration Index for Path Analytic Models

Kristine Y. Hogarty
University of South Florida

Risky Predictions, Damn Strange Coincidences, and Theory Appraisal: A Multivariate Corroboration Index for Path Analytic Models

by

KRISTINE Y. HOGARTY

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Educational Measurement and Research
College of Education
University of South Florida

Major Professor: Jeffrey D. Kromrey, Ph.D.
Robert Dedrick, Ph.D.
John Ferron, Ph.D.
Kathleen McNelis, Ph.D.

Date of Approval:
October 31, 2003

Keywords: theory testing, model fit, path analysis, verisimilitude, precision

Acknowledgement

At a time such as this, it is quite liberating to take the time to reflect on the many influences that came to bear not only on the production of this research but perhaps more importantly on the shaping of thoughts and ideas during this most challenging, yet stimulating journey.

After a rather circuitous route, I was most fortunate to find a home in the Department of Educational Measurement and Research. During all of my years studying business and criminology I never would have predicted (albeit risky or otherwise) that I would find my true passion in statistics and research. I have been truly overwhelmed by the nurturing nature of the faculty in this department, and it is with the deepest appreciation that I thank my committee members that are part of this most dedicated faculty. Dr. Bob Dedrick and Dr. John Ferron, I thank you for your thoughts, your recommendations and suggestions and your enthusiasm and support for my research. I thank Dr. Kathy McNelis for keeping my grounded in the real world and for giving me something to think about in the elevator. And, as my major professor, Dr. Jeff Kromrey, you have helped me grow in so many ways. I am so very thankful for your insight and guidance, your firm yet kind words, and of course, all of those gentle nudges along the way. Your drive and your passion for the profession are truly inspirational.

On a closing note, I would certainly be remiss if I did not acknowledge the fortitude of my family…to Mom and Kevin, we finally did it!

Dedication

For her unwavering support and enthusiasm for everything that I have ever aspired to accomplish, it is with loving memories that I dedicate this dissertation to Ethel.

Table of Contents

List of Tables

List of Figures

Risky Predictions, Damn Strange Coincidences, And Theory Appraisal: A Multivariate Corroboration Index for Path Analytic Models

Kristine Y. Hogarty

ABSTRACT

The elucidation and empirical testing of theories are important components of research in any field. Yet despite the long history of science, the extent to which theories are supported or contradicted by the results of empirical research remains ill defined. Quite commonly, support or contradiction is based solely on the "reject" or "fail to reject" decisions that result from tests of null hypotheses that are derived from aspects of theory. Decisions and recommendations based on this forced and often artificial dichotomy have been scrutinized in the past.

In recent years, such an overly simplified approach to theory testing has been challenged on logical grounds (Meehl, 1997, 1990, 1978; Serlin & Lapsley, 1985). Theories differ in the extent to which they provide precise predictions about observations. The precision of predictions derived from theories is proportional to the strength of support that may be provided by empirical evidence congruent with the prediction. However, the notion of precision linked to strength of support is surprisingly absent from many discussions regarding the appraisal of theories.

Meehl (1990a) has presented a logically sound index of corroboration to summarize the extent to which empirical tests of theories provide support or contradiction of theories. The purpose of this study was to evaluate the utility of this index of corroboration and its behavior when employing path analytic methods in the context of social science research.

The performance of a multivariate extension of Meehl's Corroboration Index (C_i) was evaluated using Monte Carlo methods. Correlational data were simulated to correspond to tests of theories via traditional path analysis. Five factors were included in the study: number of variables in the path model, level of intolerance of the theory, correspondence of the theory to the 'true' path model used for data generation, sample size and level of collinearity.

Results were evaluated in terms of the mean and standard error of the resulting multivariate C_i values. The level of intolerance was observed to be the strongest influence on mean C_i. Verisimilitude and model complexity were not observed to be strong determinants of the mean C_i. Sample size and collinearity evidenced small relationships with the mean value of C_i, but were more closely related to the sampling error.

Implications for theory and practice include alternatives and complements to tests of statistical significance, a shift from comparing findings to the null hypothesis, to the comparison of alternative theories and models, and the inclusion of additional logical components besides the theory itself. Lastly, an alternative conceptualization of the multivariate corroboration index is advanced to guide future research efforts.

Chapter One

Introduction

Background

The elucidation and empirical testing of theories are important components of research in any field. Kerlinger (1964) suggested that these components are fundamental distinctions between science and common sense. Yet despite the long history of science, the extent to which theories are supported or contradicted by the results of empirical research remains ill defined. Often such support or contradiction is reduced to the "reject" or "fail to reject" decisions resulting from tests of null hypotheses that are derived from aspects of theory. That is, a theory is "supported" by empirical evidence if null hypotheses are rejected, when the theory suggests they should be rejected. Conversely, a theory is contradicted (and may be considered "refuted," cf. Popper, 1959) if such theoretically derived null hypotheses are not rejected. The limitations of null hypothesis testing are well known (e.g., Harlow, Mulaik, & Steiger, 1997), and such testing has been the subject of much criticism and controversy (Kirk, 1972; Morrison & Henkel, 1970). Over the years, a considerable amount of doubt has been cast on the merit of null hypothesis testing as a *theoretical tool*. The use of this approach in the testing of theories presents unique conceptual challenges and interpretational dangers.

According to Thompson (2002) the field of psychology has witnessed a lengthy deliberation about the utility of statistical significance, questioning whether these tests should be banned from journals of the American

Psychological Association (APA). A task force charged with examining this issue did not endorse a ban on these tests, but rather articulated a wide-ranging set of recommendations for improved inquiry and reporting (Wilkinson & APA Task Force on Statistical Inference, 1999). Recommendations that were adopted and included in the recent fifth edition of the APA (2001) *Publication Manual*, inform potential authors to include "information about the obtained magnitude or value of the test statistic, the degrees of freedom, the probability of obtaining a value as extreme or more extreme than the one obtained, and the direction of the effect" (p. 22). Additionally, the reporting of confidence intervals (for parameter estimates and functions of parameters and for effect sizes) is strongly recommended. According to Thompson (2002), this move represents a positive step forward with respect to improved intellectual inquiry that is less reliant on null hypothesis significance tests (NHST) and requires a heightened sense of responsibility on the part of the research community.

Still others question why reform has proceeded further in some other disciplines, including medicine, than in psychology. A few researchers contend that what has happened in psychology was not inevitable. "We leave to historians and sociologists of science the fascinating and important question of why psychology has persisted for so long with poor statistical practice" (Finch, Cummings, & Thomason, 2001, p. 205-206). The persistence of poor statistical practices in a broad range of disciplines in the social sciences is particularly vexing. This conundrum suggests that it would be profitable to explore

alternatives to traditional approaches to theory testing and consider underutilized, different or yet to be developed statistical tools.

Need for Another Tool

In recent years, such an overly simplified approach to theory testing has been challenged on logical grounds (Meehl, 1997, 1990, 1978; Serlin & Lapsley, 1985). Theories differ in the extent to which they provide precise predictions about observations. The precision of predictions derived from theories is proportional to the strength of support that may be provided by empirical evidence congruent with the prediction. That is, a precise prediction that is supported by data warrants more logical evidence in support of the theory than does a weak prediction supported by data. This relationship between the precision of prediction and the strength of logical support is rooted in the relative rarity of the data, absent the theory. That is, without the theory, would we expect to see such data anyway? The extent to which we would *not* expect to see such data is what Salmon (1984) refers to as a "damn strange coincidence," and the extent to which a theory predicts such otherwise rare data is a "risky p (Meehl, 1978).

The degree to which theories differ in their precision of prediction can be illustrated with a simple example. Consider one of the most basic predictions, that is, predictions about population mean differences. A simple prediction that men and women will have *different* means on a given variable is a relatively weak prediction. A prediction that the mean of women will be *greater than* that of

men is somewhat stronger, a prediction that the means will differ by *some value between* five and ten points is stronger yet, and a prediction that the means will differ by *exactly* 7 points is even more precise. The precision of prediction can be conceptualized as existing along the continuum presented in Figure 1.

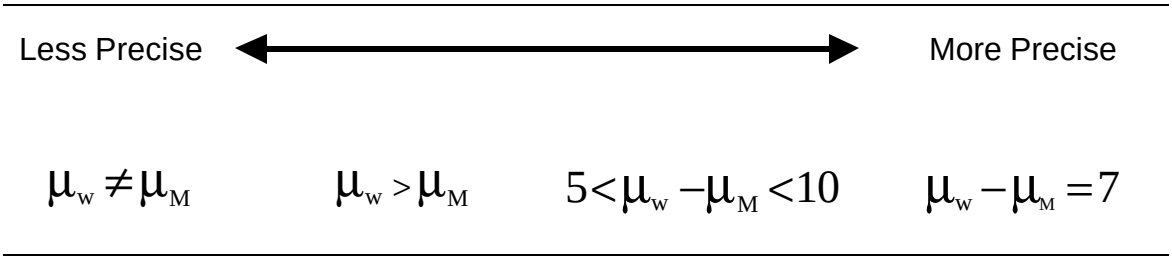

Figure 1. Continuum of Precision.

Naturally, this basic example can be extended if a researcher desires to make predictions about the direction and magnitude of relationships among variables, and with additional adjustment, can be modified to extend beyond univariate analyses to include multivariate contexts.

Meehl's Index of Corroboration (C_i)

Meehl (1997, 1990a) has proposed an index of corroboration (C_i) that may provide a standardized means of expressing the extent to which empirical research supports or contradicts a theory:

$$C_i = (Cl)(In)$$

Where *Cl* = the "closeness" of the data to the theoretical prediction (verisimilitude or truth-likeness), and

In = the "intolerance" of the theory (e.g., a standardized precision of prediction).

These terms are further defined as follows:

$$CI = 1 - (D/S)$$

Where D = the deviation of observed data from the tolerance interval of the theory

S = Spielraum (the range of data values that are expected whether or not the theory is true)

$$In = 1 - (I/S)$$

Where *I* = the interval tolerated by the theory (e.g., the raw precision of prediction).

As it is quite common to encounter the use of good fit to support theories in the research community and a host of these indices already exist, one may ask, why do we need another fit index? While the proposed index may appear to resemble methods commonly used to establish goodness of fit, the additional component that represents the precision of prediction (i.e., the intolerance of the prediction) makes this index somewhat unique. As conceptualized previously, the proposed corroboration index combines both a traditional measure of "fit", represented by the deviation of observed data from the tolerance interval of the theory and the degree of precision with which the prediction is advanced.

Further, the sheer logic of appraising a scientific theory is often more complicated than some would believe (Meehl, 1997). In addition to the

aforementioned argument regarding the precision of prediction (that is, a precise prediction that is supported by the data warrants more logical evidence of support than does a weak prediction supported by the data), the movement from theory into an empirical test necessitates the incorporation of many logical components besides the theory itself. Meehl (1997) presents these components as elements of an equation:

$$\left(T \bullet A_x \bullet C_p \bullet A_i \bullet C_n\right) \rightarrow \left(O_1 \supset O_2\right)$$

Where T = the theory being "tested,"

A_x = Auxiliary theories relied upon during the conduct of the research.

C_p = *Ceteris paribus* (all other things being equal),

A_i = Instrumental theories related to measures and controls employed,

C_n = Realized particulars (the extent to which the research was actually conducted as we think it was), and

$O_1 \supset O_2$ = the material conditional "if you observe O_1, you will observe O_2."

That which is subject to empirical test is not the theory alone, but the amalgam of these elements. Data that appear to contradict a "theory" may arise because of errors anywhere in this combination of elements (e.g., the theory may be correct but the groups we thought were equivalent were actually systematically different from each other on an important, confounding variable). In the following chapter, the elements of this amalgam are further elucidated through the use of an illustrative example representing these components in the conduct of disciplined inquiry.

Statement of the Problem

Meehl (1997, 1990a) has presented a logically sound index of corroboration to summarize the extent to which empirical tests of theories provide support or contradiction of those theories. However, the numerical properties of this index have not been investigated beyond some of the most basic predictions about population mean differences, zero order, and first-order partial correlations (Hogarty & Kromrey, 2002, 2001, 2000). Monte Carlo methods were used in the previous studies to examine the behavior of the corroboration index. These methods currently remain the most feasible way to study this index and thus a similar approach was followed in this study.

Purpose

The purpose of this study was to build upon the previous research by advancing to the next logical step, by evaluating the utility of the corroboration index and its behavior when appraising theories employing path analytic methods in the context of social science research. A simulation study was used to evaluate the behavior of the index of corroboration under a variety of encountered conditions that are typically encountered in the conduct of empirical research.

Many researchers approach path analysis by beginning with a model in which there is substantial confidence. This confidence may stem from either theoretical or substantive reasoning about the linkages between the variables under investigation. Less attention, however, is typically given to estimating the

magnitude of such linkages. Most areas of psychology do not permit a high degree of precision. According to Blaich (1998), quasi-quantitative predictions of rough magnitudes of effects could help advance the field. Therefore, the primary focus of this investigation was on the precision in the prediction of the magnitude of effects and an examination of factors that moderate the relationship between corroboration and precision.

An extension of this nature required modifications of the index as it was originally conceptualized. For example, when considering more than one parameter estimate, the formula used in the calculation of closeness was:

$$CL = \left[\prod_{j=1}^{J} \left(1 - \frac{D_j}{S_j} \right) \right]^{\frac{1}{J}}$$

Where J is equal to the number of parameters (i.e., path coefficients) being estimated.

Additionally, an earlier exploration of variations in the calculation of intolerance resulted in the following formula deemed to be most appropriate:

$$In = 1 - \prod_{j=1}^{J} \frac{I_j}{S_j}$$

For a complete discussion of the rationale behind the selection of these estimates of multivariate closeness and intolerance, consult the section entitled Multivariate Extension of C_i provided in Chapter Three.

Four Research Questions and Three Research Hypotheses

Research Questions

1. What is the relationship between mean C_i and the main effects examined in the study (i.e., verisimilitude, intolerance, model complexity, collinearity, and sample size)?
2. What is the relationship between the standard deviation of C_i and model complexity, collinearity, and sample size?
3. To what extent is the relationship between mean C_i and the precision of prediction (i.e., intolerance) influenced by the complexity of the model (i.e., the number of variables in the model)?
4. To what extent is the relationship between mean C_i and the precision of prediction (i.e., intolerance) influenced by the level of collinearity?

Research Hypotheses

1. The relationship between mean C_i and the precision of prediction (i.e., intolerance will be slightly influenced by the closeness of the data to the theory (verisimilitude).
2. The relationship between mean C_i and the precision of prediction (i.e., intolerance) will not be substantively influenced by sample size.
3. The relationship between mean C_i and precision of prediction will be substantively stronger than the relationship between mean C_i and verisimilitude, model complexity, collinearity, and sample size.

Delimitations and Limitations

This examination is limited to an exploration of the aforementioned relationships through the use of traditional path analyses employing least squares regression analysis. The focus of this inquiry is on the magnitude of path coefficients obtained by examining the relationship between observed variables, rather than latent variables, which would necessitate the employment of more sophisticated methods, such as structural equation modeling techniques. Additionally, this study includes models in which the causal flow is unidirectional, that is, the investigation of a series of recursive models.

Organization of the Study

In the first chapter, the reader is acquainted with the ongoing controversy and some of the past and current thinking with respect to tests of null hypotheses that are derived from aspects of theory. The rationale and purpose for the study are outlined and the research questions and hypotheses are advanced. Included in this chapter are the delimitations, limitations and important definitions central to the study.

Chapter Two is devoted to a review of the relevant literature that coheres around the central theme of theory testing in the social sciences, focusing on the central issues related to inquiry of this nature. Although little research to date has been conducted on the behavior of Meehl's index of corroboration, related literatures that required exploration included the philosophy of science, theory testing, path analysis, and an examination of fit indices in the context of structural

equation modeling. The index of corroboration is operationally defined in this chapter and a illustrative example is provided. Reactions to the notion of a corroboration index and past research are also discussed. Lastly, the importance of the study is explicated.

Chapter Three outlines the method employed, and describes the central design factors, the procedures and data analysis strategy employed in this study. The procedures outlined here include the selection of the components of the multivariate corroboration index, the conduct of the study, and the data generation strategy.

Chapter Four presents the results of the study. The results are organized with respect to each of the research questions and hypotheses. In addition, within each section, one the primary research questions and hypotheses have been addressed, supplementary analyses and results are examined in order to further elucidate some of the more subtle relationships evidenced in the data. The chapter concludes with a summary of the chapter key findings.

Chapter Five provides a sound set of conclusions that are firmly grounded in the results of the study, the findings of past empirical research and the body of literature that coheres around the central theme of theory testing in the social sciences. Following this recapitulation of the major findings of the study, important implications for practice and theory are advanced. The chapter concludes with recommendations for future research.

Definitions

Closeness. The "closeness" of the data to the theoretical prediction (verisimilitude or truth-likeness) defined as 1 minus the deviation divided by the Spielraum. Cl = 1 – (D/S) (Meehl, 1990a).

Collinearity. Refers to correlations among independent variables. Literally, collinearity refers to the case of data vectors representing two variables falling on the same line, that is, two variables that are perfectly correlated. However, most researchers use the term to imply *near* collinearity among a set of independent variables (Pedhazur, 1997).

Corroboration index (C_i). This index is defined as a standardized measure of the extent to which empirical research supports or contradicts a theory. The index is defined as closeness (Cl) multiplied by intolerance (In) (Meehl, 1990a).

Deviation (D). The deviation of the observed data from the tolerance interval of the theory (Meehl, 1990a).

Endogenous variable. In a causal model, an endogenous variable "is one whose variation is explained by exogenous or other endogenous variables in the model " (Pedhazur, 1997, p. 770).

Empirical Fit. The degree of congruence (or fit) between the hypothesized model and the observed data (Hu & Bentler, 1999).

Exogenous variable. In a causal model, "an exogenous variable is one whose variation is assumed to be determined by causes outside the

hypothesized model" (Pedhazur, 1997, p. 770). That is, a variable that lacks hypothesized causes in the path analysis model.

Interval (I). The interval tolerated by the theory. The unstandardized precision of prediction (Meehl, 1990a).

Intolerance (In). The standardized precision of prediction. Intolerance is defined as 1 minus the interval divided by the spielraum. In = 1 – (I/S) (Meehl, 1990a).

Model Misspecification. For this study, model misspecification can occur when "true" paths are omitted or ancillary paths are included in the model. In this context, model misspecification is reflected in the level of verisimilitude.

Path coefficient. A standardized regression coefficient indicating the direct effect of one variable on another variable in path analysis. "For each independent variable in the equation, there is a path coefficient indicating the amount of expected change in the dependent variable as a result of a unit change in the independent variable" (Pedhazur, 1997, p. 772).

Path model. A diagram that graphically displays 'the hypothesized pattern of causal relations among a set of variables" (Pedhazur, 1997, p. 770).

Recursive model. A model that considers only unidirectional causal relationships, that is "reciprocal causation between variables is ruled out" (Pedhazur, 1997, p. 771).

Spielraum (S). The range of data values that are expected whether or not the theory is true (Meehl, 1990a).

Theoretical Fit. "The degree of isomorphism between a theoretical model and a true model" (Olssson, Troye, & Howell, 1999, p. 31).

Theory. "A set of interrelated constructs, definitions, and propositions that present a systematic view of phenomena by specifying relations among variables, with the purpose of explaining phenomena" (Kerlinger, 1964).

Variance Inflation Factor (VIF). This component indicates the inflation of the variance of *b (*i.e., the estimated path coefficient*)*, resulting from the correlation between two or more independent variables (Pedhazur, 1997).

Verisimilitude. The closeness of the observed data to the theoretical prediction (truth-likeness). "Verisimilitude is an ontological concept; that is, it refers to the relationship between theory and the real world" (Meehl, 1990a, p. 133).

Chapter Two

Review of the Literature

Overview

The review of the literatures coheres around the central theme of theory testing in the social sciences. This chapter is divided into six major sections: Philosophy of Science, Theory Testing, Path Analysis, Fit Indices, Meehl's Index of Corroboration C_i, and Importance of the Study. The chapter is organized in this manner to facilitate communication of the central issues by eliciting insight from the extant literature, and to develop a balanced landscape for the presentation of competing methodologies. Although the chapter is physically divided into major and minor subsections, at times, no true conceptual boundaries may exist.

The review of the literature begins with a broad overview, recounting the history of the philosophy of science by tracing the evolution of thought and practice that have characterized theory appraisal over the past few decades. Once the contemporary origins of this discipline have been explored, the review is naturally extended through an examination of traditional methods employed in theory testing. In this section, common approaches are described, methodological obstacles and objections are advanced, and observed deficiencies inherent in hypothesis testing approaches are uncovered. In addition, there is a brief review of common features of theories that are often deemed desirable. This section invites readers to extend their thinking beyond

the consideration of a single or traditional indicator of verisimilitude, suggesting a broader milieu that might include supplementary indices and/or other variables.

Following this section, attention is then directed to a précis of path analysis, one of the most widely employed methods in the testing of theories. This section examines some of the methodological nuances of this popular method and considers certain conditions that need to be satisfied when utilizing this type of statistical technique. To further elucidate this approach, a brief review of structural equation modeling is presented. A central element of this statistical modeling method, the determination of empirical fit, is introduced via a commentary on the similarities and differences, as well as the inherent strengths and weaknesses of some of the more commonly employed indices of goodness of fit. This section concludes by addressing the question, "How persuasive is a

In the fifth section of this chapter, the corroboration index is reintroduced and an argument for the incorporation or adoption of an index of corroboration is advanced (Meehl, 1997, 1990a). This discussion is augmented with some of the comments and criticisms presented by contemporary scholars. This major section of the literature review concludes with a summary of the recent empirical research that establishes a firm foundation for the current research endeavor. The review of the literature is brought to a close by addressing the importance of the current study and thus reveals the potential utility of a corroboration index in the evaluation of theories across a vast array of domains and disciplines.

Philosophy of Science

According to Kuhn (1962), the history of any science can be described by a succession of incommensurable paradigms. In this view, competing paradigms do not agree on what constitutes knowledge or the meaning of truth, with empirical work done in one paradigm having little importance or relevance to another. Beliefs constituting a paradigm are so fundamental they are immune from empirical testing. In this regard, experimental failures may lead to the rejection of specific theories, but the paradigm remains untouched, directing the construction of new theories. The recent work of other philosophers of science such as Lakatos and Lauden stand in contrast to those views held by Kuhnians, suggesting that research programs are not incommensurable, and evolve in ways not predicted by Kuhn (Gholson & Barker, 1985).

Lakatos (1970) substituted the Kuhnian paradigm with a "research program" that involves a succession of theories. A "hard core" of shared commitments links theories, each successive theory introducing a new and more detailed articulation of these commitments. Accordingly, a protective belt of dispensable hypotheses provides shelter from immediate empirical refutation. Dispensable features are modified by successive theories with the core assumptions remaining intact. The ability to stimulate the development of complex and adequate theories is viewed as an objective feature and important characteristic of any research program.

Lauden (1977) replaced the notion of a "research program" with a research tradition. This extension involves families of theories sharing a

common ontology and methodology, that is, a shared vision of reality and agreement regarding appropriate ways to investigate that reality. In addition to empirical factors, conceptual factors are viewed as important in theory appraisal, and independent of experimental success or failure. Lauden also offered a solution to Lakatos's unrealistic requirement that core commitments pass unchanged through successor theories, contending that core principles are not functionally metaphysical and can be modified in response to empirical testing.

Theory Testing

For nearly three-quarters of a century, statistical significance testing has been the most widely used method of analysis in psychological experiments (Nickerson, 2000). In many areas of psychology, refutation of the null hypothesis has been the sole theory-testing procedure employed (Meehl, 1990aa). Over the years, a considerable amount of doubt has been cast on the merit of null hypothesis testing as a *theoretical* tool. Commenting on the slow progress in soft psychology, Dar (1987) stated that null hypothesis tests are destructive to theory building. According to Lykken (1968), "theory corroboration requires testing of multiple predictions because the chance of getting statistically significant results for the wrong reasons in any given case is surprisingly high" (p. 158). For example, in the social sciences we are typically concerned with many variables, some that are within our control whereas others are not. Many of these variables, although not of direct interest or central to a study, have been shown to be nuisance variables, variables that may have a significant influence, or may

interact with each other (Meehl, 1978). Additionally, the well-known influence of sample size on statistical significance tests may well be in itself cause for skepticism of unlikely statistically significant results.

Analyses need to be designed to shed light on whether a model is consistent with the data, if not, then doubt is cast about the theory from which the model was derived. Consistency, however, does not constitute proof, it merely lends support. According to Popper (1959), all research can accomplish is falsification of theory--those theories that survive are not disconfirmed. Gigerenzer (1998) has argued that the institutionalization of null hypothesis significance testing has permitted surrogates for theories to flourish resulting in one-word explanations, redescriptions, vague dichotomies, and data fitting.

Arguments against the use of tests of statistical significance abound. According to Meehl (1990a), any null hypothesis of zero correlation between two variables or of zero difference between two sample means may confidently be set up by an investigator as a straw man which often can be 'refuted', even when conceptually meaningless predictors are chosen at random. Carver (1978) contends, "statistical significance tells us nothing directly relevant to whether the results we found are large or small, and it tells us nothing with respect to whether the sampling error is large or small" (p. 291). Over the years, there has been a concerted effort aimed toward encouraging researchers to standardly provide some indication of effect size along with or in place of the results of statistical significance tests. Effect sizes have been viewed as consistent with null hypothesis significance testing and as an important complement. This move,

alluded to in the introductory chapter, represents a positive step forward with respect to improved intellectual inquiry; that is less reliance on null hypothesis significance tests (Thompson, 2002).

Meehl (1990a) contends the way in which a theory accumulates "money in the bank" is by passing several stiff tests; claiming that "the main way a theory gets money in the bank is by predicting facts that, absent the theory, would be antecedently improbable" (p 115). Theoretical support depends on a variety of factors, including the relative uniqueness of the prediction, how surprising the prediction is, the precision of prediction, and degree of correspondence between the prediction and the observed data (Nickerson, 2000).

The role of theory in the formation of causal models was perhaps most forcefully expressed by Hanson:

> Causes are connected with effects; but this is because our theories connect them, not because the world is held together by cosmic glue. The world may be glued together by imponderables, but that is irrelevant for understanding causal explanations. The notions behind "the cause *x*" and "the effect *y*" are intelligible only against a pattern of theory, namely one which puts guarantees on inferences from *x* to *y*. Such guarantees distinguish truly causal sequences from mere coincidence" (1958, p.64)

Faust and Meehl (2002) contend that in the evaluation of theories, researchers need to develop predictors of the success of theories or their

ld add rigor, qualitative diversity or breadth, reducibility upward or downward, and elegance or mathematical beauty" (p. S187). Further, these authors contend that no credible philosopher of science has ever claimed that any one of these features is a guarantee of truth or that any one feature is always superior to another.

As an aid to the future evaluation of theories, Faust and Meehl (2002) reintroduced an index of "predictive accuracy in relation to risk" (a.k.a., Meehl's C_i) and proposed that additional indices could be developed to rate qualitative diversity and parsimony. By working with a host of potential variables related to theory status, perhaps some of those traditionally advanced combined with

others less traditional or yet to be developed, we can get a better sense of predictive power, how variables are best combined, and how we might begin to cope with inconsistencies.

A contemporary debate with respect to theory appraisal and theory development can be found within the pages of a recent issue of P*sychological Methods*. Within these journal pages, scholars revisited this vital issue and offered a host of recommendations and suggestions. These include yet another statistical approach to strong appraisal of truth or verisimilitude that involved a class of path diagrams (Meehl & Waller, 2002); an evaluation of tests of statistical significance (Markus, 2002); a treatise on just-identified, recursive models as compared to the delete one and add one models proposed by Meehl and Waller (Reichardt, 2002); and commentaries on the proposed Meehl and Waller approach to path analysis and verisimilitude (see for example MacCallum, Browne, & Preacher 2002; and Mulaik, 2002). In sum, it appears that the controversy is still alive and well and continues to be on the minds of prominent scholars in the field.

Path Analysis

Causal modeling is a tremendously popular method and valuable analytical tool used extensively in the social sciences. This method is not intended to discover causes, but to shed light on the tenability of causal models. Causal models must specify both the relationships between independent and dependent variables, as well as explicitly state the relationships among all

variables considered. Each link between the variables under investigation implicitly represents a hypothesis that would be tested by estimating the magnitude of the relationship (Asher, 1983). Predicated on the assumptions of valid design and execution, this method holds greater promise of increasing awareness and understanding of more complex phenomenon than simple examinations of correlations between variables without attention to mediating or spurious relationships. The examination of causal relationships, that is the cause *x* and the effect *y*, are intelligible only against a pattern of theory, one that puts guarantees on inference from x to y. As stated previously, such guarantees distinguish truly causal sequences from mere coincidence (Hanson, 1958).

Path analysis is an extremely popular statistical method and there has been a substantial increase in the use of this type of modeling technique over the years by social and behavioral scientists. Path analysis falls within the general category of methods referred to as structural equation modeling or covariance structure analysis. This method is commonly used for analyzing systems of structural equations and allows researchers to shed light on questions regarding whether or not a proposed causal model is consistent with the data. One advantage of this technique is that it allows a researcher to investigate the utility of a proposed theoretical framework. Accordingly, a proposed theory is represented by a mathematical model. This mathematical model conveys the nature of the relationships among the variables under investigation. Consistency of a model with the data, does not however constitute proof of a theory, but rather provides support for a particular theory.

Competing or equivalent models can also be consistent with the same data. For any given covariance structure model, there will often be alternative models that may be indistinguishable from the suggested model in terms of goodness of fit (MacCallum, Wegener, Uchino, & Fabrigar, 1993). The decision regarding the tenability of such models rests not on the data but on the theory from which the model was generated.

According to Bollen (1989), contemporary applications of path analysis emphasize three components: the path diagram, the decomposition of covariances and correlations in terms of model parameters, and the distinction between direct, indirect, and total effects of one variable on another. Pedhazur (1997) claims that although a path diagram is not essential for the numerical analysis employed in path analysis, it provides a useful venue for visibly displaying hypothesized patterns of causal relationships among a set of variables. Estimates of model parameters, path coefficients, provide information with respect to the magnitude of the direct effect, or expected amount of change in a dependent variable resulting from a unit change in the independent variable, holding all others constant (Pedhazur, 1997). Path coefficients represent the individual components that result when we decompose the correlation between two endogenous variables or between an exogenous and endogenous variable. The distinction between total, direct and indirect effects arises from the relationships represented by the causal model. Consider the simple model depicted in Figure 2.

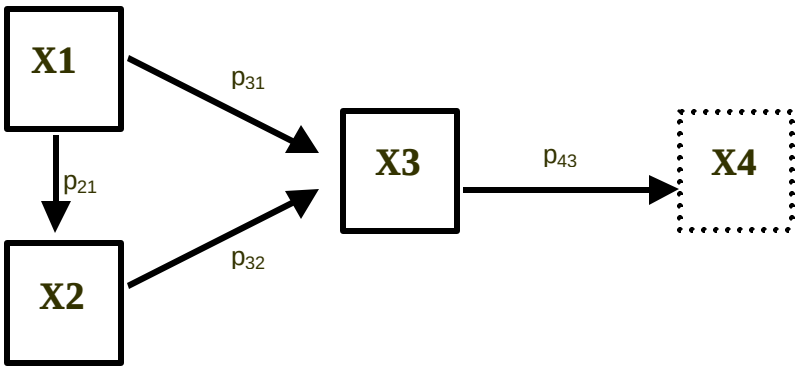

Figure 2. Mediated Causal Model.

Consider the total effect of X_1 on X_3. In this model we observe that X_1 has a direct effect on X_3, as described by p_{31}. We also observe that X_1 has an indirect effect on X_3, as mediated by X_2 (p_{21} and p_{32}). This results in a two-part decomposition, that is the direct effect of X_1 on X_3, and the indirect effect of X_1 on X_3 via X_2. The total effect of one variable on another is the sum of both direct and indirect effects (Alwin & Hauser, 1975; Duncan, 1975; Finney, 1972). Therefore, the total effect of X_1 on X_3 is calculated by adding the direct effect (p_{31}) to the product of the paths constituting the indirect effects (p_{21}*p_{32}). Additionally, direct, indirect and total effects can be produced for other variables in the model, for example the total effect of X_1 if we add X_4 to the model. The total effect of X_1 on X_4 would be obtained by calculating the product of the indirect effects (i.e., p_{21}*p_{32}*p_{43} + p_{31}* p_{43}).

Fit indices

One benefit of a structural equation modeling approach is that fit indices are available for assessing the tenability of the estimated model. Because models are only approximations of reality, they can be expected to fit only approximately. If the relationships implied by the model (as evidenced in the variance/covariance matrix) are not consistent with those observed in the data, it becomes clear that the model is inadequate and that the effect estimates should be questioned.

Hu and Bentler (1999) contend that the two most commonly employed methods of evaluating model fit are those that involve the c^2 goodness-of-fit statistics and fit indices. There are a variety of goodness of fit measures and a host of methods used in determining the goodness of fit of a proposed theoretical model. These indices generally fall into two broad categories, absolute and incremental fit indices (Bollen, 1989; Gerbing & Anderson, 1993; Hu & Bentler, 1995; Marsh, Ball, & McDonald, 1988; Tanaka, 1993). Absolute measures of goodness of fit assign a numerical value to the degree to which the proposed model reproduces the sample data (i.e., variance/covariance structure), that is, the degree of misspecification of a particular model. According to Hu and Bentler (1999) reference models are not typically used to assess the amount of increment in model fit, however, it is possible to make a comparison to a saturated model, one that accurately reproduces the sample covariance structure. Incremental indices, on the other hand, are used to compare the fit of the proposed model to the fit of a null model. If goodness of fit is adequate, it is

considered evidence for the plausibility of the model, that is, the theoretical model may accurately represent the true model. Again, for any given covariance structure model, there will often be alternative models that may fit the data equally as well (MacCallum, 1993).

The chi-square goodness of fit index is frequently used to assess the fit between the variance/covariance matrix implied by the model and the observed covariance matrix of the sample. Early investigations of the behavior of c^2 (Boomsma, 1982) revealed that this index was too dependent on sample size to be useful in many situations. There is now general acceptance of the unsatisfactory behavior of the c^2 statistic for the assessment of model fit (Thompson & Daniel, 1996). It is important to note, however, that although sample size will not cause a good model to have poor fit, with larger sample sizes minor deviations in parameters estimates are often detected.

Absolute Indices of Fit

A host of fit indices have been designed to avoid some of the pitfalls involved with sample size, violations of distributional assumptions and model misspecification, hazards that have traditionally plagued the traditional overall test of fit (i.e., the c^2 statistic).

> Absolute indices of goodness-of-fit include the Goodness-of-Fit Index (GFI) and the Adjusted Goodness-of-Fit Index (AGFI; Bentler, 1983; Joreskog & Sorbom, 1984; Tanaka & Huba, 1985), Steiger's Gamma Hat, a rescaled version of

> Akaike's information criterion (CAK, Cudeck & Browne, 1983), a cross validation index (CK, Brown & Cudeck, 1989), McDonald's (1989) Centrality Index (Mc), Hoelter's (1983) Critical N (CN), a standardized version of Joreskog and Sorborm's (1981) root mean squared residual (SRMR; Bentler, 1995), and the root mean square error of approximation (RMSEA; Steiger & Lind, 1980). (Hu & Bentler, 1999, p. 2)

There is little empirical support that these other fit indices can more unambiguously point to model accuracy as compared to the c^2 test (Hu & Bentler, 1999). Further, in their investigation of the effects of sample size, estimation methods and model specification, Fan, Thompson, and Wang (1999) issued concerns about the behavior of certain fit indices and the information that they provide with respect to misspecified models, specifically their noncomparable nature and the strong influence of estimation method. Additionally, often cited problems exist with various measures, as they are affected by sample size, and may indicate good overall model fit even when one or more of the parameters in the model is poorly determined and fail to provide information regarding what is wrong with the model (Fraas & Newman, 1994).

Incremental Fit Indices

In contrast to absolute fit indices, incremental fit indices measure the improvement in fit by comparing a target model with a more restricted, nested baseline model (Hu & Bentler, 1999). Most typically, a null model, one in which there are no correlations between the observed variables is considered (Bentler & Bonett, 1980). Examples of incremental fit indices include the Normed Fit Index (NFI, Bentler & Bonett, 1980), Bollen's fit index (BL86, 1986), the Tucker-Lewis Index (TLI, 1973), another index developed by Bollen (BL89, 1989) Bentler's (1989, 1990) and McDonald and Marsh's (1990) Relative Noncentrality Index (RNI) and Bentler's Comparative Fit Index (CFI). The formulas for some of the aforementioned indices are provided in Hu and Bentler (1999).

Hu and Bentler (1999) claim that there are two pressing issues that must be considered in the proper application of fit indices for model evaluation. These issues are important considerations for applied researchers and methodologists. The first important issue concerns the behavior of fit indices under various data and model conditions, including a host of commonly encountered situations in general practice. These conditions include "sensitivity of fit index to model misspecification, small sample bias, estimation effects, effects of violations of normality and independence, and bias of fit indexes resulting from model complexity" (p. 4). The second issue involves the judicious applica
of thumb". As with many rules of thumb, little consensus exists with respect to conventional cut off criteria, and often recommendations are diverse and or inconsistent. In light of the lack of empirical evidence, questions remain with

respect to the adequacy of these conventionally advanced cutoffs. A recent examination of this issue (Hu & Bentler, 1999) revealed that for some of the recommended fit indices, the cutoff criterion was evidenced to be greater (or smaller) than conventional rules of thumb required for model evaluation or selection.

Binomial Index of Model Fit

In contrast to some of the more traditional indices, Fraas and Newman (1994) proposed a binomial test of model fit as an alternative method for determining the goodness of fit of a proposed theoretical model. This method, employing an index referred to as the binomial index of model fit value, is based on the application of the binomial distribution to the number of paths in a model that are supported by the data. This approach requires that an event be classified into one of two categories according to certain criteria, that is, the determination of whether the data provide support for a given path. This determination can be made in a number of ways. For example, the decision for support can be based on (a) the parameter estimate for a path exceeds an a priori effect size, (b) the parameter estimate is statistically significant, and (c) the parameter estimate reflects a hypothesized algebraic sign or any combination of these. After criteria have been established to determine whether a given path is supported by the data, the second step involves testing the actual number of paths supported by the data. Using a binomial test, the probability of obtaining at least the number of paths supported by the data is calculated. If the calculated

probability is less than the alpha level (i.e., less than would be expected to occur by chance), the conclusion drawn is that the data are supportive of the model.

These authors contend that the use of a binomial test to estimate how well data support a theoretical model differs conceptually from other goodness-of-fit approaches. They purport that this method is better described as the degree to which the paths support the nomological net of a theory, rather than being based on the reproduction of a variance-covariance matrix.

Concerns and criticisms surrounding this approach include the lack of independence between events, the effect of sample size when employing statistical significance as a criterion, and the limited capacity of the index to provide insight regarding path misspecification. However, the most salient problem for this line of inquiry is the differential application of criteria in determining support for a given path. Freedom to adjust this criteria will likely result in contradictory conclusions regarding model fit, leading to inconsistency across studies and thus failing to provide a standardized estimate of the precision of prediction.

Theoretical and Empirical Fit

In an investigation designed to compare the performance of different maximum likelihood and generalized least squares estimation techniques, Olsson, Troye, and Howell (1999) examined both measures of theoretic fit and empirical fit. According to these authors, in research the goal is often to construct models that reflect the structures and parameters of some

unobservable causal mechanism. "The degree of isomorphism between such a theoretic model and a "true" model can be labeled "theoretic fit" (p. 31). Alternatively, and most commonly, measures of empirical fit are employed because they serve as the only available evidence of the adequacy of the theoretical structure and accuracy of the parameter estimates and hence provide indirect support for a theory. However, in a Monte Carlo study, when the true population parameter values are known, the discrepancy between the true values and the estimated parameter values can be calculated. For example, in a simulation study we can construct theoretical models that reflect the structures (M_{true}) and parameters (P_{true}) of some unobservable true model of the underlying causal mechanisms assumed to generate the empirical observations—to achieve theoretical fit. However, in realistic settings, M_{true} and P_{true} are unknown—and there is no direct evidence of theoretical fit. Therefore, researchers make use of indicants of the theoretic model's ability to account for the structures of the data employing indices of overall-fit (Chi-square, RMSEA, etc.), in addition to significance tests of the parameters. "If the goodness of fit is adequate, it is considered as evidence for the plausibility of the model; that is the theoretic model M_{theory} may accurately represent M_{true}. To the extent that M_{theory} is wrong (i.e., the theoretic model), an ideal estimation procedure would provide an accurate estimate of "model error" (Olsson, et al., 1999, p. 34-35).

How persuasive is a Good Fit?

According to Roberts and Pashler (2000), it is a mistake to allow good fits to provide substantive support for a theory. These authors contend that the practice of using good fits to support theories is flawed in several ways, advancing several possible reasons for their continued use. These reasons include, a desire to imitate physics, the presence of confirmation bias, theory complexity, neglect of basic principles and a popularity based at least partly on repetition and inertia. "A good fit reveals nothing about the flexibility of the theory (how much it cannot fit), the variability of the data (how firmly the data rule out what the theory cannot fit), or the likelihood of other outcomes (perhaps the theory could have fit any plausible result)" (p. 358). In order to determine how much "the fit" should increase our belief in a proposed theory one must employ all three of the aforementioned pieces of information. Showing that a theory fits data is not only not enough, it is nearly meaningless. These authors also contend that it is necessary to compare plausible alternative outcomes with what the tested theory could explain through an examination of both the flexibility of the tested theory and the variability of the actual results. Further, the resultant evidence will not be very convincing if either is large compared to the range of plausible outcomes (Roberts & Pashler, 2000).

Meehl's C_i

Meehl (1997, 1990a) has proposed an index of corroboration (C_i) that may provide a standardized means of expressing the extent to which empirical research supports or contradicts a theory:

$$C_i = (Cl)(In)$$

where *Cl* = the "closeness" of the data to the theoretical prediction (verisimilitude or truth-likeness), and

In = the "intolerance" of the theory (e.g., a standardized precision of prediction).

These terms are further explicated as follows:

$$Cl = 1 - (D/S)$$

where *D* = deviation of observed data from the tolerance interval of the theory

S = Spielraum (the range of data values that are expected whether or not the theory is true)

$$In = 1 - (I/S)$$

where *I* = the interval tolerated by the theory (e.g., the raw precision of prediction).

An Example

To build on the simple example presented in Chapter One, (i.e., a prediction about population mean differences), let us now consider a theory that

posits a relationship between two variables. Recall that large values of C_i should result from strong theories making tight predictions in which data are very similar to predicted values. Let's suppose a researcher has made a prediction that a positive correlation exists between attitude toward computers and integration of computers in the classroom. This prediction is somewhat stronger than a simple prediction that a correlation exists, because a directional relationship is predicted. However, the prediction is less precise than a prediction that posits a positive relationship between .5 and 1.0. Further, knowing that the plausible values of a correlation range from -1.0 to +1.0, whether or not the theory is true, the Spielraum (S) is thus 2.

In this example, the simple directional prediction of a positive correlation between attitude and integration suggests a tolerance interval of 1.0 (any correlation greater than zero is consistent with this "flabby" prediction) and an intolerance (In) of 1 – 1/2 or 0.50. If the sample correlation between attitude and integration is found to be .50, the data do not deviate from the prediction (Cl = 1.0) and Meehl's C_i = (Cl)(In) = (1.0)(.50) = .50. If the prediction was not simply a positive correlation but a positive correlation between .5 and 1.0, then the tolerance interval is .5 and In = 1 – .5/2 or .75. The same observed data (a correlation of .50) are also consistent with this prediction, but C_i = (1.0)(.75) = .75. The latter theory receives more corroboration from the data because it made a "riskier" prediction that was consistent with the observations.

Suppose the observed data evidenced a correlation of -.5, indicating an inverse relationship between attitude and integration. Such data are not

consistent with the predictions of either theory. For the theory providing a directional prediction only, the data deviate (D) from the lower bound of the tolerance interval by .5 and Cl = 1 – D/S = .75. These data provide a corroboration index value of (Cl)(In) = (.75)(.50) = .375. For the "riskier" prediction of a positive correlation between .50 and 1.0 the data deviate by 1.0 and Cl = 1 – D/S = .50. For this theory, the data also provide a corroboration index value of (Cl)(In) = (.50)(.75) = .375. Although the observed data deviate to a greater extent from the prediction of the latter theory, the corroboration is the same, in this particular case, because the prediction was more precise.

Table 1 and Figure 3 present the values of C_i that would be realized for the values of sample correlation under four levels of precision of prediction. Note that as predictions become more accurate (the observed correlation is closer to the predicted correlation), higher values of C_i are obtained with more precise predictions. When the prediction is far from the observed value, higher values were observed from looser predictions. Further, the intolerance of the theoretical prediction presents an upper bound for C_i (i.e., precision of prediction limits the degree of corroboration regardless of the magnitude of the observed correlation).

Table 1. *Values of C_i under Four Levels of Precision*

Prediction	S	I	In	r	C_i
$r < -.10$ *or* $r > .10$	2	1.8	.10	.5	.10
$r > 0$	2	1	.5	.5	.50
$.50 < r =< 1.00$	2	.5	.75	.5	.75
$.50 < r < .70$	2	.2	.9	.5	.90

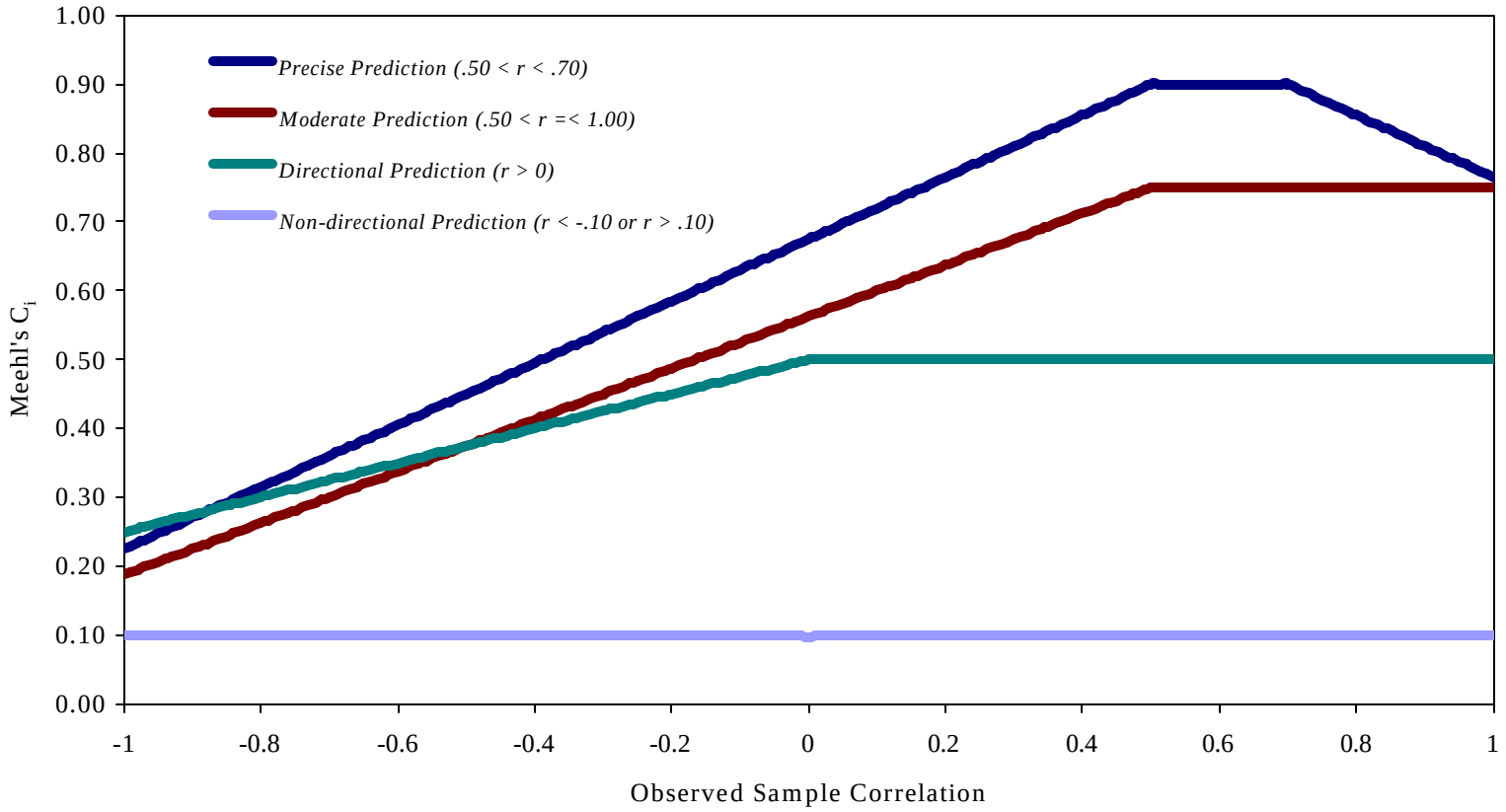

Figure 3. Predicted Values of C_i under Varying Levels of Precision.

Amalgam

Recall, from the previous chapter, that movement from theory into an empirical test necessitates the incorporation of many logical components besides the theory itself. That which is subject to empirical test is not the theory alone, but the amalgam of these elements. Meehl (1997, 1990a) presents these components as elements of an equation:

$$\left(T \bullet A_x \bullet C_p \bullet A_i \bullet C_n\right) \rightarrow \left(O_1 \supset O_2\right)$$

Where T = the theory being "tested,"

A_x = Auxiliary theories relied upon during the conduct of the research.

C_p = *Ceteris paribus* (all other things being equal),

A_i = Instrumental theories related to measures and controls employed,

C_n = Realized particulars (the extent to which the research was actually conducted as we think it was), and

$O_1 \supset O_2$ = the material conditional "if you observe O_1, you will observe O_2."

Auxiliary theories (A_x) lie at the periphery of the theory being tested and are somewhat distinct from the "hard core" concepts or postulates of the theory under investigation. Although central portions of a particular theory may not be rigorously defined, there will likely exist key critical components as well as non-central elements. These tangential components (although not central to the theory being explored) are still, in fact, a part of the theory.

For example, in an investigation of the relationship between nutrition and anxiety in which anxiety is measured using responses to Likert-type items written in English, the use of participants whose primary language is not English necessitates an auxiliary theory that the inferences from the scores derived from the anxiety instrument retain their validity in such a population. If data obtained from such research fail to support theoretical predictions, the failure may be attributable to the core theory being incorrect or simply that the auxiliary theory did not hold.

The concept of verisimilitude (truth-likeness) is closely related to this core-peripheral distinction. Meehl (1990a) suggests that a theory that is false in its core postulates has lower verisimilitude than one that, while correct in its core concepts, is incorrect in several of its peripheral ones. As even the best theories are likely to be approximations of the true state of reality, verisimilitude then, refers to the relationship between the theory and the real world.

Ceteris paribus does not mean that all factors not mentioned are equal for all participants, but rather that there are no systematic factors left unmentioned. This clause amounts to a very strong and highly improbable negative assertion that "nothing else is at work except factors that are totally random and therefore subject to being dealt with by our statistical methods" (Meehl, 1990aa, p. 111).

The instrumental auxiliary theories (A_I) are related to measures and controls employed by the researcher. These are distinguished from A_X in that they do not contain any psychological constructs. Thus, if anxiety is measured by changes in

galvanic skin response rather than by a Likert instrument, the auxiliary theory at work is within A_I rather than A_x.

The realized particulars (C_n) represent the extent to which the research was actually conducted as we think it was. This element of the amalgam represents treatment integrity. For example, if we plan to manipulate participant nutritional status to examine it's relationship with anxiety, but the participants do not adhere to their dietary "treatment," then the variable actually applied in the research is not what we think it is. Data that contradict our theory may arise because of this perturbation in C_n.

Reaction to Meehl's C_i

As might be expected, the mere mention of an index of corroboration, or an attempt to quantify meta-theory, inspired considerable debate. Campbell (1990) contended that verisimilitude needs to be considered in the context of a pattern of predictions, one that can be matched to a pattern referred to as the "s-c-facts". The s-c-facts represent the "focal-collective scientific-consensual "facts", allowing for the connotation "so-called facts" (Campbell, 1990, p. 144). The so-called facts can arise from earlier tests of theories, or from theoretical sources such as exploratory experimentation or refined folk observations. In Campbell's view, the incorporation of the so-called facts results in reducing the exaggerated role of theory. Competing theories would thus be compared based on their goodness of fit to the shared s-c-facts they provide predictions for. Campbell (1990) suggested a simple correlation might be provided as evidence of

verisimilitude. It would seem that comparing correlation coefficients (a standardized measure) would indeed be similar to comparing the component of C_i that represents closeness (*Cl*), as this estimate represents a standardized measure of verisimilitude.

Chow (1990) asserted that Meehl's argument (i.e., theory appraisal based on numerical predictions in various situations) is only appropriate "when a theory is being tested with non-experimental methods or in an *ex post facto* manner" (p. 147). He further argues that given the inherent difference between experimental and non-experimental studies, the proposed corroboration index would not be appropriate when a theory is being tested experimentally.

In agreement with Meehl, Humphreys (1990) contended that the target article not be restricted to courses and seminars on psychological theory, but rather it should be required reading for every graduate course in quantitative methods. Additionally, this researcher asserted that substantive advances in psychological research would occur, "if psychologists were to plan their research, analyze their data, and discuss their findings in congruence with the current target article" (p. 155).

In response to the aforementioned commentary, Meehl (1990b) addressed each of the commentators and advanced a more focused discussion of the corroboration index and verisimilitude. In general, these comments served to clarify certain references and specific claims, and to underscore the intended purpose of the corroboration index. In particular, Meehl (1990b) noted the nearly wholesale lack of enthusiasm for the proposed index despite agreement with the

critical aspects of his overall position. In a concluding remark, Meehl (1990b) offered the following notion, “In employing any useful numerification of an open concept in the social sciences, one is properly alert to the caveats, but not frightened into cognitive paralysis by them” (p.177). Clearly, disapproving responses were likely anticipated, as “we know from the history of science that radically novel ideas regularly meet with resistance, and statisticising metatheory is certainly a new – and radical – idea” (p. 177).

Past Research on C_i

An initial examination of the utility of the index and its behavior in theory testing was conducted in the context of a simple theory, the core of which predicted a difference in means between two groups (Hogarty & Kromrey, 2000). This effort was aimed toward illuminating the relationship between the closeness of the observed data or verisimilitude and the precision of prediction. The relationships explored in this study included factors related to the nature of the theory being tested (i. e., predicted mean difference between groups, the raw tolerance interval of the theory, and the Spielraum), the degree of correspondence of the theory to the actual populations simulated (i.e., population difference in means and variance ratios between the two populations), and research design factors (i.e., sample size, reliability of the dependent variable, and the confounding effect of an extraneous variable). An important limitation of this research, however, was that the investigation considered only relationships

for the most basic of predictions, that is, predictions about population mean differences.

Under these very limited circumstances, the mean index of corroboration was seemingly unaffected by sample size, and notably more influenced by the level of verisimilitude and the level of intolerance specified by the theory. In addition, the reliability of the dependent measure was shown to have but a slight influence on the mean C_i, and only when predictions were very close to truth. Although sample size and measurement reliability were not important determinants of mean C_i, both factors were related to the variability of this statistic, with larger samples and more reliable measures providing greater stability across samples. Although such sampling variability is important, one would anticipate that the degree of support for a theoretical prediction that was tested with a large sample should be greater than that provided by a small sample. This finding clearly illuminated the need for additional work aimed at incorporating a sample size component into an index such as C_i.

A second study conducted by Hogarty and Kromrey (2001) was designed to investigate the relationship between theoretical predictions and empirical results through a consideration of Meehl's index of corroboration in the context of hypothetical theories that made relatively simple predictions (magnitude of a zero-order correlation) and those that made more statistically complex predictions (magnitude of a first-order partial correlation). This investigation served to advance knowledge about the behavior of Meehl's index of corroboration beyond the most basic theoretical predictions of differences in

population means. The relationships explored in this study included factors related to the nature of the theory being tested (i.e., predicted magnitude and direction of correlation, the raw tolerance interval of the theory, and the type of correlation), the degree of correspondence of the theory to the actual populations simulated (i.e., true population correlation, both zero and partial correlation, and the magnitude of correlation between the two focal variables and the variable being partialed) and research design factors (i.e., sample size and reliability).

Surprisingly, nearly identical values of the statistic were obtained for both types of prediction across the various levels of the design factors that were employed. As with the evaluation of mean differences, the major influence on C_i was the precision of the prediction. This factor far outweighed the impact of closeness (verisimilitude), with theories that made tight predictions obtaining notably higher values of C_i than those making loose predictions even with extreme differences between the prediction and the true population parameter. Verisimilitude was less influential in determining C_i than that observed in the assessment of mean differences (Hogarty & Kromrey, 2000). In addition to the building evidence regarding the influence of the precision of prediction and verisimilitude, insight was gained about the impact of measurement reliability when employing zero-order and first-order partial correlations. In our earlier investigation, less reliable measures evidenced slightly smaller values of C_i . In this study, the influence of measurement reliability was found to depend on the relationship between the true population correlation and the prediction. When a theory's prediction was precisely correct, the largest mean value of Meehl's C_i

results from using measures with the highest reliability, with progressively smaller values resulting from the use of successively less reliable measures. Similarly, if the theory predicted a correlation greater than the true value, more reliable measures produced larger values of Meehl's C_i with the difference in values becoming somewhat greater as the verisimilitude decreases (i.e., a greater difference between the prediction and the reality). However, when the predicted correlation was less than the true correlation, less reliable measures provided larger values of Meehl's C_i with the difference increasing as the predicted value approaches zero. The observed result, that theories with lower verisimilitude may obtain greater corroboration than theories with higher verisimilitude, if the measurement of the relevant variables is not reliable is a function of the attenuation of the sample correlation (Pedhazur, 1997).

These results suggest that caution is needed regarding the interpretation of the magnitude of C_i without regard to the context of the application. Although once again sample size was not deemed an important determinant of mean C_i, it was seen to influence the variability of this statistic. Similar to the results obtained by Hogarty and Kromrey (2000) in the investigation of prediction of mean differences, larger samples evidenced less variability in C_i across samples. Again, this suggests additional efforts should be aimed at incorporating a sample size component into an index such as C_i.

Despite the obvious need to reduce the emphasis upon statistical significance and null hypothesis testing, sample size requirements remain important considerations in the interpretation of research evidence. Therefore,

the most recent work by Hogarty and Kromrey (2002) included a sample size adjustment to the calculation of C_i. For this study, the sample size requirement was conceptualized as the smallest sample size that a researcher(s) would no longer be substantively concerned with sampling error (the smallest size at which sampling error may be considered trivial). This was considered the *"fail-safe N."* In this context, a weight for C_i was computed as the square root of the ratio of the study's sample size to the "*fail-safe N*." That is,

$$Weight = Relative\ Size = \sqrt{\frac{N_{study}}{N_{failsafe}}}$$

Incorporating this weight in Meehl's C_i provides the Weighted C_i

$$Weighted\ C_i = (Cl)(In)(RS) = \left(1-\frac{D}{S}\right)\left(1-\frac{I}{S}\right)\left(\sqrt{\frac{N_{study}}{N_{failsafe}}}\right)$$

Through the incorporation of a sample size component into Meehl's index of corroboration, a statistic that more closely approximated the desired behavior was suggested.

In this study, six factors were manipulated. First, three factors related to the theory being tested were included. The predicted mean difference between groups was examined at five levels, the raw tolerance interval of the theory was examined at four levels, and the Spielraum was examined at three levels. These values of raw tolerance and Spielraum yield intolerance (*In*) values ranging from 0.50 (the value of intolerance for a simple directional prediction of effects) to 0.98 (reflecting a tight, risky prediction). Second, two factors related to the true populations simulated were manipulated. The population difference in means

was examined at five levels, and variance ratios between the two populations were manipulated at four levels. These population mean differences, crossed with the theory's predictions provided conditions ranging from those in which the theory's prediction exactly represented the true populations (perfect verisimilitude), to those in which the theory deviated from the true population conditions by effect sizes as large as two standard deviations. Finally, the sample size of each study, a characteristic of research design, was investigated at four levels.

Once again, these findings shed light upon the relationships of these components in the context of only the most basic of predictions, that is, predictions about population mean differences. Under these very limited circumstances, the Weighted C_i index of corroboration was profoundly affected by sample size, only slightly influenced by the level of verisimilitude, and severely limited by the level of intolerance. These findings suggested that the major influence on C_i was the precision of the predictions. This factor far outweighed the impact of closeness (verisimilitude), with theories that make tight predictions obtaining notably higher values of C_i than those making loose predictions, even when the predictions were substantially wrong. As anticipated, the Weighted C_i (in contrast to Meehl's original formulation of C_i) provides a greater degree of support for a theoretical prediction that was tested with a large sample than that provided by a small sample.

The importance of theoretical intolerance as a determinant of degree of corroboration highlights the need for the development of precise theories in the

social sciences. Additionally, the results of the analyses focusing on more traditional approaches to theory appraisal underscore the need to extend our thinking beyond the common “reject” or “fail to reject” decisions resulting from tests of null hypotheses that are derived from aspects of theory. Jointly, these results suggest that efforts to develop theories in the social sciences that enjoy greater precision of prediction may concomitantly provide critical tests with greater potential for corroboration.

A theory’s merit is a matter of degree, rather than a yes or no question, as it is treated in null hypothesis testing (Meehl, 1990aa). A natural extension of this previous line of research should involve the examination of these relationships when making more complex predictions from theories. An extension of the components of C_i to multivariable problems, such as encountered in path analysis, is worthy of investigation.

Importance of the Study

The use of path analysis in the appraisal of theories was most recently debated among the pages of a topical issue of *Psychological Methods* (2002). In fact, the entire issue was devoted to a conversation regarding theory appraisal, causal models, tests of statistical significance, empirical fit and verisimilitude (see for example, Markus, 2002; MacCallum, et al., 2002; Meehl & Waller, 2002; Mulaik, 2002; Reichardt, 2002; Waller & Meehl, 2002). This present study is designed to contribute to this conversation, by building upon the previous research conducted in the context of a simple theory through an exploration of

the utility and behavior of the corroboration index when testing more complex predictions, such as applications of path analysis. It is anticipated that the results will support the incorporation of Meehl's C_i in the planning of empirical studies as well as the interpretation of research results. It is also important to note that the intention is not to encourage abandonment of other supplementary approaches or tools or to use the corroboration index in isolation, rather the index is advanced with the understanding that it be employed in an auxiliary or complementary role. It is hoped that the use of a corroboration index may help in reducing the "hypnotic fascination" with null hypothesis significance testing (Meehl, 1990aa). Its use should serve to move the arguments surrounding theory testing away from the testing of null hypotheses into a consideration of the complexity of the research context, the degree of "risk" entailed by the theory's predictions, and the extent to which the obtained data (absent the theory) represent a "damn strange coincidence."

The index of corroboration is unique in that it combines both a measure of the closeness (or verisimilitude) of the data and the precision with which the prediction is made. Additionally, unlike some of the other indices that are typically employed in the conduct of research, the index of corroboration is not context bound or discipline specific. In this vein, it might be viewed as behaving like an effect size that is computed differently given different circumstances or situations. The univariate corroboration index is available if that is the type of measure that is appropriate (i.e., in testing population mean differences). This multivariate extension expands the utility of the index to the next logical level, by

exploring the versatility of the index beyond the limited applications previously examined.

Chapter Three

Method

Organization

The purpose of this chapter is to elucidate the method for this study. The chapter opens with a restatement of the purpose of the study and the research questions and hypotheses. A brief overview of the utility of Monte Carlo studies follows. A description of common applications and uses of simulation methods is then presented. After the efficacy of this approach has been established the research design is described. Illustrations of the models under consideration and population correlation matrices are included to demonstrate two of the central design factors, that is, model complexity and collinearity. The justification for the multivariate extension of the corroboration index is advanced and supported by results from a series of data simulations. The conduct of the Monte Carlo study is then explained through an illustration of the data generation strategy. The chapter concludes with a discussion of the interpretational framework that guides the reporting of the results.

Purpose

Meehl (1997, 1990a) has presented a logically sound index of corroboration to summarize the extent to which empirical tests of theories provide support or contradiction of those theories. However, the numerical properties of this index have not been investigated beyond some of the most basic predictions about population mean differences, zero order correlations and first-order partial

correlations. This study is the next logical step, providing an evaluation of the utility of the index and its behavior in the testing of theories employing path analysis in the context of social science research.

Four Research Questions and Three Research Hypotheses

Research Questions

1. What is the relationship between mean C_i and the main effects examined in the study (i.e., verisimilitude, intolerance, model complexity, collinearity, and sample size)?
2. What is the relationship between the standard deviation of C_i and model complexity, collinearity, and sample size?
3. To what extent is the relationship between mean C_i and the precision of prediction (i.e., intolerance) influenced by the complexity of the model (i.e., the number of variables in the model)?
4. To what extent is the relationship between mean C_i and the precision of prediction (i.e., intolerance) influenced by the level of collinearity?

Research Hypotheses

1. The relationship between mean C_i and the precision of prediction (i.e., intolerance will be slightly influenced by the closeness of the data to the theory (verisimilitude).
2. The relationship between mean C_i and the precision of prediction (i.e., intolerance) will not be substantively influenced by sample size.

3. The relationship between mean C_i and precision of prediction will be substantively stronger than the relationship between mean C_i and verisimilitude, model complexity, collinearity, and sample size.

Monte Carlo Studies

The behavior of Meehl's C_i was evaluated using Monte Carlo methods. A series of simulations were conducted that related theoretical predictions to empirical results. The use of simulation methods allows the control and manipulation of research design factors and the incorporation of sampling error into the analyses. The study was designed in the context of hypothetical theories, the cores of which predict a single outcome from various configurations of exogenous and endogenous variables. The resulting path coefficients were the parameter estimates of primary interest.

The utility of Monte Carlo studies is derived, in large part, from their ability to evaluate the properties of statistical procedures and help researchers select appropriate analytical procedures under varying design conditions. Monte Carlo studies have been employed to investigate the behavior of a variety of parameter estimates of interest to researchers, as well as the Type I and Type II error rates of statistical tests, coverage probabilities of confidence intervals, the bias and variability of IRT item parameter estimates, factor loadings, path coefficients, and goodness-of-fit indices (Serlin, 2000). In this type of study, the conditions that researchers are likely to encounter in the conduct of applied research are manipulated and the properties of the estimates are examined under each of the

varied scenarios. The results serve to inform researchers of reasonable approaches and proper cautions to exert when particular conditions are confronted.

Research Design

The choice of characteristics of the sampled populations (or factors) in Monte Carlo studies is typically determined by examining conditions that are likely to be encountered by researchers working in applied settings. Five factors were manipulated in these simulations: factors related to the theory being tested, the degree of correspondence of the theory to the actual populations simulated and research design factors. The two factors related to the theory being tested were the number of variables in the model and the size of the tolerance interval or level of intolerance. The number of variables in the model was examined at three levels, the simplest model containing four variables, a more sophisticated model with six variables, and the most complex model containing eight variables. The set of 'true' models is fully illustrated in Figures 4-9.

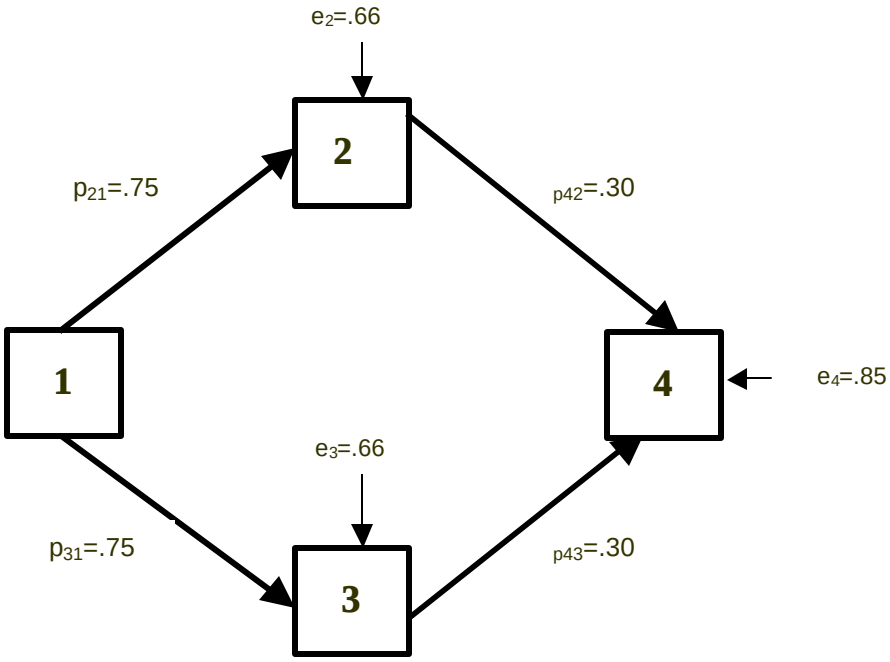

Figure 4. Truth for Four Variable Model, Low Collinearity.

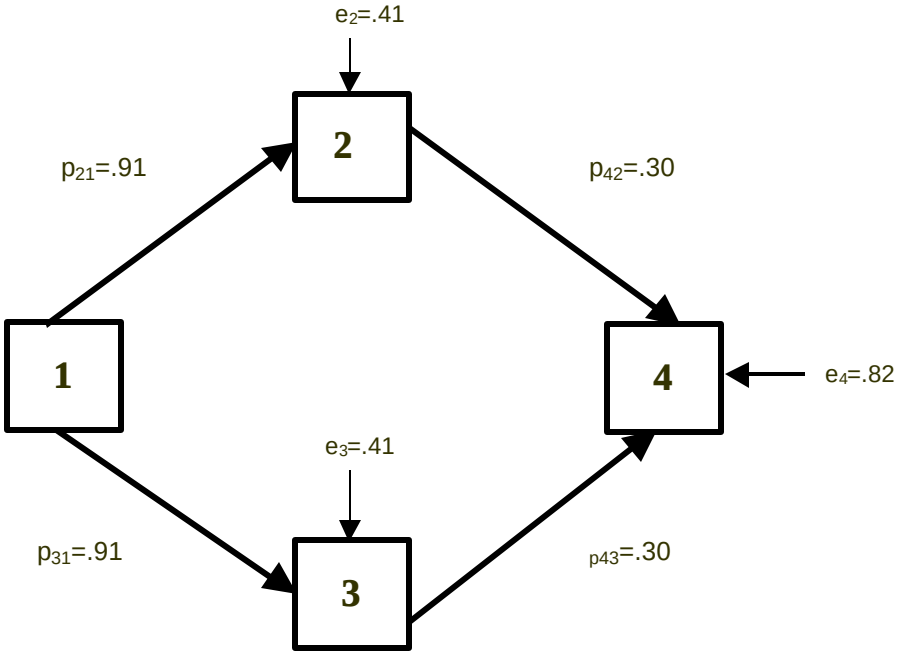

Figure 5. Truth for Four Variable Model, Moderate Collinearity.

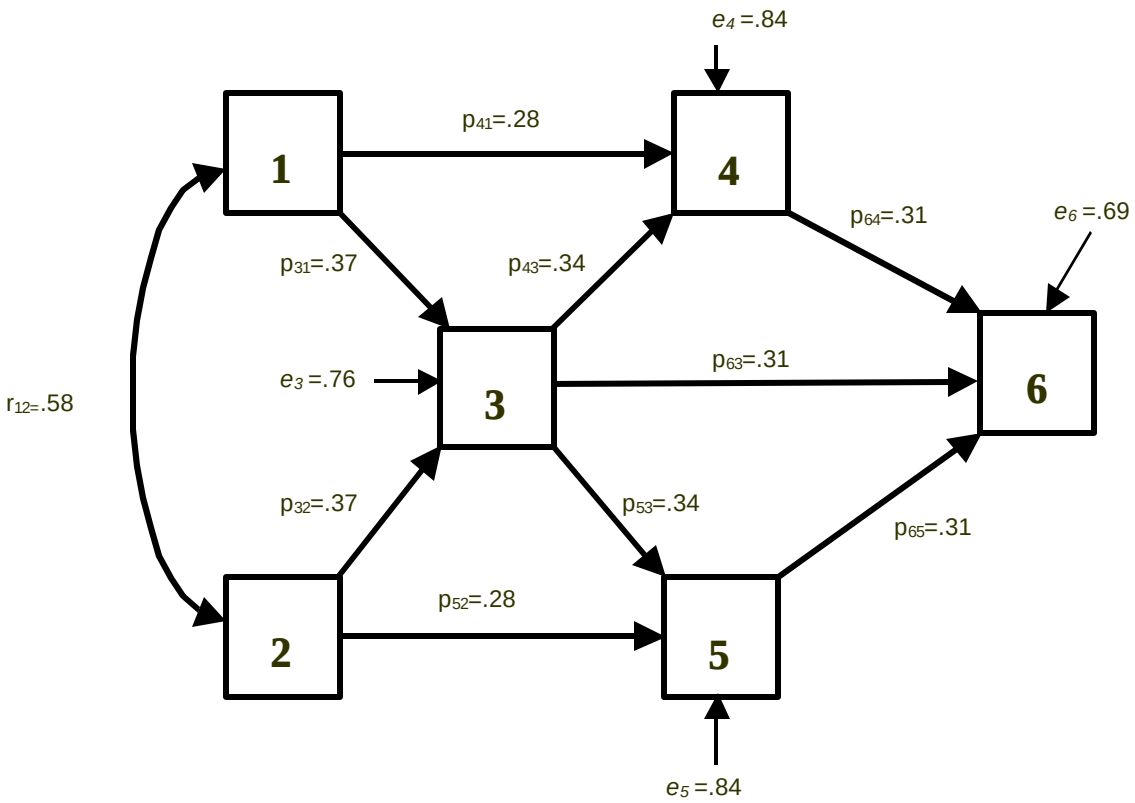

Figure 6. Truth for Six Variable Model, Low Collinearity.

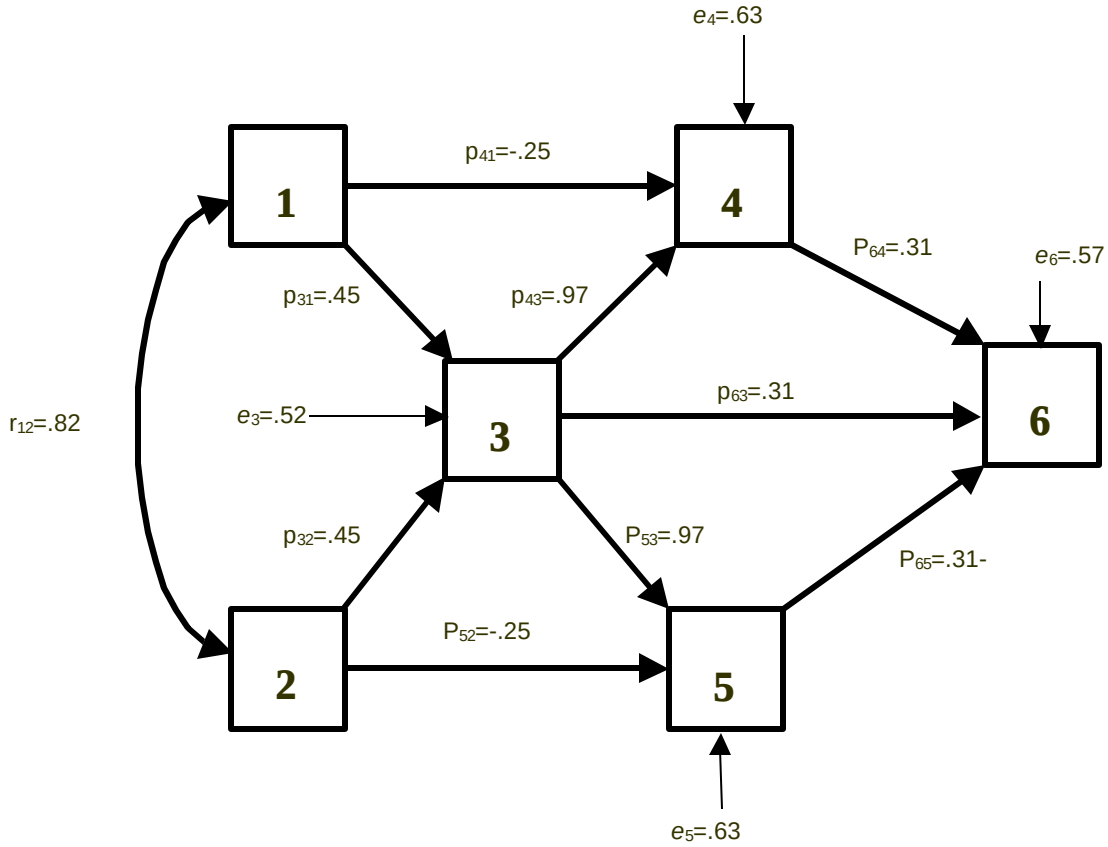

Figure 7. Truth for Six Variable Model, Moderate Collinearity.

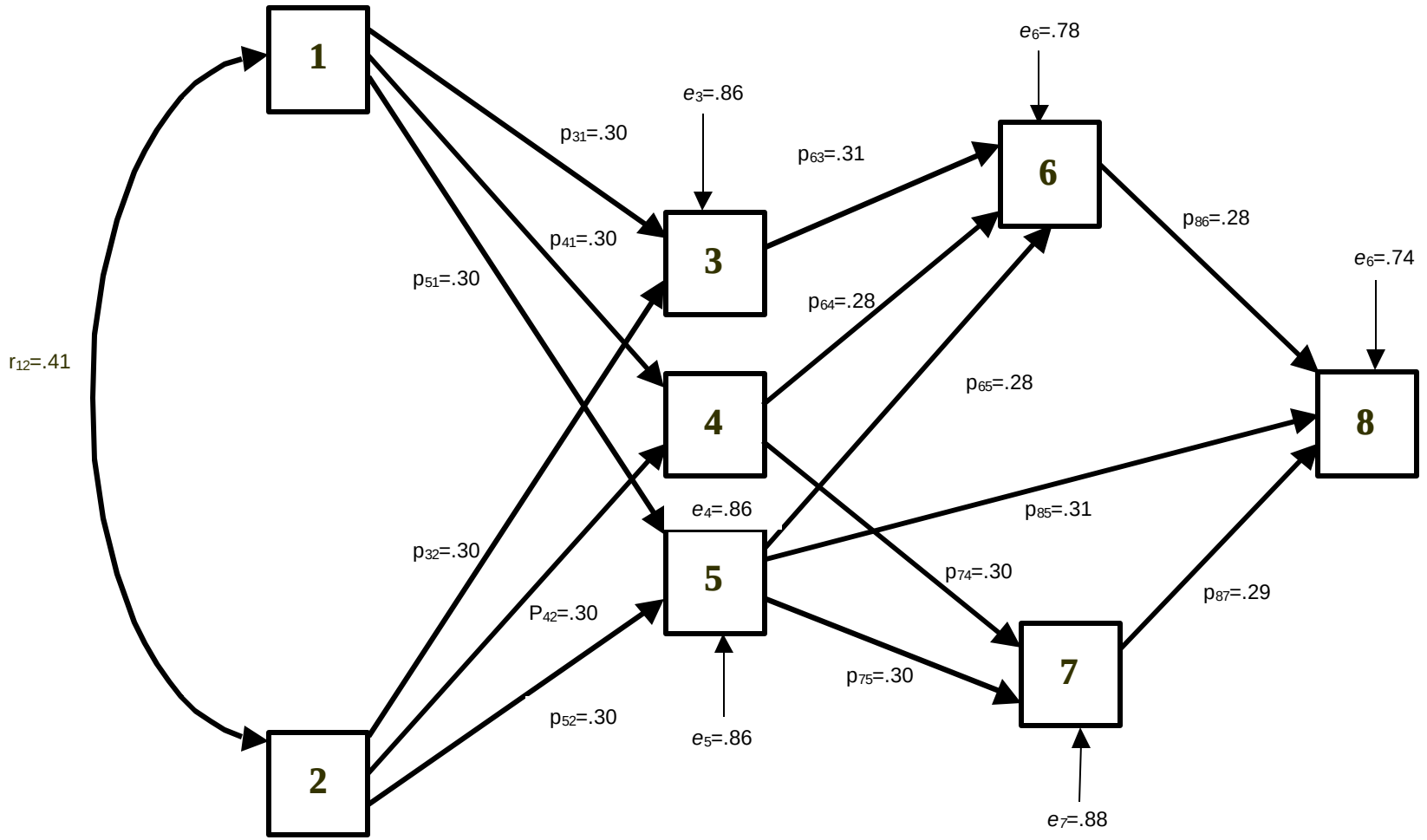

Figure 8. Truth for Eight Variable Model, Low Collinearity.

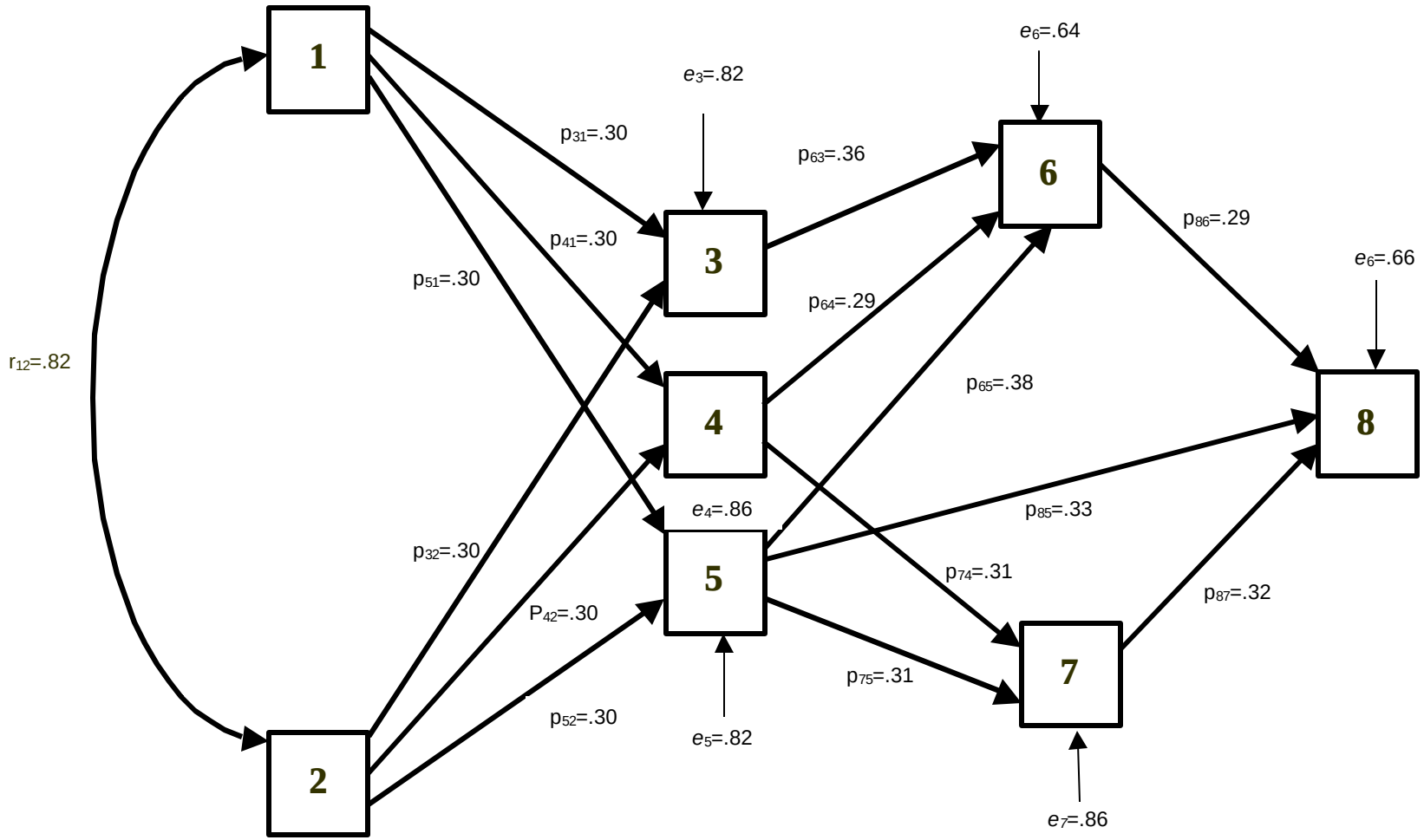

Figure 9. Truth for Eight Variable Model, Moderate Collinearity.

The complexity of the analysis is reflected, in large part, by the number of variables under investigation. For example, an examination of a model with eight variables involves exploring $\left(\frac{k*(k-1)}{2}\right)$ or 28 possible pairwise correlations among the variables. The Spielraum for each standardized path coefficient in all models was necessarily fixed, ranging from –1.0 to +1.0. The raw tolerance interval of the theory was examined at three levels of precision: a non null condition with $\beta \neq 0$, a directional condition employing ½the Spielraum, with $\beta < 0 \text{ or } \beta > 0$, and an interval prediction equal to ¼ of the Spielraum, that is $.05 < |\beta| < .55$. Therefore, these values of raw tolerance and Spielraum yield intolerance (*In*) values for each coefficient ranging from 0.10 (a non null condition) to 0.50 (the value of intolerance for a simple directional prediction of effect) to 0.75 (reflecting a tighter, riskier prediction).

As even the best theories are likely to be approximations of the true state of reality, several levels of verisimilitude or truth-likeness were also explored. Different levels of verisimilitude may result, in part, from misspecified models where one or more of the “true” paths have been omitted, or when one or more ancillary paths are included. For future reference, these models or conditions are referred to as exclusionary and supplementary, respectively. This phenomenon, the degree of correspondence of the theory to the actual populations simulated, was examined at 3 levels (low, moderate, and high). The levels of verisimilitude were kept comparable across the types of models. For example, for the four variable model, the high verisimilitude condition was constructed to mirror truth.

When examining moderate verisimilitude, one path was deleted, representing 1/6 or approximately 17% of the potential paths in the model. For the lowest level of verisimilitude, two paths were deleted, representing 1/3 or approximately 33% of the total number of potential paths in the model. Figures 10-12 provide an illustration for the six variable exclusionary model. For this six variable model, moderate verisimilitude reflects the deletion or addition of two paths, that is, approximately 13% of the potential paths. Low verisimilitude required the deletion or addition of five paths, again, approximately 33% of the total number of potential paths. Further, the levels of verisimilitude for the six variable supplementary model are illustrated in Figures 13-14. For the most complex model, four paths were added or deleted to represent moderate verisimilitude (14% of the paths), and a total of nine paths were added or deleted for the lowest level of verisimilitude, reflecting approximately 32% of the potential paths.

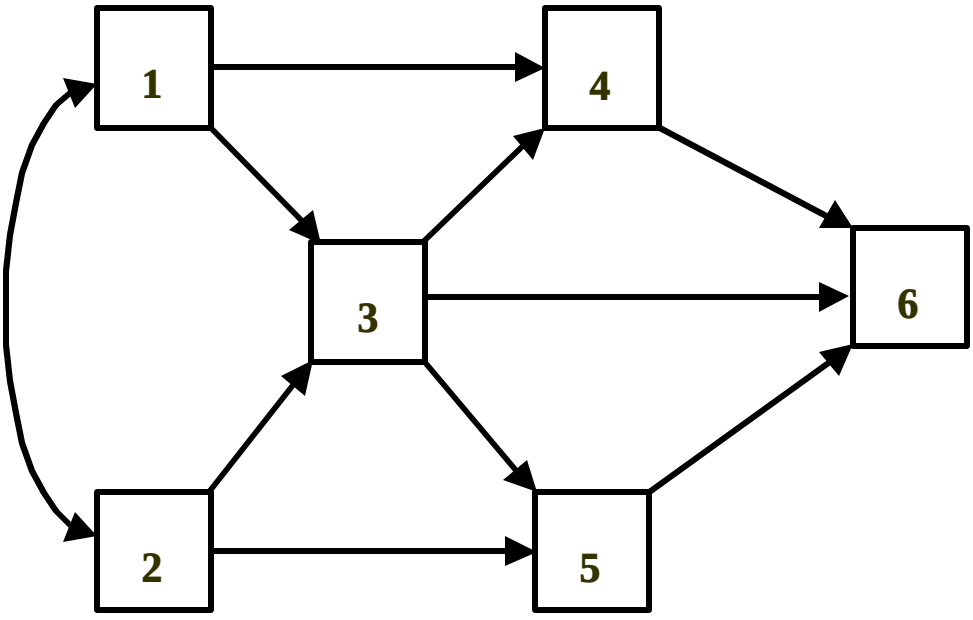

Figure 10. Six Variable Exclusionary Model, High Level of Verisimilitude.

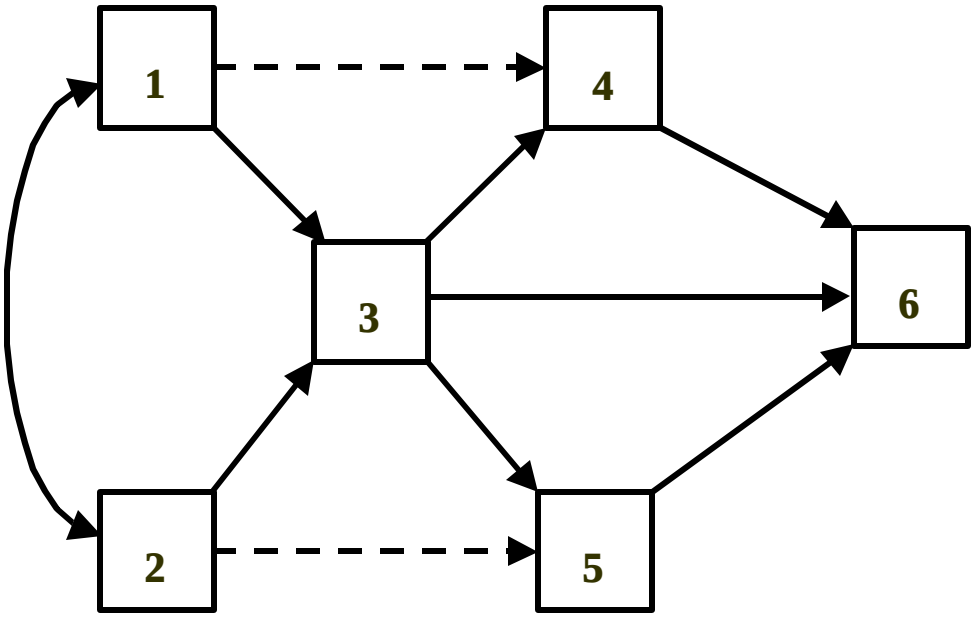

Figure 11. Six Variable Exclusionary Model, Moderate Level of Verisimilitude.

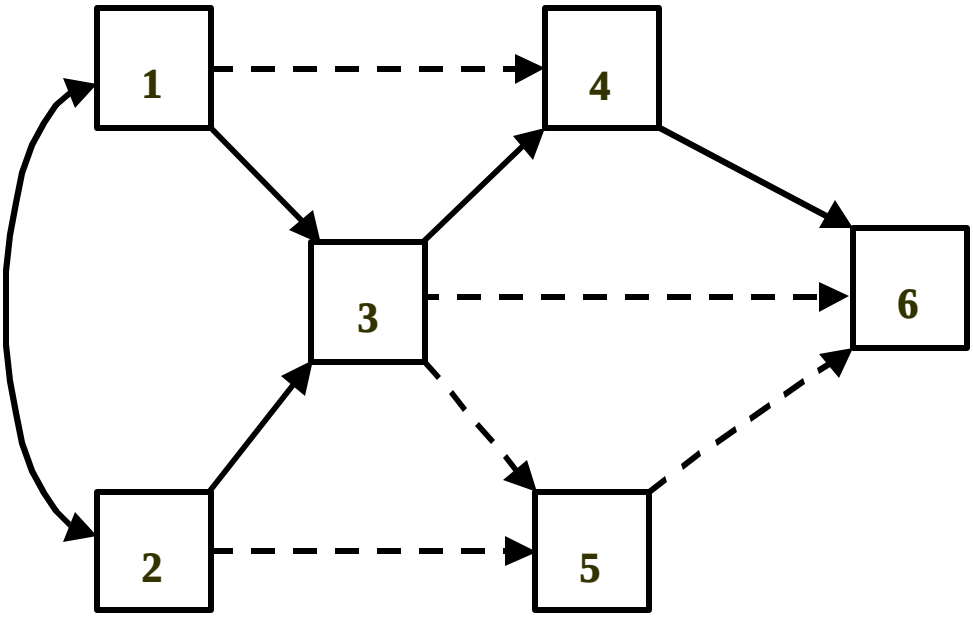

Figure 12. Six Variable Exclusionary Model, Low Level of Verisimilitude.

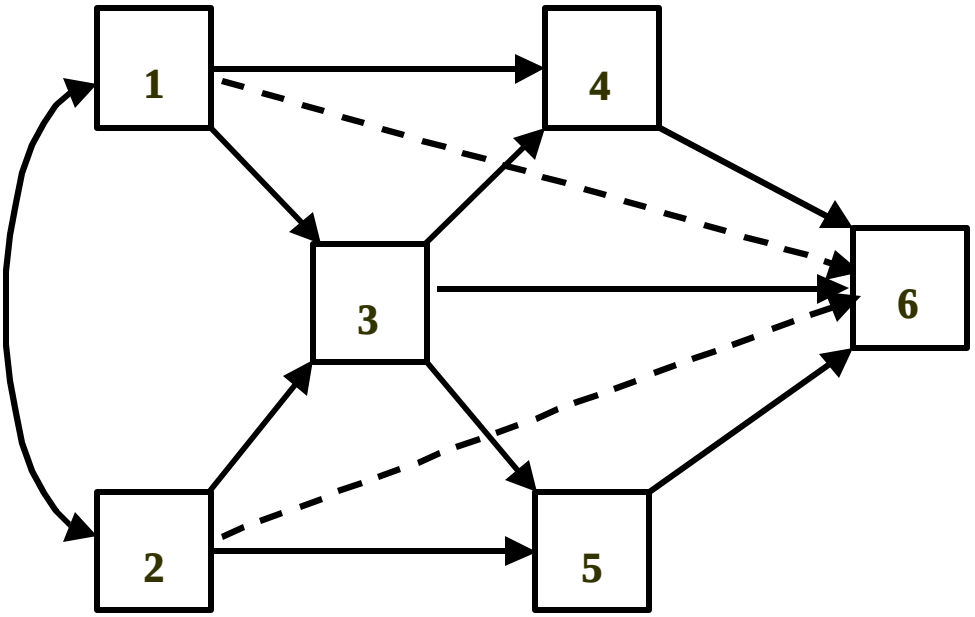

Figure 13. Six Variable Supplementary Model, Moderate Level of Verisimilitude.

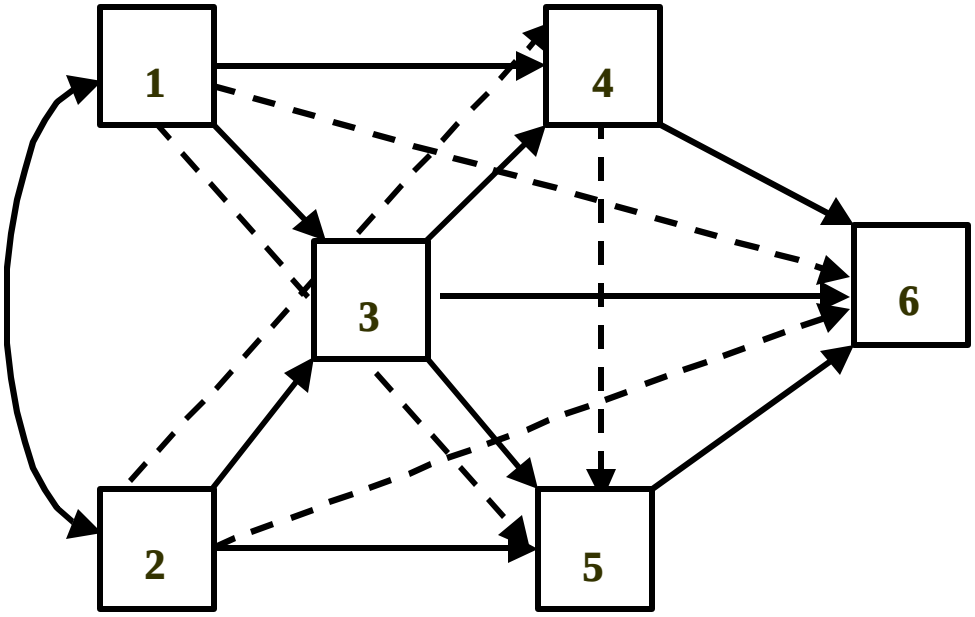

Figure 14. Six Variable Supplementary Model, Low Level of Verisimilitude.

Lastly, two factors related to the design of empirical research were included in the simulations. Sample size was examined at 3 levels (100, 200 and 500 observations) and the correlation between variables was examined at 2 levels. The level of correlation between variables is an important consideration, and when designs involve more than two independent variables it is necessary to look beyond zero-order correlations to diagnose this relationship. In regression analyses this particular issue is referred to as collinearity.

Collinearity may have devastating effects on regression statistics, manifesting in imprecise estimates of regression coefficients (Pedhazur, 1997), and thus is an important consideration given the nature of this investigation. This influence is illustrated by examining the formula for the standard error of a regression coefficient for the case of two independent variables. The standard error for b_1 is given by:

$$s_{b_{y1.2}} = \sqrt{\frac{s^2_{y.12}}{\sum x_1^2 \left(1 - r_{12}^2\right)}}$$

where $s^2_{y.12}$ = variance of estimate; $\sum x_1^2$ =sum of squares of X_1; and r_{12}^2 =squared correlation between independent variables X_1 and X_2. One method commonly employed in the diagnosis of collinearity, focuses on the variance of *b*, which is the square of the formula provided as:

$$s^2_{b_{y1.2}} = \frac{s^2_{y.12}}{\sum x_1^2 \left(1 - r_{12}^2\right)} = \frac{s^2_{y.12}}{\sum x_1^2}\left[\frac{1}{1 - r_{12}^2}\right]$$

In the preceding formula, the term in the brackets is called the variance inflation factor (VIF). This component indicates the inflation of the variance of *b*, resulting from the correlation between the two independent variables. The lower bound of the VIF is one, that is when r_{12}^2=.00. The VIF gets larger (and variance of *b* more inflated) as the correlation between independent variables increases.

Further, when standardized variables are used (i.e., correlations), the following equation illustrates the relationship between the regression coefficients and the correlation matrix:

$$b = R^{-1}r$$

where β is a column vector of standardized coefficients, R^{-1} is the inverse of the correlation matrix of regressors; and r is a column vector of correlations between each independent variable and the dependent variable. The inverted R matrix (R^{-1}) will contain the VIF values along the principal diagonal. For the two variable case, this can be seen as:

$$R^{-1} = \begin{bmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{21}}{1-r_{12}^2} \\ \frac{-r_{21}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{bmatrix}$$

There does not appear to be a single rule of thumb that has been widely accepted with respect to varying levels of VIF, that is, how big is too big? Belsley (1984) contended that the value of 10 is offered frequently, yet without meaningful foundation. This suggests that the VIF needs to be considered with respect to the factors operating within a particular study or context.

The levels of collinearity examined in this study included a low level of collinearity, which would not tend to indicate a deleterious influence with respect to the analyses and results, and a moderate level of collinearity. The levels of VIF in this study were chosen for investigation based upon levels that would likely be encountered in applied research. VIF was set to 1.5 for the low collinearity condition and 3.0 for the moderate collinearity condition. More extreme values were examined but not selected for inclusion due to the likelihood of redundancy among the variables that would not be well suited to this type of statistical analysis.

Multivariate Extension of C_i

Recall that Meehl (1997) initially proposed an index of corroboration (C_i) that provides a standardized means of expressing the extent to which empirical research supports or contradicts a theory:

$$C_i = (CI)(In)$$

where *CI* = the "closeness" of the data to the theoretical prediction (verisimilitude or truth-likeness), and

In = the "intolerance" of the theory (e.g., a standardized precision of prediction).

These terms are further explicated as follows:

$$CI = 1 - (D/S)$$

where *D* = deviation of observed data from the tolerance interval of the theory

S = Spielraum (the range of data values that are expected whether or not the theory is true)

$$In = 1 - (I/S)$$

where *I* = the interval tolerated by the theory (e.g., the raw precision of prediction).

For this study, a multivariate extension of this index was required. The multivariate extension of C_i, investigated in the context of path analyses through the use of multiple regression analysis is defined as follows:

$$Intolerance = 1 - \prod_{j=1}^{J} \frac{I_j}{S_j}$$

$$Closeness = \left[\prod_{j=1}^{J} \left(1 - \frac{D_j}{S_j} \right) \right]^{\frac{1}{J}}$$

where *j* indexes the set of relationships being tested (i.e., I_j and S_j are the tolerance interval and Spielraum for path coefficient *j* and D_j is the distance between the theoretical value and the observed value).

As the proposed corroboration index has not previously been employed in a multivariate context, alternative approaches were explored in order to determine the most effective method to employ. There initially appeared to be two alternative approaches to the composite Meehlian corroboration index, considering a multivariate situation. The first method involves the computation of C_i separately for each path coefficient and then multiplying the obtained values for an overall index:

$$C_i = \prod_{J=1}^{J}\left[\left(1-\frac{D_j}{S_j}\right)\left(1-\frac{I_j}{S_j}\right)\right] \qquad [C_i=1]$$

where the D_i, S_i and I_i are treated for each path independently. An alternative approach that was considered involved the calculation of the product of the distances for each variable and the product of the standardized tolerances. These products would then be subtracted from the value 1.

$$C_i = \left[1-\prod_{J=1}^{J}\left(\frac{D_J}{S_J}\right)\right]\left[1-\prod_{J=1}^{J}\left(\frac{I_J}{S_J}\right)\right] \qquad [C_i=2]$$

where the D_i, S_i and I_i are treated for each variable independently. It was apparent that both of these formulae will necessarily have problems at the extreme values. For example, in the first formula, if $I_1 = S_1$ then $C_i = 0$ regardless of the status of the other variables. In the second formula, if $D_1 = 0$ then CL = 0 regardless of the other variables. Discounting these extreme conditions, however, they appeared worth pursuing. Further consideration of the two approaches revealed that when using the first approach, the obtained index of corroboration was reduced, in most cases, as additional parameters were included in the specified model. Naturally, the second approach was then chosen for this investigation.

The next logical step was to investigate possible representations of the two components of the multivariate C_i, that is, an estimate of multivariate closeness (or verisimilitude) and intolerance (i.e., precision of prediction). Two

methods of calculating closeness were examined. The first method, was developed based on Pythagorean thinking, and is represented by:

$$CL = 1 - \left(\frac{1}{J} \sum_{j=1}^{J} \left(\frac{D_j}{S_j} \right)^2 \right)^{\frac{1}{2}} \qquad \text{[CL=1]}$$

The second method involved the computation of individual closeness estimates for each of the obtained path coefficients, calculation of the product of these terms, and taking the J^{th} root of the product term (with J equal to the number of path coefficients).

$$CL = \left[\prod_{j=1}^{J} \left(1 - \frac{D_j}{S_j} \right) \right]^{\frac{1}{J}} \qquad \text{[CL=2]}$$

To illuminate the difference in these two formulations of closeness let us consider a simple model with four variables and a tolerance interval of .5. For this example let us assume a lower limit of .25 and an upper limit of .75. For a given sample, the following path coefficients are obtained: p_{21}=.30, p_{31}=.04, p_{42}=.10, and p_{43}=.85. For each estimated path coefficient we must first calculate the deviation (d_j) from the lower or upper limit of the tolerance interval (e.g., d_4 = .85 - .75 = .10). Therefore the obtained deviations would be d_1=0, as .30 falls within the interval that ranges from .25 to .75; d_2=.21, d_3=.15 and d_4=.10. Given the first formulation of closeness (Cl=1), the obtained value of multivariate closeness was estimated to be:

$$CL = 1 - \left(\frac{1}{J} \sum_{j=1}^{J} \left(\frac{D_j}{S_j} \right)^2 \right)^{\frac{1}{2}}$$

$$CL = 1 - \left(\left(\frac{1}{4} \right) \left(\left(\frac{0}{2} \right)^2 + \left(\frac{.21}{2} \right)^2 + \left(\frac{.15}{2} \right)^2 + \left(\frac{.10}{2} \right)^2 \right) \right)^{\frac{1}{2}} = .93$$

For the second formulation of closeness (Cl=2), the obtained value of multivariate closeness was calculated as:

$$CL = \left[\prod_{j=1}^{J} \left(1 - \frac{D_j}{S_j} \right) \right]^{\frac{1}{J}}$$

$$CL = \left[\left(1 - \frac{0}{2} \right) \left(1 - \frac{.21}{2} \right) \left(1 - \frac{.15}{2} \right) \left(1 - \frac{.10}{2} \right) \right]^{\frac{1}{4}} = .94$$

Although the obtained values of closeness for these two calculations appear very similar in magnitude, the results of a small simulation that was conducted to evaluate the utility of the two proposed methods highlighted the difference in the formulations across a broader range of conditions. The results of this investigation are illustrated in Figure 15. Examination of this figure revealed the superior performance of the second method. For the first method, there appeared to be a sharp decline in the resultant value of C_i with small departures from truth. The more gradual, linear decline, consistent with the behavior of the univariate C_i , was deemed to be more representative of how this component of the corroboration index should contribute to the calculation of multivariate corroboration.

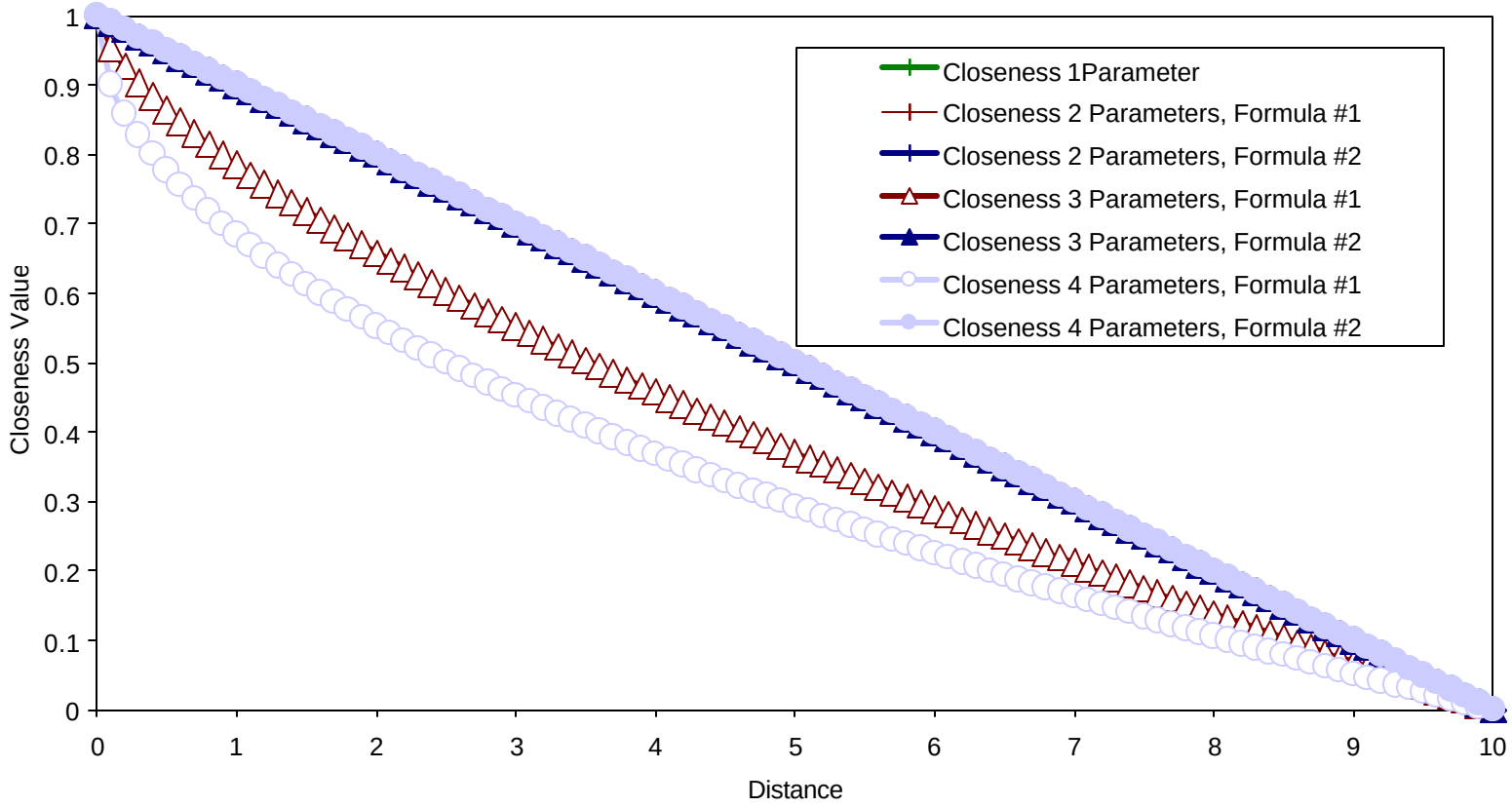

Figure 15. Alternative Methods for Computing Closeness, One, Two, Three, and Four Parameter Models.

Following the examination of these two approaches to closeness, alternative strategies for calculating intolerance were explored. The formula initially considered was simply the product of the standardized intolerances for the individual path coefficients:

$$In = \prod_{J=1}^{J}\left[1-\left(\frac{I_J}{S_J}\right)\right] \qquad \text{[In=1]}$$

This approach was not deemed profitable because the intolerance was found to get smaller as more variables were added to the model. For example, with a four variable model and a tolerance interval of 1, the application of this formula would yield an intolerance = .065, as calculated by $\left[\left(1-\frac{1}{2}\right)\left(1-\frac{1}{2}\right)\left(1-\frac{1}{2}\right)\left(1-\frac{1}{2}\right)\right]=.0625$. An alternative approach was to obtain the product of the tolerances for all of the parameters and then subtract that value from one.

$$In = 1-\prod_{j=1}^{J}\frac{I_j}{S_j} \qquad \text{[In=2]}$$

If we apply the same example to the second formulation of multivariate intolerance, we see that we obtain a very different answer. Applying the second approach to intolerance yields an intolerance = .9375, calculated as $1-\left[\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\right]=.9375$. Both methods of calculating intolerance are illustrated in Figure 16.

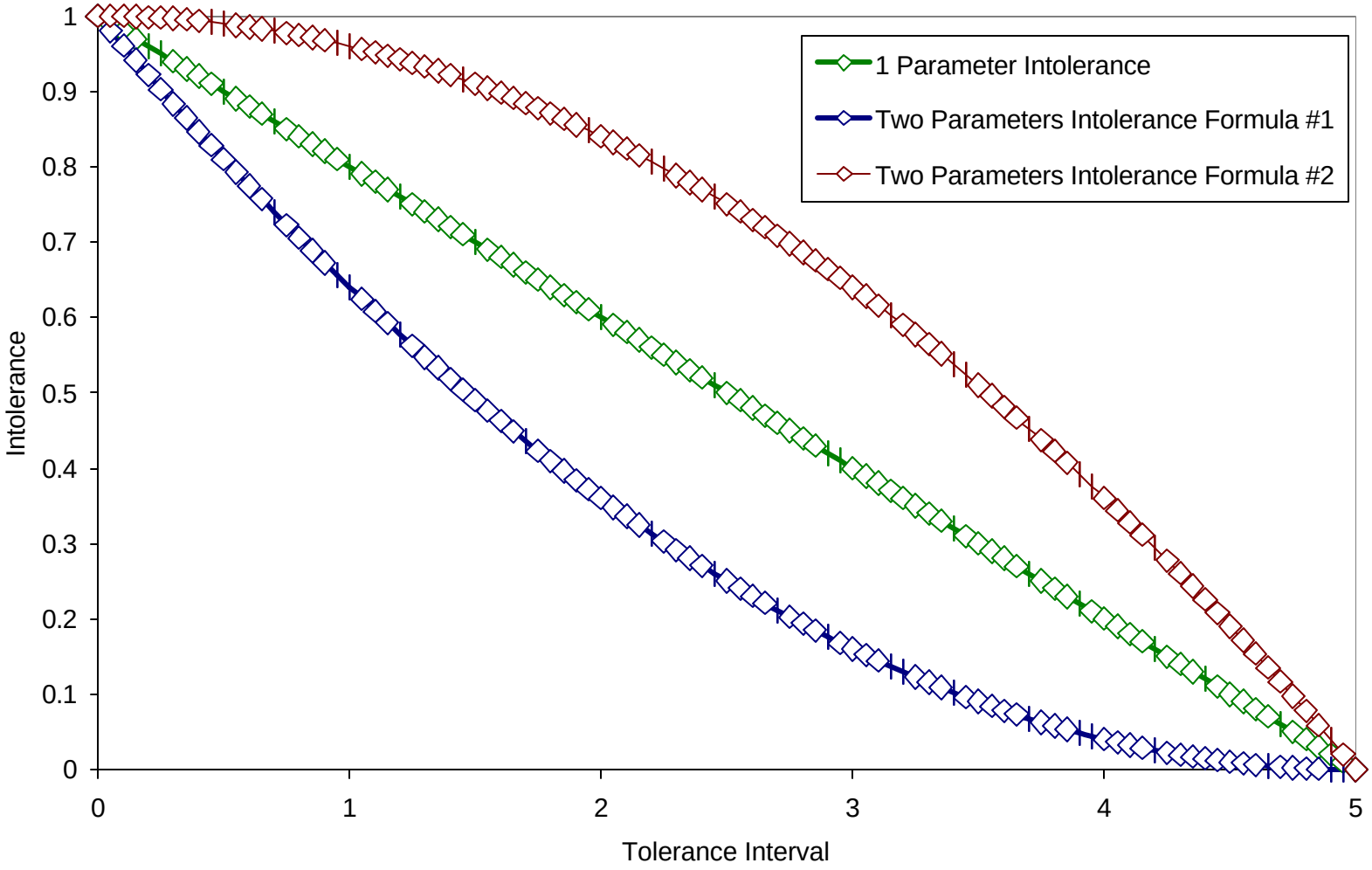

Figure 16. Alternative Methods for Computing Intolerance, One and Two Parameter Models.

However, a preliminary analysis of the application of the chosen formulate for intolerance revealed that this statistic approached its upper limit of 1.0 much too quickly as the number of paths evaluated increased. Therefore, an approach to 'tuning' the level of intolerance was investigated. In this approach, a root operation was performed on the product term:

$$In = 1 - \left(\prod_{j=1}^{J} \frac{I_j}{S_j} \right)^{\frac{1}{J-(J-1)X}} \qquad \text{[In=3]}$$

where X = some constant between 0 and 1.

If we consider the limits of such an exponent, if $X = 0$ then the exponent will reduce to $1/J$ or the j^{th} root. This most extreme case will not allow the multivariate intolerance to increase as the number of parameters increases. At the other extreme ($X = 1$), the exponent will reduce to 1.0 regardless of the number of parameters, which is the multivariate intolerance formula that was originally proposed. To help guide the selection of an appropriate level of adjustment, another small simulation study was conducted. These results are illustrated in Figure 17 for an intolerance level = .50 (i.e., directional prediction).

Examination of this figure reveals the incremental influence of various tuning adjustments to multivariate intolerance as parameters are added to a model. The calculation of intolerance, and hence mean Ci, was then submitted to a series of tuning adjustments to explore the influence of tuning on the multivariate index. Variability across three of the levels examined appeared relatively insignificant. A small sample of these values is provided in Table 2. To avoid the appearance of either overly downward or upward bias, it was decided

that a tuning factor of .50 would provide the appropriate correction to the index of multivariate intolerance.

Table 2

Obtained Value of Mean Ci for Three Levels of Tuning Multivariate Intolerance by Level of Verisimilitude, Precision of Prediction. Six Variable Model, Low Collinearity, Sample Size = 100

		Tune Level		
Precision of Prediction	Verisimilitude	.4	.5	.6
Non null	High	.08	.09	.10
	Moderate (MVD)	.08	.09	.10
	Low (LVD)	.07	.08	.09
	Moderate (MVA)	.08	.09	.11
	Low (LVA)	.08	.09	.11
Directional	High	.66	.71	.77
	Moderate (MVD)	.65	.70	.76
	Low (LVD)	.63	67	.72
	Moderate (MVA)	.66	.72	.78
	Low (LVA)	.66	.72	.79
Interval	High	.88	.92	.95
	Moderate (MVD)	.88	.91	.94
	Low (LVD)	.86	.89	.92
	Moderate (MVA)	.88	.92	.95
	Low (LVA)	.88	.91	.94

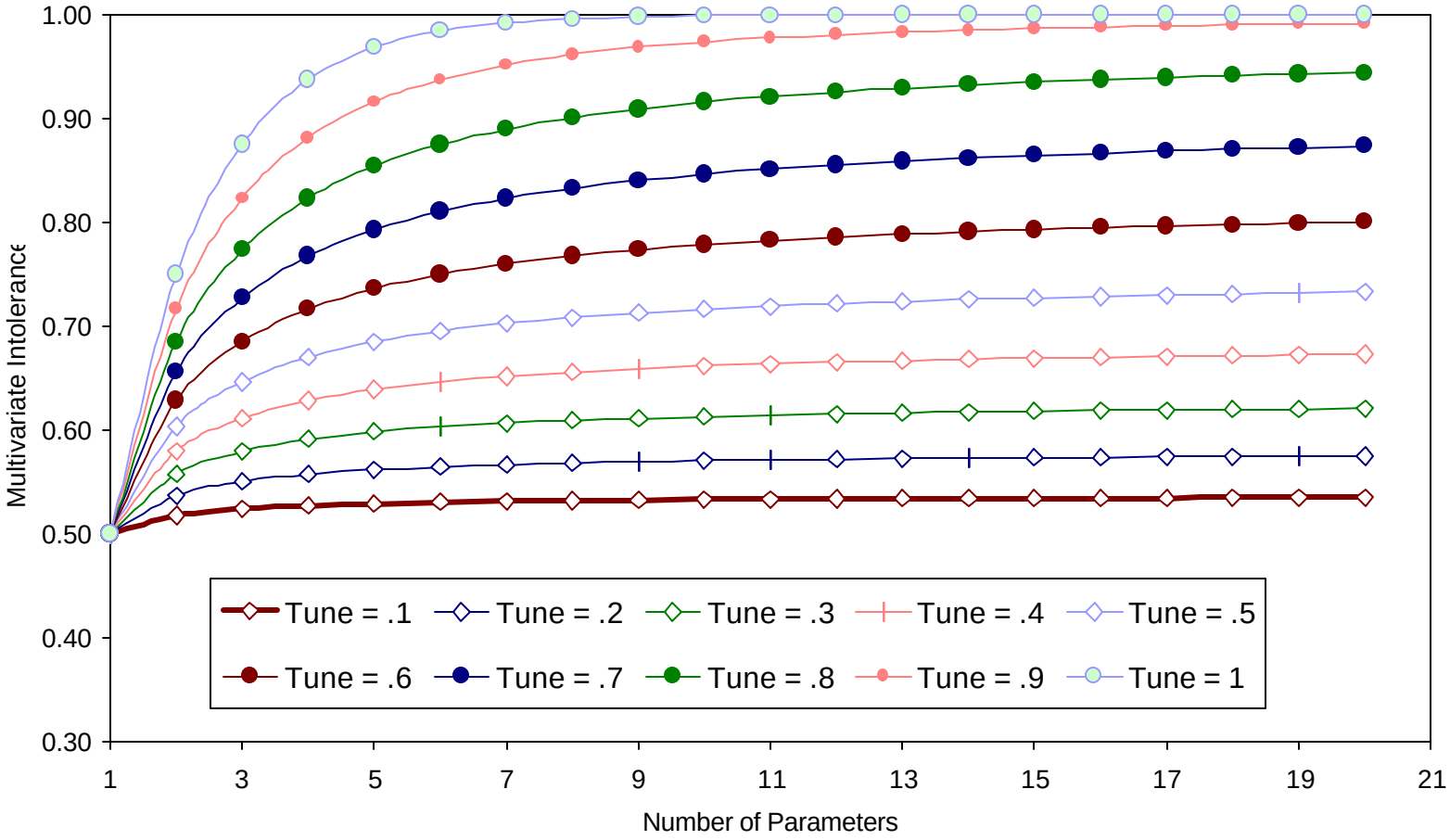

Figure 17. Tuning Multivariate Intolerance, Intolerance = .50.

Conduct of the Monte Carlo Study

This research was conducted using SAS/IML version 8.2. Conditions for the study were run under Windows 2000. For this study, six population correlation matrices were constructed based on a specified number of variables, level of collinearity, and true path model. The true population correlation matrices are exhibited in Tables 3-8.

Table 3

Population Correlation Matrix, 4 Variable Model, (VIF ≅ 1.5)

	X1	X2	X3	X4
X1	1.00			
X2	.75	1.00		
X3	.75	.56	1.00	
X4	.45	.47	.47	1.00

Table 4

Population Correlation Matrix, 4 Variable Model, (VIF ≅ 3.0)

	X1	X2	X3	X4
X1	1.00			
X2	.91	1.00		
X3	.91	.82	1.00	
X4	.54	.55	.55	1.00

Table 5

Population Correlation Matrix, 6 Variable Model, (VIF ≅ 1.5)

	X1	X2	X3	X4	X5	X6
X1	1.00					
X2	.58	1.00				
X3	.58	.58	1.00			
X4	.47	.36	.50	1.00		
X5	.36	.47	.50	.27	1.00	
X6	.43	.43	.61	.54	.54	1.00

Table 6

Population Correlation Matrix, 6 Variable Model, (VIF ≅ 3)

	X1	X2	X3	X4	X5	X6
X1	1.00					
X2	.82	1.00				
X3	.54	.54	1.00			
X4	.54	.48	.54	1.00		
X5	.48	.54	.54	.36	1.00	
X6	.49	.49	.65	.59	.59	1.00

Table 7

Population Correlation Matrix, 8 Variable Model, (VIF ≅ 1.5)

	X1	X2	X3	X4	X5	X6	X7	X8
X1	1.00							
X2	.41	1.00						
X3	.42	.42	1.00					
X4	.42	.42	.25	1.00				
X5	.42	.42	.25	.25	1.00			
X6	.37	.37	.46	.43	.43	1.00		
X7	.25	.25	.15	.38	.38	.26	1.00	
X8	.31	.31	.26	.32	.54	.49	.48	1.00

Table 8

Population Correlation Matrix, 8 Variable Model, (VIF ≅ 3.0)

	X1	X2	X3	X4	X5	X6	X7	X8
X1	1.00							
X2	.82	1.00						
X3	.54	.54	1.00					
X4	.54	.54	.33	1.00				
X5	.54	.54	.33	.33	1.00			
X6	.56	.56	.58	.53	.59	1.00		
X7	.34	.34	.21	.42	.42	.35	1.00	
X8	.45	.45	.34	.39	.64	.60	.56	1.00

Data Generation Strategy

Multivariate normal samples were generated from each population R matrix and a sample correlation matrix was computed for each sample. Each of the sample correlation matrices was then analyzed using a series of regression equations. These regression equations were determined by the desired level of verisimilitude that was being examined (i.e., the path model implied by the theory). It was the theoretical model that determined the appropriate regression equations to employ. The series of regression equations were applied to each sample and the resulting parameter estimates were used in the calculation of the closeness component of the corroboration index. In the final computation of the corroboration indices, the size of the tolerance interval was manipulated. The program code was verified by hand-checking results from benchmark datasets. The data resulting from each path analysis were pooled and the average value of C_i was evaluated in the context of the central design factors. The method for data simulation is illustrated in Figure 18.

For each population matrix, 10,000 samples were generated. The use of 10,000 samples provided adequate precision of estimates of the sampling behavior of the corroboration index**.** For example, 10,000 samples provide a maximum 95% confidence interval width around an observed proportion that is $\pm$ .0098 (Robey & Barcikowski, 1992).

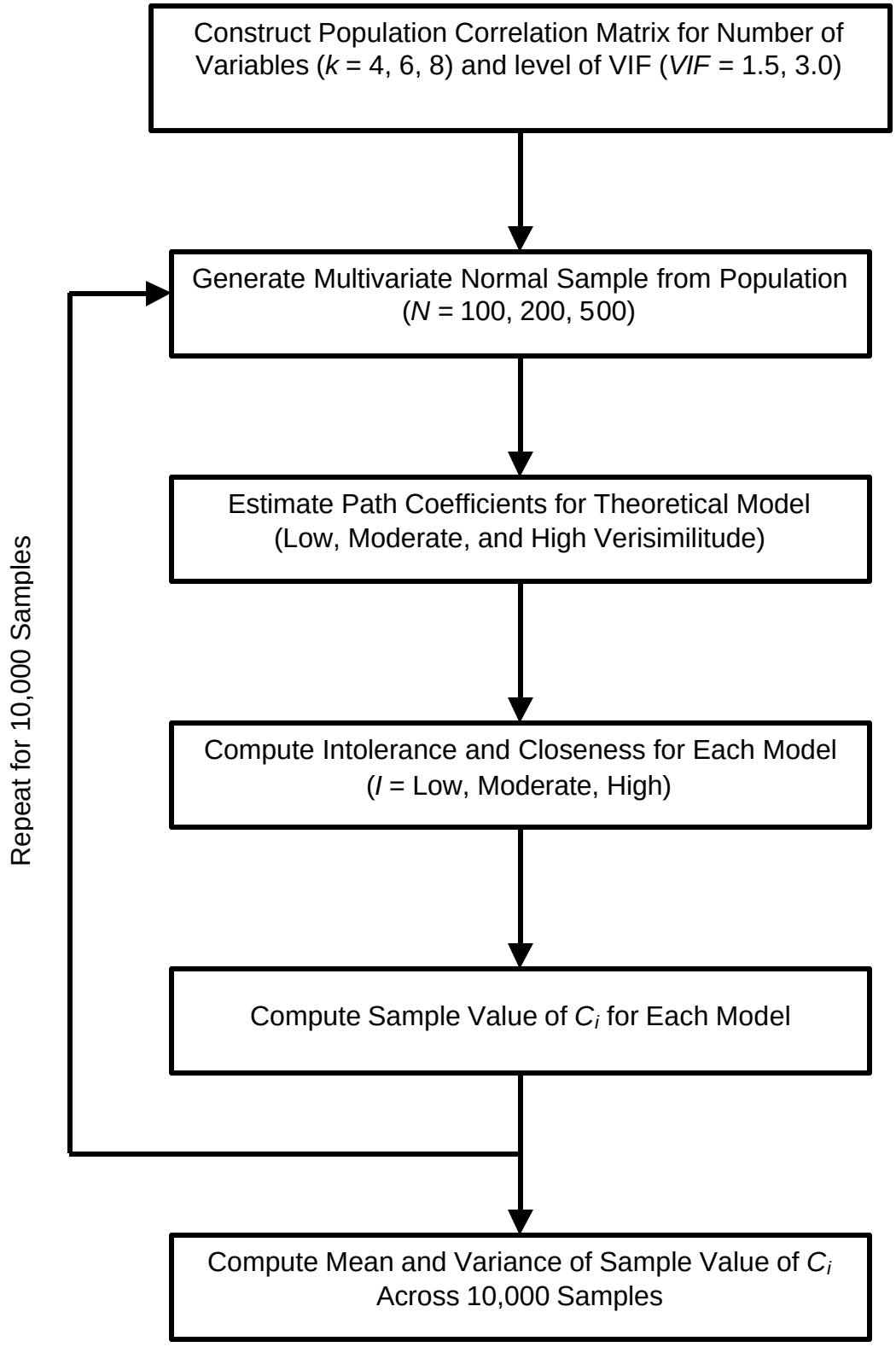

Figure 18. Data Generation Strategy.

Data Analysis

To guide the interpretation of the simulation results, C_i was treated as a dependent variable and a factorial ANOVA was conducted. The independent variables for this ANOVA were the five Monte Carlo design factors: (a) the number of variables in the model (i.e., model complexity), (b) level of intolerance (c) level of verisimilitude, (d) sample size, and (e) level of collinearity. In addition to the main effects, the interactions of these factors were examined. For each of these analyses, an effect size estimate, omega-squared ($\hat{w}^2$), was used to estimate the proportion of variance accounted for in the population by each effect (Maxwell & Delaney, 1990). For the calculation of this effect, within-cell variability was provided by the variance of the 10,000 replications of each condition.

The estimate of $\hat{w}^2$ is given by

$$\hat{w}^2 = \frac{SS_{effect} - (df_{effect})(MS_{wg})}{SS_T + MS_{wg}}$$

The results of this research are presented in tables and graphs that address each of the research questions and hypotheses through an illustration of the relationship between the central design facets and the resultant mean C_i and standard deviation of C_i. Further, a series of tables and graphs are employed to illustrate the extent to which the relationship between mean C_i and level of intolerance was influenced by verisimilitude, collinearity and sample size. When deemed appropriate, supplementary analyses and results are presented to further explain some of the more unexpected relationships evidenced in the data.

Chapter Four

Results

Organization

The purpose of this chapter is to present the results of this study. As the results are organized with respect to the research questions and hypotheses, the chapter opens with a restatement of the proposed research questions and hypotheses. Within each section, after each of the primary research questions and hypotheses have been addressed, supplementary analyses and results are examined in order to further elucidate some of the more subtle relationships evidenced in the data. At the end of the chapter key findings are underscored and summarized.

Four Research Questions and Three Research Hypotheses

Research Questions

1. What is the relationship between mean C_i and the main effects examined in the study (i.e., verisimilitude, intolerance, model complexity, collinearity, and sample size)?
2. What is the relationship between the standard deviation of C_i and model complexity, collinearity, and sample size?
3. To what extent is the relationship between mean C_i and the precision of prediction (i.e., intolerance) influenced by the complexity of the model (i.e., the number of variables in the model)?

4. To what extent is the relationship between mean C_i and the precision of prediction (i.e., intolerance) influenced by the level of collinearity?

Research Hypotheses

1. The relationship between mean C_i and the precision of prediction (i.e., intolerance will be slightly influenced by the closeness of the data to the theory (verisimilitude).
2. The relationship between mean C_i and the precision of prediction (i.e., intolerance) will not be substantively influenced by sample size.
3. The relationship between mean C_i and precision of prediction will be substantively stronger than the relationship between mean C_i and verisimilitude, model complexity, collinearity, and sample size.

Relationship Between Mean C_i *and the Central Design Factors*

As stated earlier, in order to guide the interpretation of the simulation results, the average C_i was treated as a dependent variable and a factorial ANOVA was conducted. The independent variables for this ANOVA were the five Monte Carlo design factors: (a) the number of variables in the model (i.e., model complexity), (b) level of intolerance (c) level of verisimilitude, (d) sample size, and (e) level of collinearity. In addition to these main effects, the interactions of these factors were also examined. For each of these analyses, an effect size estimate, omega-squared ($\hat{w}^2$), was used to estimate the proportion of variance accounted for in the population by each effect. The results of these

analyses are presented in Table 9. An examination of the obtained effect sizes revealed that only a single effect evidenced considerable influence on the average C_i. As anticipated, the level of intolerance was the most salient influence on average C_i, with an estimated $\hat{w}^2$=.55. The remaining main effects and interaction effects exercised negligible influence on mean C_i. The residual mean square value presented along with the obtained values of omega-squared represents the average variability within each condition (or cell) under examination.

Probing Deeper: The Influence of Verisimilitude, Model Complexity, Collinearity, and Sample Size after Controlling for Intolerance

In light of the very strong influence of this single design factor, it appeared fruitful to examine the other main effects and interaction effects after controlling for the level of intolerance. Therefore, three additional analyses were conducted. Again, mean C_i was treated as the dependent variable and three separate ANOVAs, one for each level of intolerance, were conducted with the remaining four design factors (i.e., the number of variables in the model, verisimilitude, sample size and level of collinearity) treated as independent variables. Consistent with the initial analysis, the interactions of these factors were also examined. The results of this set of analyses are presented in Tables 10-12. The resultant values of $\hat{w}^2$, suggest that the number of variables in the model ($\hat{w}^2$=.26) and level of verisimilitude ($\hat{w}^2$=.21) were somewhat influential, but only for lowest level of precision (i.e., non null predictions). These analyses also revealed the lack of influence of any of the other central design factors examined.

Table 9

Estimated DF, SS, and Omega Squared by Design Factors

Effect	DF	SS	$\hat{w}^2$
N of Variables (k)	2	0.07	<.01
Verisimilitude (V)	4	0.05	<.01
k*V	8	0.05	<.01
Collinearity (C)	1	<.01	<.01
k*C	2	<.01	<.01
V*C	4	<.01	<.01
k*V*C	8	<.01	<.01
N of Observations (N)	2	<.01	<.01
k*N	4	<.01	<.01
V*N	8	<.01	<.01
k*V*N	16	<.01	<.01
C*N	2	<.01	<.01
k*C*N	4	<.01	<.01
V*C*N	8	<.01	<.01
k*V*C*N	16	<.01	<.01
Intolerance (I)	2	32.40	0.55
k*I	4	0.02	<.01
V*I	8	0.01	<.01
k*V*I	16	<.01	<.01
C*I	2	<.01	<.01
k*C*I	4	<.01	<.01
V*C*I	8	<.01	<.01
k*V*C*I	16	<.01	<.01
k*I	4	<.01	<.01
k*N*I	8	<.01	<.01
V*N*I	16	<.01	<.01
k*V*N*I	32	<.01	<.01
C*N*I	4	<.01	<.01
k*C*N*I	8	<.01	<.01
V*C*N*I	16	<.01	<.01
k*V*C*N*I	32	<.01	<.01
Residual MS	26	2699729	<.01

Table 10

Estimated DF, SS, and Omega Squared,

Intolerance = Non Null Prediction

Effect	DF	SS	$\hat{w}^2$
N of Variables (k)	2	<.01	0.26
Verisimilitude (V)	4	<.01	0.21
k*V	8	<.01	0.01
Collinearity (C)	1	<.01	<.01
k*C	2	<.01	<.01
V*C	4	<.01	<.01
k*V*C	8	<.01	<.01
N of Observations (N)	2	<.01	<.01
k*N	4	<.01	<.01
V*N	8	<.01	<.01
k*V*N	16	<.01	<.01
C*N	2	<.01	<.01
k*C*N	4	<.01	<.01
V*C*N	8	<.01	<.01
k*V*C*N	16	<.01	<.01
Residual MS	<.01	89909	<.01

Table 11

Estimated DF, SS, and Omega Squared,

Intolerance = Directional Prediction

Effect	DF	SS	$\hat{w}^2$
N of Variables (k)	2	0.06	0.01
Verisimilitude (V)	4	0.04	0.01
k*V	8	<.01	<.01
Collinearity (C)	1	<.01	<.01
k*C	2	<.01	<.01
V*C	4	<.01	<.01
k*V*C	8	<.01	<.01
N of Observations (N)	2	<.01	<.01
k*N	4	<.01	<.01
V*N	8	<.01	<.01
k*V*N	16	<.01	<.01
C*N	2	<.01	<.01
k*C*N	4	<.01	<.01
V*C*N	8	<.01	<.01
k*V*C*N	16	<.01	<.01
Residual MS	6	899909	<.01

Table 12

Estimated DF, SS, and Omega Squared,

Intolerance = Interval Prediction

	DF	SS	$\hat{w}^2$
Effect			
N of Variables (k)	2	0.03	<.01
Verisimilitude (V)	4	0.02	<.01
k*V	8	<.01	<.01
Collinearity (C)	1	<.01	<.01
k*C	2	<.01	<.01
V*C	4	<.01	<.01
k*V*C	8	<.01	<.01
N of Observations (N)	2	<.01	<.01
k*N	4	<.01	<.01
V*N	8	<.01	<.01
k*V*N	16	<.01	<.01
C*N	2	<.01	<.01
k*C*N	4	<.01	<.01
V*C*N	8	<.01	<.01
k*V*C*N	16	<.01	<.01
Residual MS	19.98	899909	<.01

Estimates of Mean Ci

To facilitate the interpretation of the results in this section, and to gain a better understanding of the nature of the models under investigation, the level of model complexity, model misspecification and the number of estimated paths under examination are displayed in Table 13. In this table and in all of the tables and figures that follow, HV represents conditions with *high verisimilitude*; MVD represents models with *moderate verisimilitude*, with model misspecification occurring as paths are deleted; and LVD represents models with *low*

verisimilitude resulting from additional paths being deleted. MVA represents a *moderate level of verisimilitude* that resulted when one or more paths are added to the model, while LVA represents, l*ow verisimilitude*, occurring when further paths are added.

To further aid in the interpretation of the results it may be useful to reconsider the relationship between the raw tolerance interval of the theory and the level of intolerance examined in this study. Recall that the raw tolerance interval of the theory was examined at three levels of precision: a non null condition with $\beta \neq 0$, a directional condition employing ½the Spielraum, with $\beta < 0 \text{ or } \beta > 0$, and an interval prediction equal to ¼ of the Spielraum, that is $.05 < |\beta| < .55$. Translated into numerical terms, a non null prediction employs 95% of the range of expected values; a directional prediction equates to 50% of the expected values, while an interval prediction compares obtained values to a targeted 25% of the Spielraum.

Table 13
Model Complexity, Verisimilitude, and Number of Estimated Paths

Model	Verisimilitude				
Complexity	HV	MVD	LVD	MVA	LVA
Low (4)	4	3	2	5	6
Moderate (6)	9	7	4	11	14
High (8)	14	10	5	18	23

From Table 13 it is easy to discern the exact number of paths estimated in each model. For example, in both the low complexity, high verisimilitude model (HV) and the moderate complexity low verisimilitude exclusionary model (LVD)

four paths are estimated. Similarly, both the moderate complexity model low verisimilitude auxiliary model (LVA) and the high complexity high verisimilitude model (HV) both include 14 estimated paths. For illustrative purposes, diagrams representing each level of model complexity and misspecification are provided in Figures 19-30.

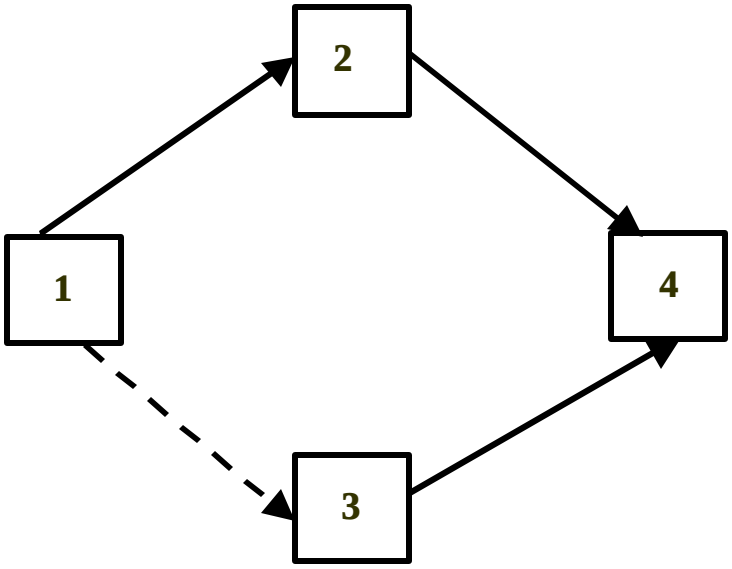

Figure 19. Four Variable Exclusionary Model, Moderate Level of Verisimilitude (MVD).

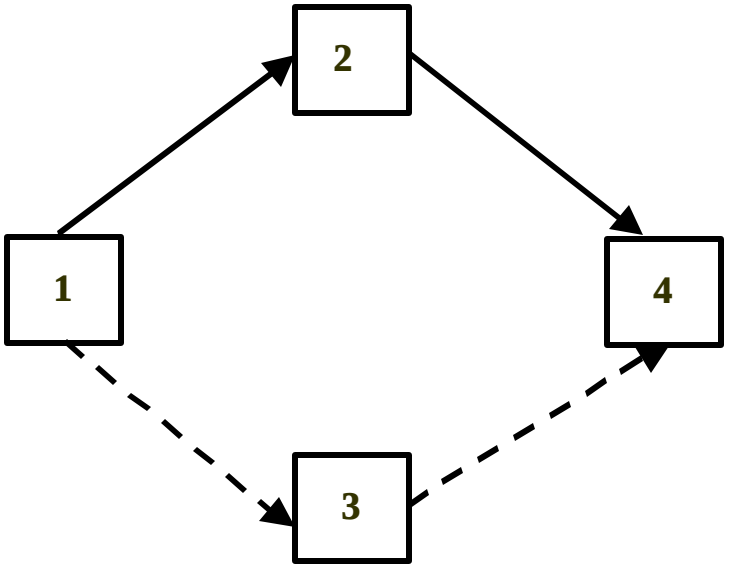

Figure 20. Four Variable Exclusionary Model, Low Level of Verisimilitude (LVD).

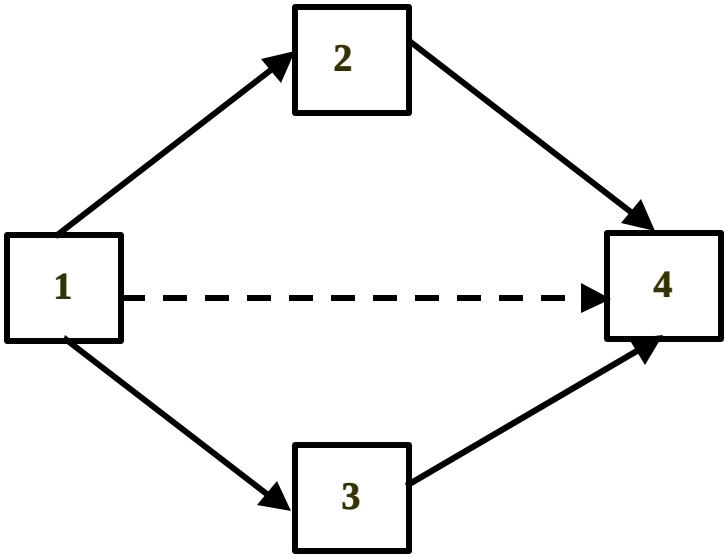

Figure 21. Four Variable Supplementary Model, Moderate Level of Verisimilitude (MVA).

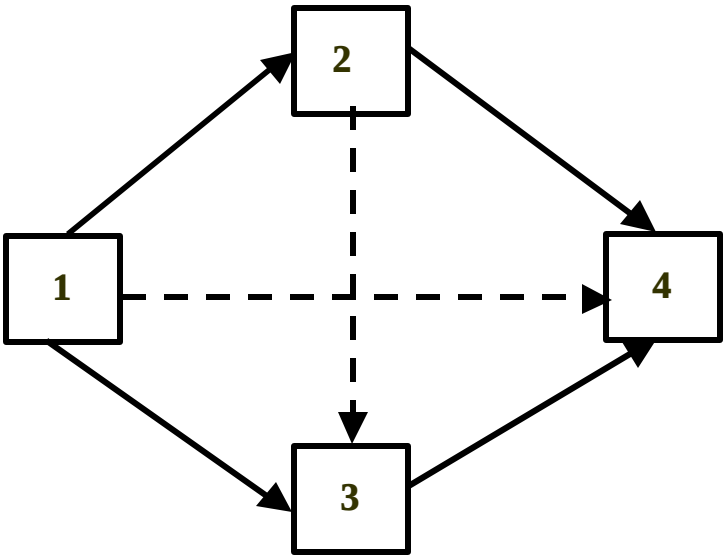

Figure 22. Four Variable Supplementary Model, Low Level of Verisimilitude (LVA).

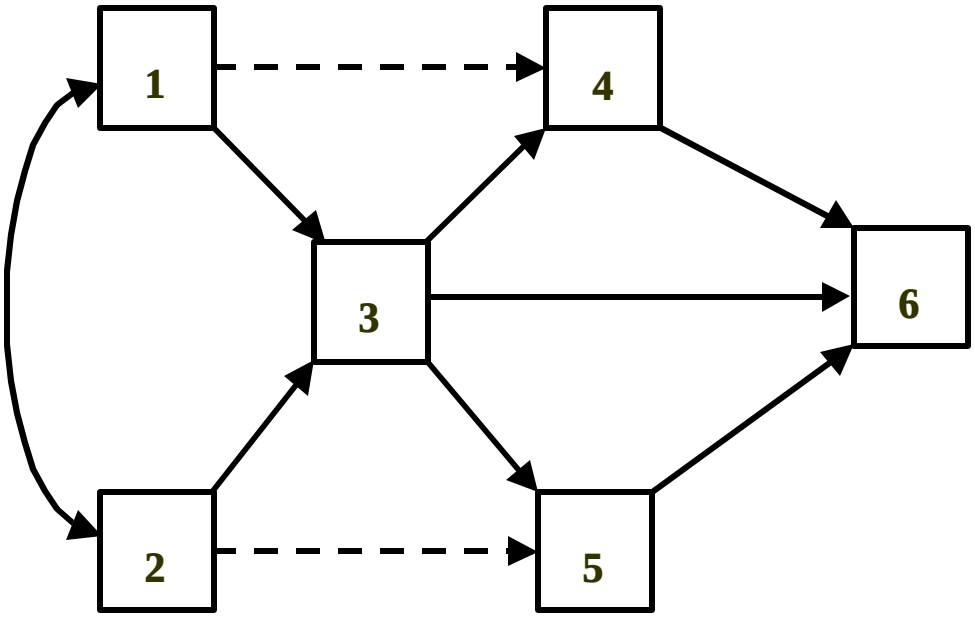

Figure 23. Six Variable Exclusionary Model, Moderate Level of Verisimilitude (MVD).

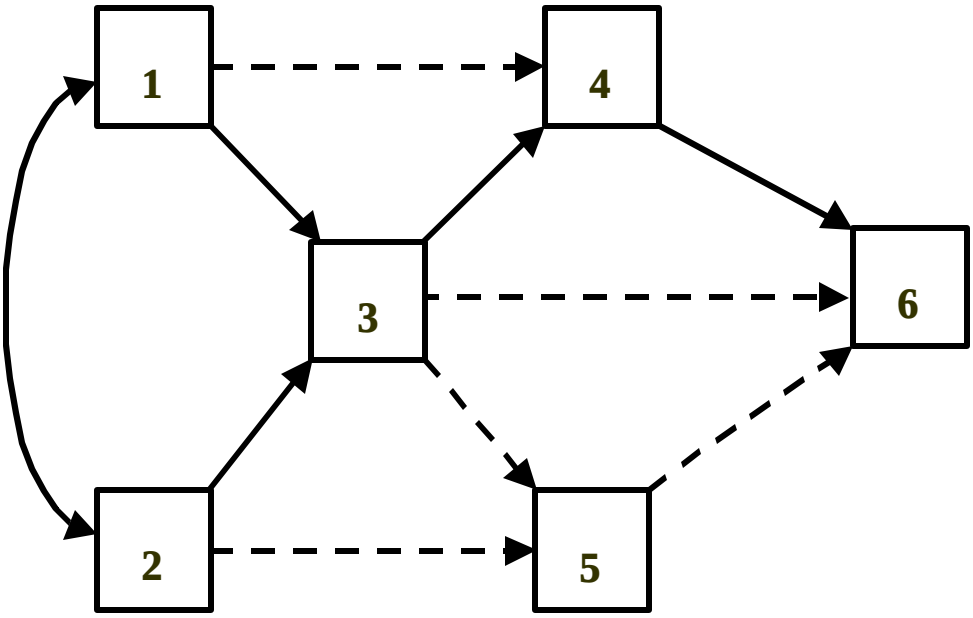

Figure 24. Six Variable Exclusionary Model, Low Level of Verisimilitude (LVD).

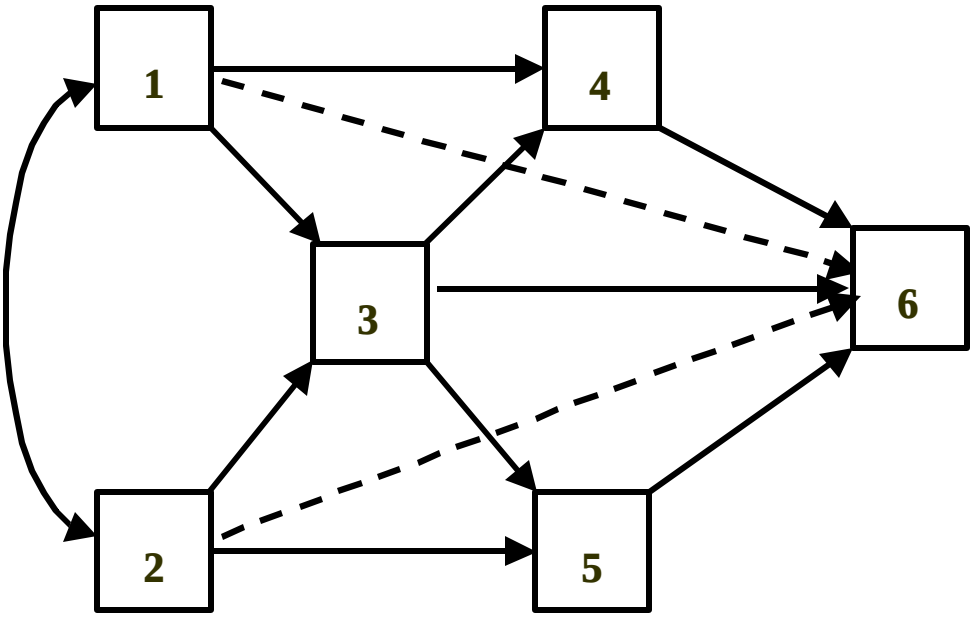

Figure 25. Six Variable Supplementary Model, Moderate Level of Verisimilitude (MVA).

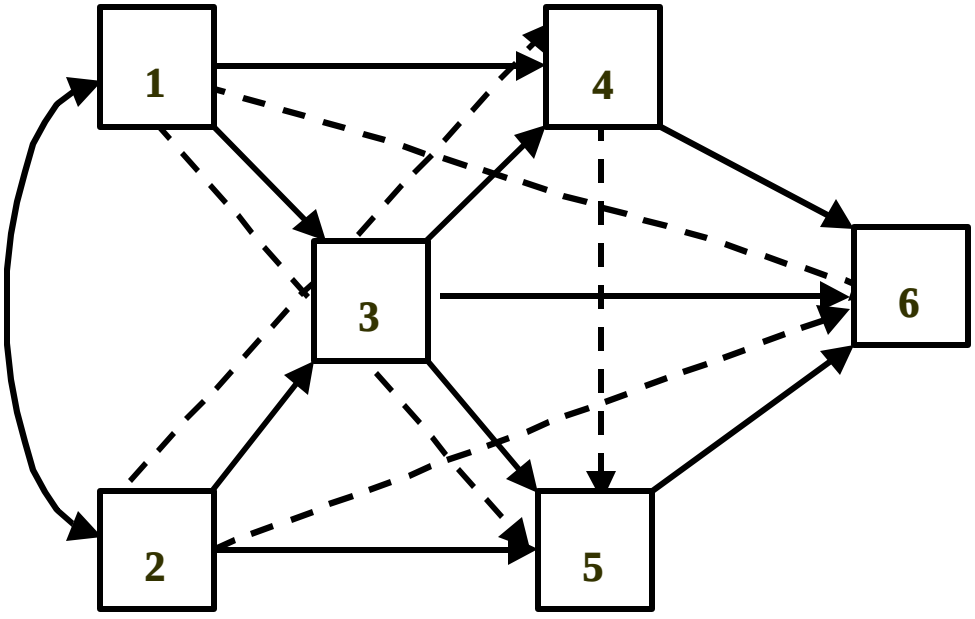

Figure 26. Six Variable Supplementary Model, Low Level of Verisimilitude (LVA).

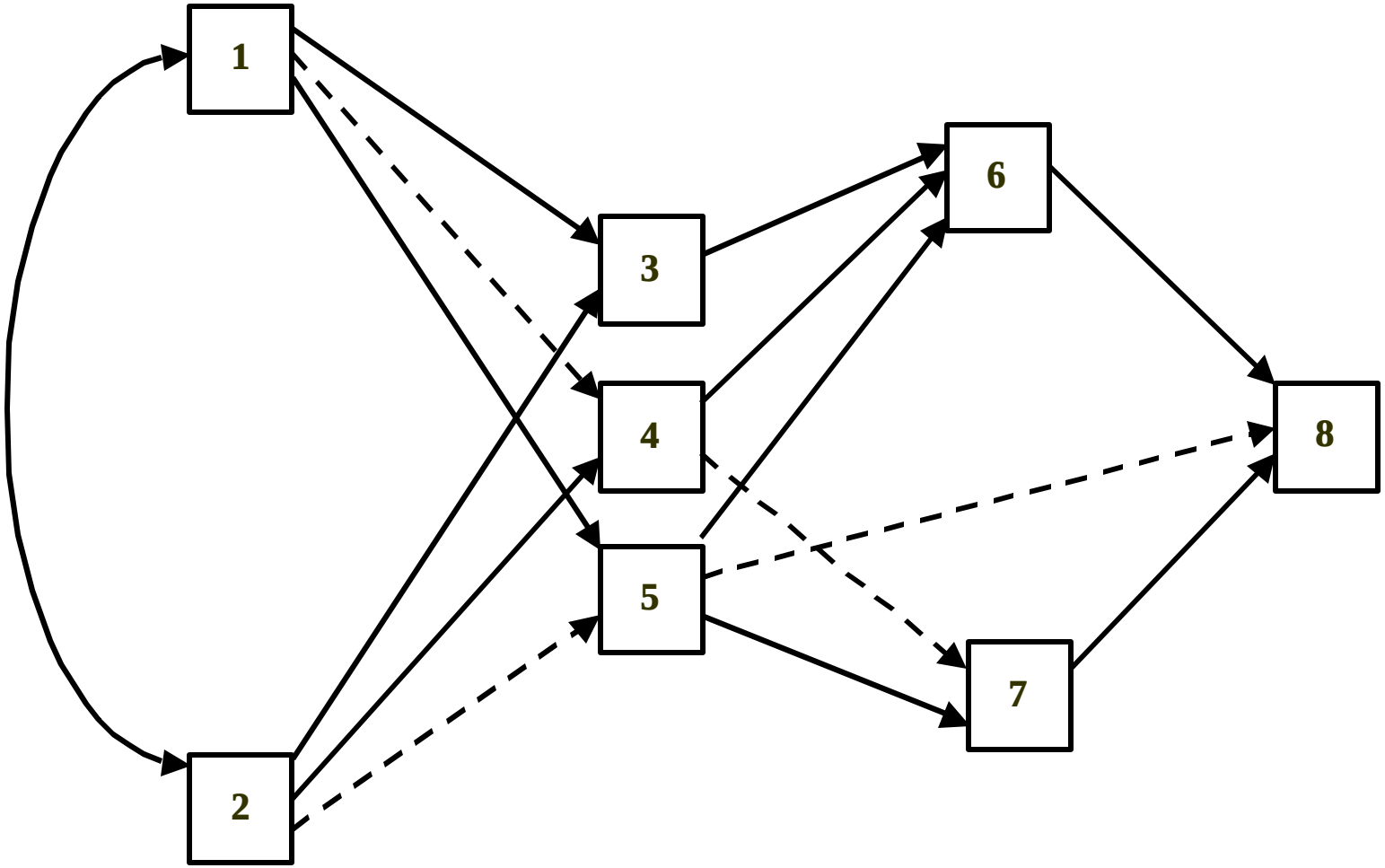

*Figure 27.*Eight Variable Exclusionary Model, Moderate Level of Verisimilitude (MVD).

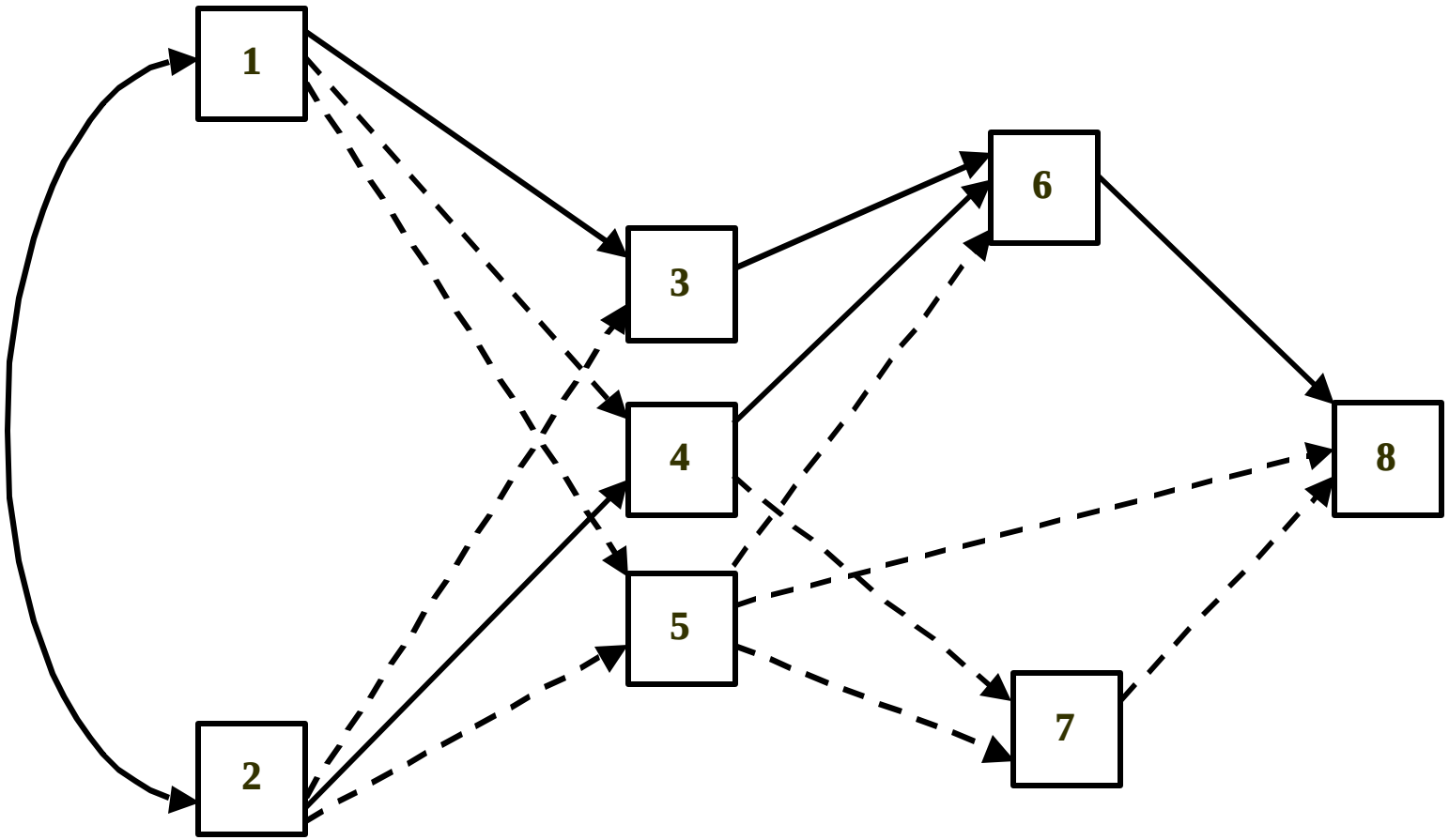

Figure 28. Eight Variable Exclusionary Model, Low Level of Verisimilitude (LVD).

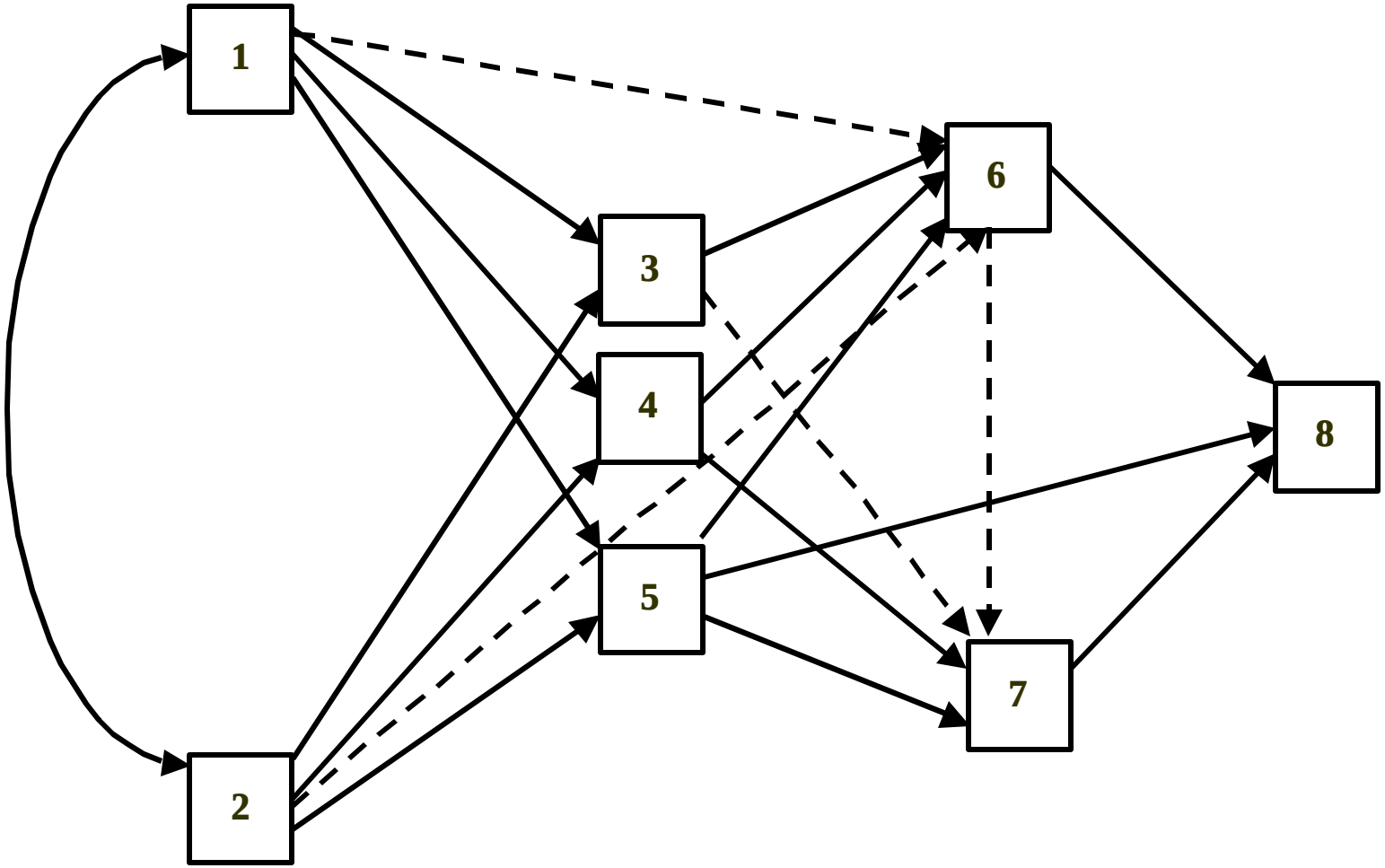

Figure 29. Eight Variable Supplementary Model, Moderate Level of Verisimilitude (MVA).

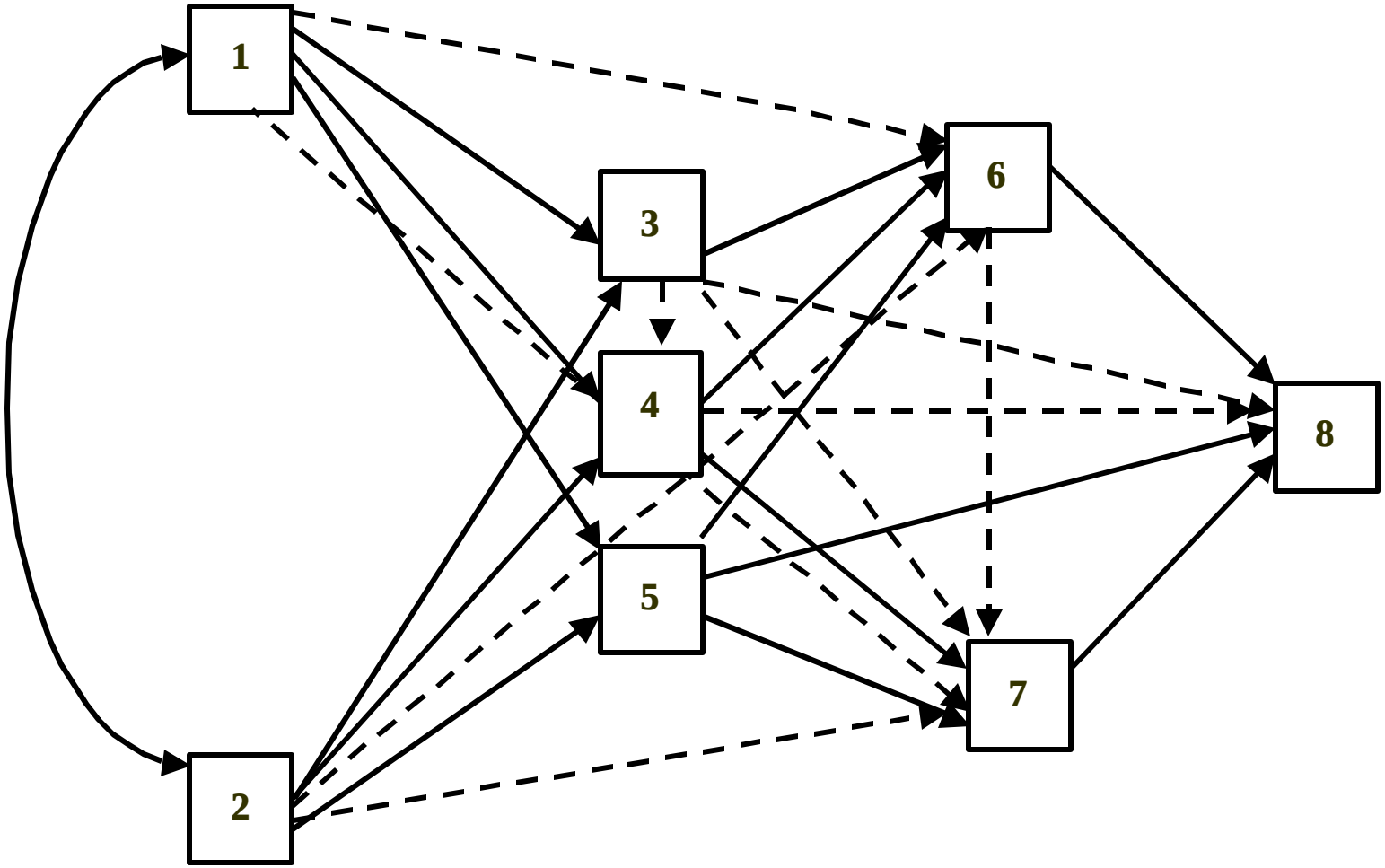

Figure 30. 8 Variable Supplementary Model, Low Level of Verisimilitude (LVA).

Mean Ci by Precision of Prediction and Level of Verisimilitude

The obtained values of mean C_i by level of verisimilitude, collinearity, sample size and precision of prediction are presented in Tables 14-16. As trends across this number of conditions may be challenging to visualize, a series of box and whisker plots is also provided (see Figures 31-33).

As these figures clearly illustrate, there is a very strong relationship between the magnitude of mean C_i and the level of intolerance. These figures also reveal a general lack of variability across the various levels of verisimilitude for both the non null and interval predictions. For example, for the weakest predictions (i.e., non null in nature), mean C_i was estimated to range from .07 to .09, and for the most precise predictions, mean C_i ranged from .84 to .92. However, for the directional prediction, mean C_i was observed to range from .60 to .73. For both directional and interval predictions the variability in mean C_i was more pronounced for the low verisimilitude, exclusionary models (LVD) than for any of the other models examined.

Viewed from a slightly different perspective, the considerable influence of intolerance on mean C_i is also evident if we examine this relationship across the number of estimated paths in the various models under consideration (see Figure 34). In this illustration, we can readily observe that as the number of estimated paths increases there is very little variability within each level of intolerance, yet with increasingly precise predictions, the magnitude of mean C_i rises dramatically.

Table 14
Mean C_i by Intolerance, Verisimilitude, Collinearity, and Sample Size
Model Complexity = Low

Verisimilitude	Collinearity	Sample Size	Mean C_i		
			Non Null	Directional	Interval
HV	Low	100	0.08	0.67	0.89
MVD			0.07	0.65	0.87
LVD			0.07	0.60	0.84
MVA			0.08	0.68	0.89
LVA			0.08	0.69	0.89
HV		200	0.08	0.67	0.89
MVD			0.07	0.65	0.87
LVD			0.07	0.60	0.84
MVA			0.08	0.68	0.89
LVA			0.08	0.69	0.90
HV		500	0.08	0.67	0.89
MVD			0.07	0.65	0.88
LVD			0.07	0.60	0.84
MVA			0.08	0.68	0.90
LVA			0.08	0.69	0.90
HV	Moderate	100	0.08	0.67	0.89
MVD			0.07	0.65	0.87
LVD			0.07	0.60	0.84
MVA			0.08	0.68	0.89
LVA			0.08	0.68	0.88
HV		200	0.08	0.67	0.89
MVD			0.07	0.65	0.87
LVD			0.07	0.60	0.84
MVA			0.08	0.68	0.89
LVA			0.08	0.68	0.88
HV		500	0.08	0.67	0.89
MVD			0.07	0.65	0.87
LVD			0.07	0.60	0.84
MVA			0.08	0.68	0.89
LVA			0.08	0.68	0.89

Table 15

Mean C_i by Intolerance, Verisimilitude, Collinearity, and Sample Size
Model Complexity = Moderate

Verisimilitude	Collinearity	Sample Size	Mean C_i		
			Non Null	Directional	Interval
HV	Low	100	0.09	0.71	0.92
MVD			0.09	0.70	0.91
LVD			0.08	0.67	0.89
MVA			0.09	0.72	0.92
LVA			0.09	0.72	0.91
HV		200	0.09	0.71	0.92
MVD			0.09	0.70	0.91
LVD			0.08	0.67	0.89
MVA			0.09	0.72	0.92
LVA			0.09	0.72	0.91
HV		500	0.09	0.71	0.92
MVD			0.09	0.70	0.91
LVD			0.08	0.67	0.89
MVA			0.09	0.72	0.92
LVA			0.09	0.72	0.92
HV	Moderate	100	0.09	0.71	0.92
MVD			0.09	0.70	0.91
LVD			0.08	0.67	0.87
MVA			0.09	0.72	0.91
LVA			0.09	0.72	0.91
HV		200	0.09	0.71	0.92
MVD			0.09	0.70	0.91
LVD			0.08	0.67	0.87
MVA			0.09	0.72	0.92
LVA			0.09	0.72	0.91
HV		500	0.09	0.71	0.92
MVD			0.09	0.70	0.91
LVD			0.08	0.67	0.87
MVA			0.09	0.72	0.92
LVA			0.09	0.72	0.92

Table 16

Mean C_i by Intolerance, Verisimilitude, Collinearity, and Sample Size

Model Complexity = High

Verisimilitude	Collinearity	Sample Size	Mean C_i		
			Non Null	Directional	Interval
HV	Low	100	0.09	0.73	0.92
MVD			0.09	0.72	0.92
LVD			0.08	0.69	0.90
MVA			0.09	0.73	0.92
LVA			0.09	0.73	0.92
HV		200	0.09	0.73	0.92
MVD			0.09	0.72	0.92
LVD			0.08	0.69	0.90
MVA			0.09	0.73	0.92
LVA			0.09	0.73	0.92
HV		500	0.09	0.73	0.92
MVD			0.09	0.72	0.92
LVD			0.08	0.69	0.90
MVA			0.09	0.73	0.92
LVA			0.09	0.73	0.92
HV	Moderate	100	0.09	0.73	0.92
MVD			0.09	0.72	0.92
LVD			0.08	0.69	0.90
MVA			0.09	0.73	0.92
LVA			0.09	0.73	0.91
HV		200	0.09	0.73	0.92
MVD			0.09	0.72	0.92
LVD			0.08	0.69	0.90
MVA			0.09	0.73	0.92
LVA			0.09	0.73	0.92
HV		500	0.09	0.73	0.92
MVD			0.09	0.72	0.92
LVD			0.08	0.69	0.90
MVA			0.09	0.73	0.92
LVA			0.09	0.73	0.92

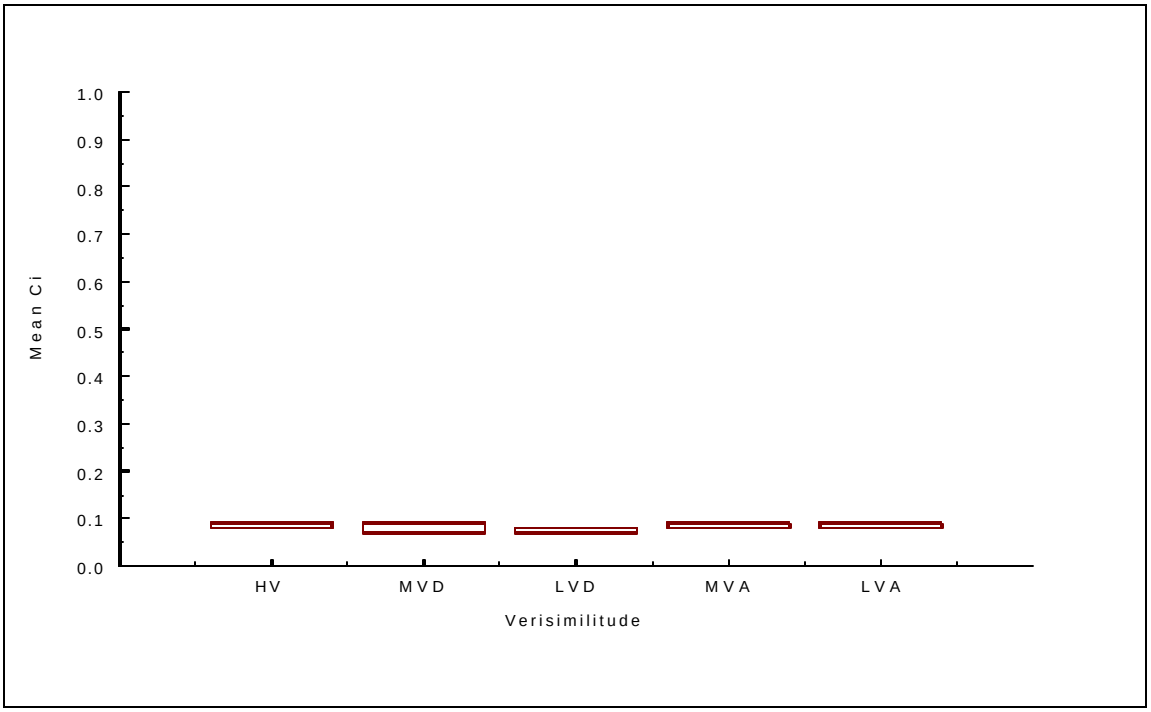

Figure 31. Mean C_i by Level of Verisimilitude, Non Null Prediction.

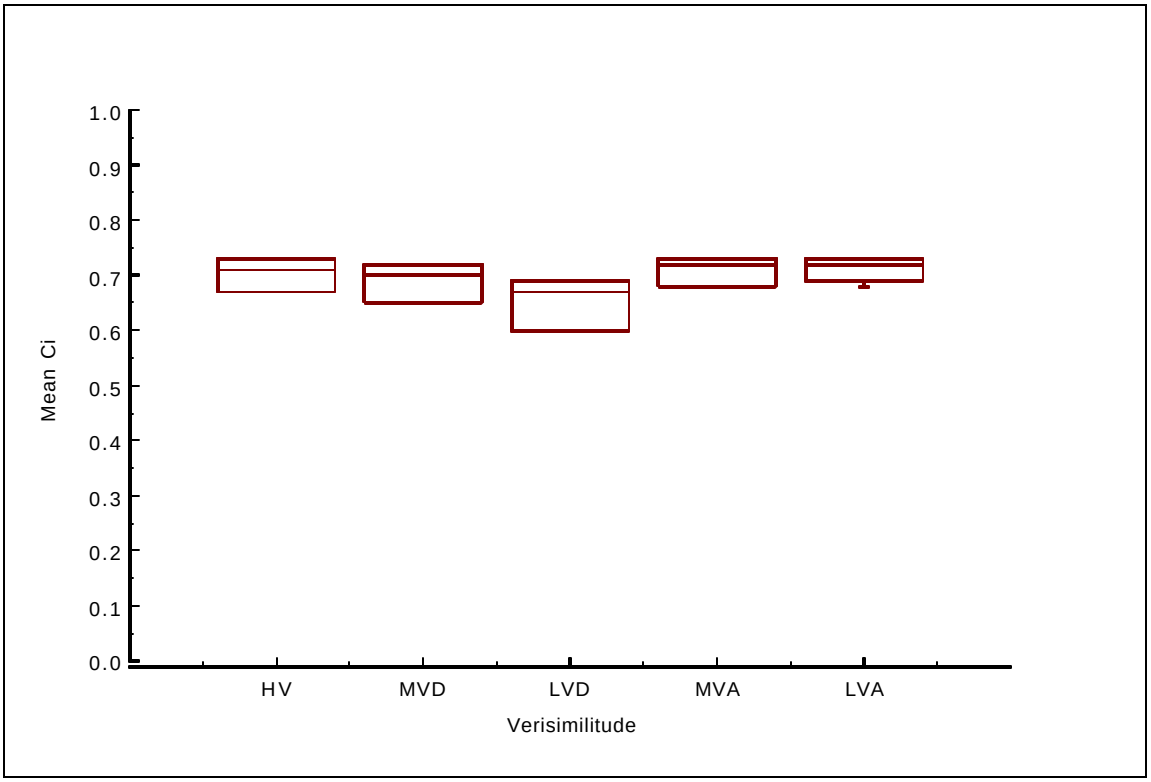

Figure 32. Mean C_i by Level of Verisimilitude, Directional Prediction.

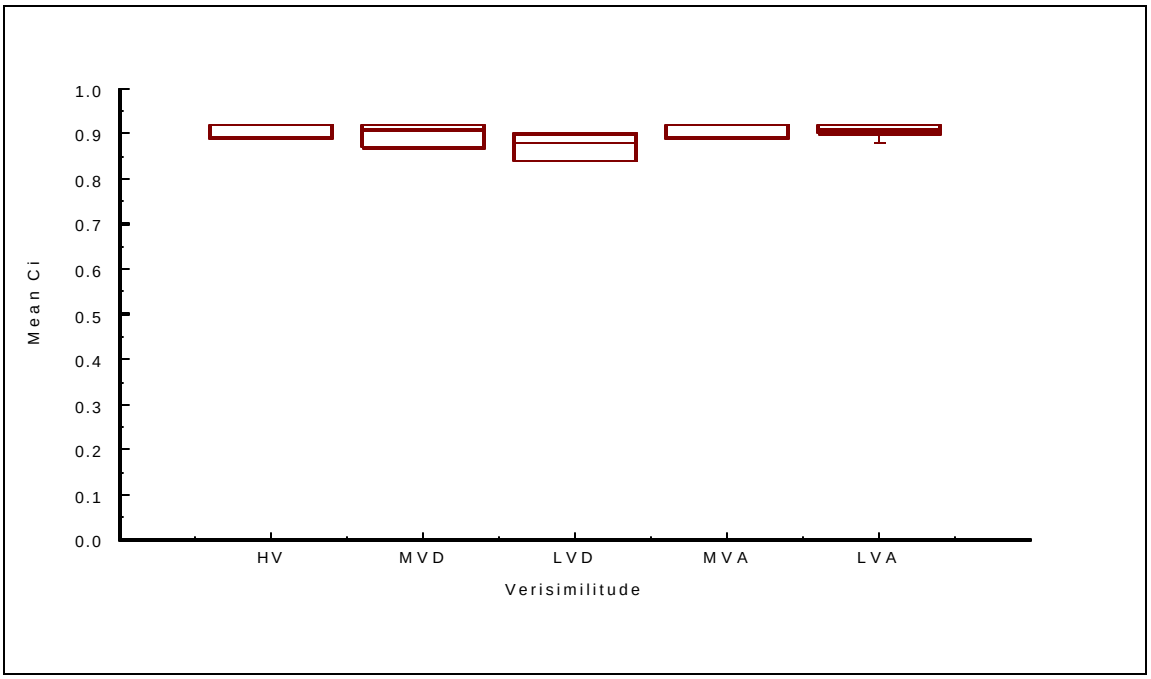

Figure 33. Mean C_i by Level of Verisimilitude, Interval Prediction.

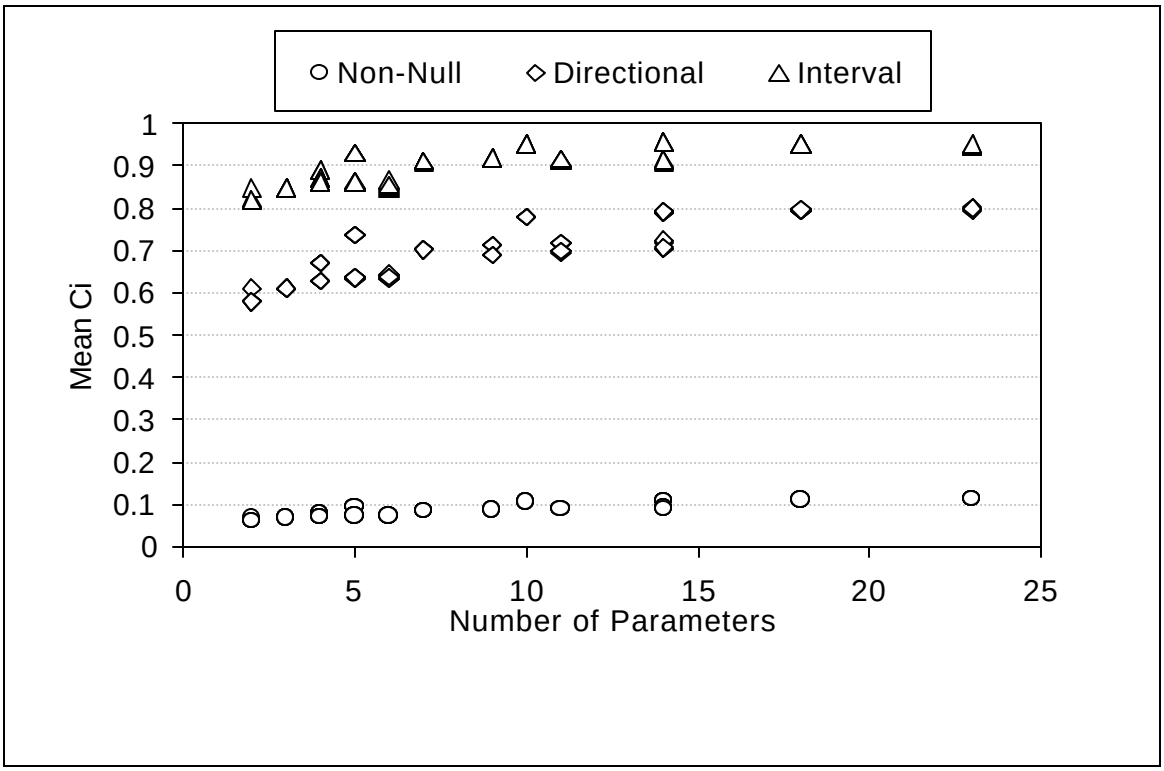

Figure 34. Mean C_i by Level of Intolerance and Number of Estimated Paths.

Relationship between the Standard Deviation of C_i and Model Complexity, Collinearity, and Sample Size

An initial examination of the magnitude of the standard deviation of C_i and the central design factors revealed very little variability across conditions. These values are organized by the central design factors and are presented in Tables 17-19. Figure 35 provides a box and whisker plot of the distribution of the standard deviation of C_i across all 270 conditions examined. The standard deviation of mean C_i was estimated to range from <.01 to 0.02. In only two conditions, did the estimated standard deviation of C_i obtain a magnitude greater than .01, with both of these conditions occurring in the low complexity models under the most severe condition of model misspecification (i. e., low verisimilitude). This striking lack of variability suggests that no practically significant relationship exists between the central design factors and the standard deviation of C_i.

Table 17

Standard Deviation of C_i by Intolerance, Verisimilitude, Collinearity, and Sample Size Model Complexity = Low

Verisimilitude	Collinearity	Sample Size	Standard Deviation of C_i		
			Non Null	Directional	Interval
HV	Low	100	<.01	<.01	<.01
MVD			<.01	<.01	<.01
LVD			<.01	<.01	<.01
MVA			<.01	<.01	0.01
LVA			<.01	0.01	0.01
HV		200	<.01	<.01	<.01
MVD			<.01	<.01	<.01
LVD			<.01	<.01	<.01
MVA			<.01	<.01	0.01
LVA			<.01	0.01	0.01
HV		500	<.01	<.01	<.01
MVD			<.01	<.01	<.01
LVD			<.01	<.01	<.01
MVA			<.01	<.01	<.01
LVA			<.01	<.01	0.01
HV	Moderate	100	<.01	<.01	<.01
MVD			<.01	<.01	0.01
LVD			<.01	<.01	0.01
MVA			<.01	<.01	0.01
LVA			<.01	0.01	0.02
HV		200	<.01	<.01	<.01
MVD			<.01	<.01	<.01
LVD			<.01	<.01	0.01
MVA			<.01	<.01	0.01
LVA			<.01	0.01	0.02
HV		500	<.01	<.01	<.01
MVD			<.01	<.01	<.01
LVD			<.01	<.01	<.01
MVA			<.01	<.01	<.01
LVA			<.01	0.01	0.01

Table 18

Standard Deviation of C_i by Intolerance, Verisimilitude, Collinearity, and Sample Size, Model Complexity = Moderate

Verisimilitude	Collinearity	Sample Size	Standard Deviation of C_i		
			Non Null	Directional	Interval
HV	Low	100	<.01	<.01	<.01
MVD			<.01	<.01	<.01
LVD			<.01	<.01	<.01
MVA			<.01	<.01	<.01
LVA			<.01	<.01	0.01
HV		200	<.01	<.01	<.01
MVD			<.01	<.01	<.01
LVD			<.01	<.01	<.01
MVA			<.01	<.01	<.01
LVA			<.01	<.01	<.01
HV		500	<.01	<.01	<.01
MVD			<.01	<.01	<.01
LVD			<.01	<.01	<.01
MVA			<.01	<.01	<.01
LVA			<.01	<.01	<.01
HV	Moderate	100	<.01	0.01	<.01
MVD			<.01	<.01	<.01
LVD			<.01	<.01	<.01
MVA			<.01	<.01	<.01
LVA			<.01	<.01	<.01
HV		200	<.01	<.01	<.01
MVD			<.01	<.01	<.01
LVD			<.01	<.01	<.01
MVA			<.01	<.01	<.01
LVA			<.01	<.01	<.01
HV		500	<.01	<.01	<.01
MVD			<.01	<.01	<.01
LVD			<.01	<.01	<.01
MVA			<.01	<.01	<.01
LVA			<.01	<.01	<.01

Table 19

Standard Deviation of C_i by Intolerance, Verisimilitude, Collinearity, and Sample Size, Model Complexity = High

Verisimilitude	Collinearity	Sample Size	Standard Deviation of C_i		
			Non Null	Directional	Interval
HV	Low	100	<.01	<.01	<.01
MVD			<.01	<.01	<.01
LVD			<.01	<.01	<.01
MVA			<.01	<.01	<.01
LVA			<.01	<.01	<.01
HV		200	<.01	<.01	<.01
MVD			<.01	<.01	<.01
LVD			<.01	<.01	<.01
MVA			<.01	<.01	<.01
LVA			<.01	<.01	<.01
HV		500	<.01	<.01	<.01
MVD			<.01	<.01	<.01
LVD			<.01	<.01	<.01
MVA			<.01	<.01	<.01
LVA			<.01	<.01	<.01
HV	Moderate	100	<.01	<.01	<.01
MVD			<.01	<.01	<.01
LVD			<.01	<.01	<.01
MVA			<.01	<.01	<.01
LVA			<.01	<.01	<.01
HV		200	<.01	<.01	<.01
MVD			<.01	<.01	<.01
LVD			<.01	<.01	<.01
MVA			<.01	<.01	<.01
LVA			<.01	<.01	<.01
HV		500	<.01	<.01	<.01
MVD			<.01	<.01	<.01
LVD			<.01	<.01	<.01
MVA			<.01	<.01	<.01
LVA			<.01	<.01	<.01

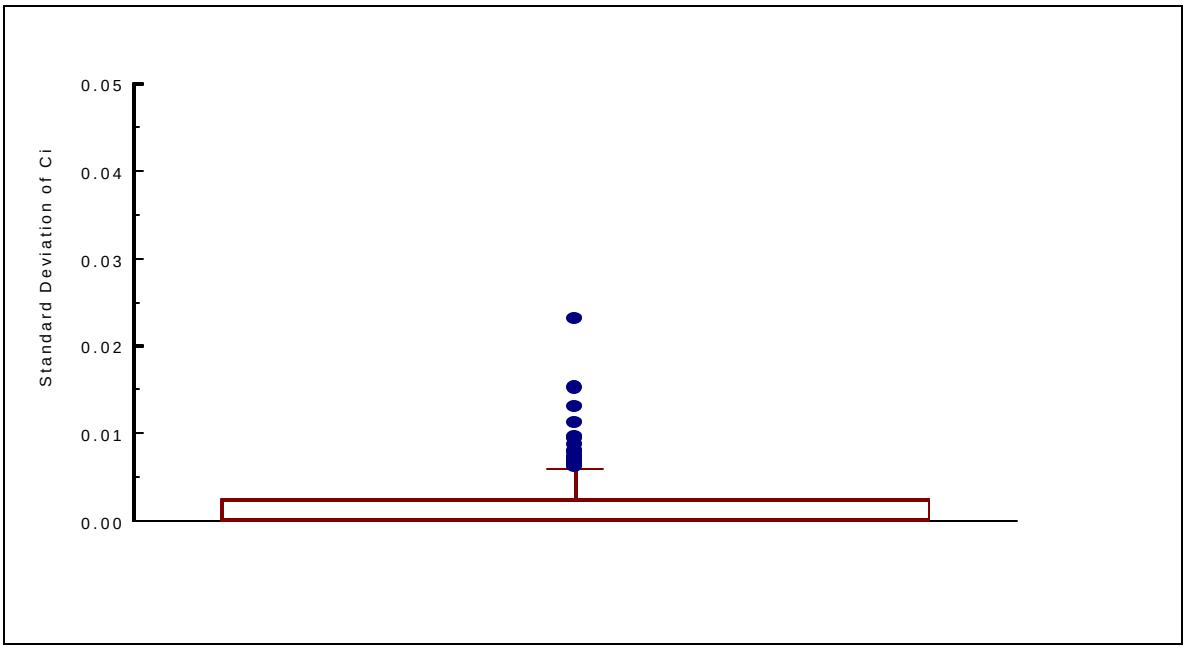

Figure 35. Box and Whisker Plot of Estimated Standard Deviations.

Probing Deeper: An Examination of the Variability in Path Coefficients

To investigate the lack of variability evidence by mean C_i, the variability of the average expected path coefficients, was examined by calculating the standard error. These analyses were conducted for each of the 'true' models, for each level of model complexity and collinearity. The results are displayed as a stem and leaf plot in Figure 36. An examination of the distribution of resultant standard errors revealed a moderate degree of variability, however it was observed that more than half of the standard errors were estimated to be less than .10. Further examination of these results revealed that the more complex models evidenced more sampling error than the moderate and simple models. Standard errors of this magnitude, across so many of the conditions examined, led many of the estimated path coefficients to fall within the specified tolerance

interval. The lack of deviation from the tolerance interval resulted in multivariate closeness estimates that approached 1.0 across most of the conditions examined. Given an invariant intolerance estimate for each condition, the within-cell variability of C_i was entirely dependent on the estimate of closeness and failed to vary appreciably across samples.

Stem	Leaf
.36	00
.34	
.32	
.30	
.28	
.26	00
.24	00000
.22	
.20	000000
.18	000000
.16	000000
.14	0000000000000000
.12	000000
.10	00000000000000000000000000
.08	0000000000000000
.06	000000000000000000000000000
.04	00000000000000000000
.02	000000000000000

Figure 36. Stem and Leaf Plot of Standard Errors of Regression Coefficients.

Relationship between Mean C_i, Precision of Prediction and Verisimilitude

To examine the relationship between mean C_i, the precision of prediction, and verisimilitude, it was first necessary to collapse the data across the other design factors (i.e., model complexity, sample size, and colli nearity) and compute marginal values of mean C_i. The results of this analysis are presented in Figure 37. Examination of this figure once again reveals the profound influence of the precision of prediction, and the negligible influence of verisimilitude or "truth likeness". For each level of verisimilitude we see a dramatic increase in mean C_i as the level of precision increases. For example, with high verisimilitude (HV), the mean C_i is only approximately .09 for the non null prediction, yet reaches .91 for the interval prediction. However, if we look across the various level of verisimilitude the obtained values of mean C_i vary very little. For the non null conditions, mean C_i was observed to range from .08 to .09, whereas for the directional condition, mean C_i ranges from .65 (LVD) to .71 (for both MVA and LVA). For the most precise predictions, mean C_i ranged from .87 (LVD) to .91 (HV, MVA, and LVA).

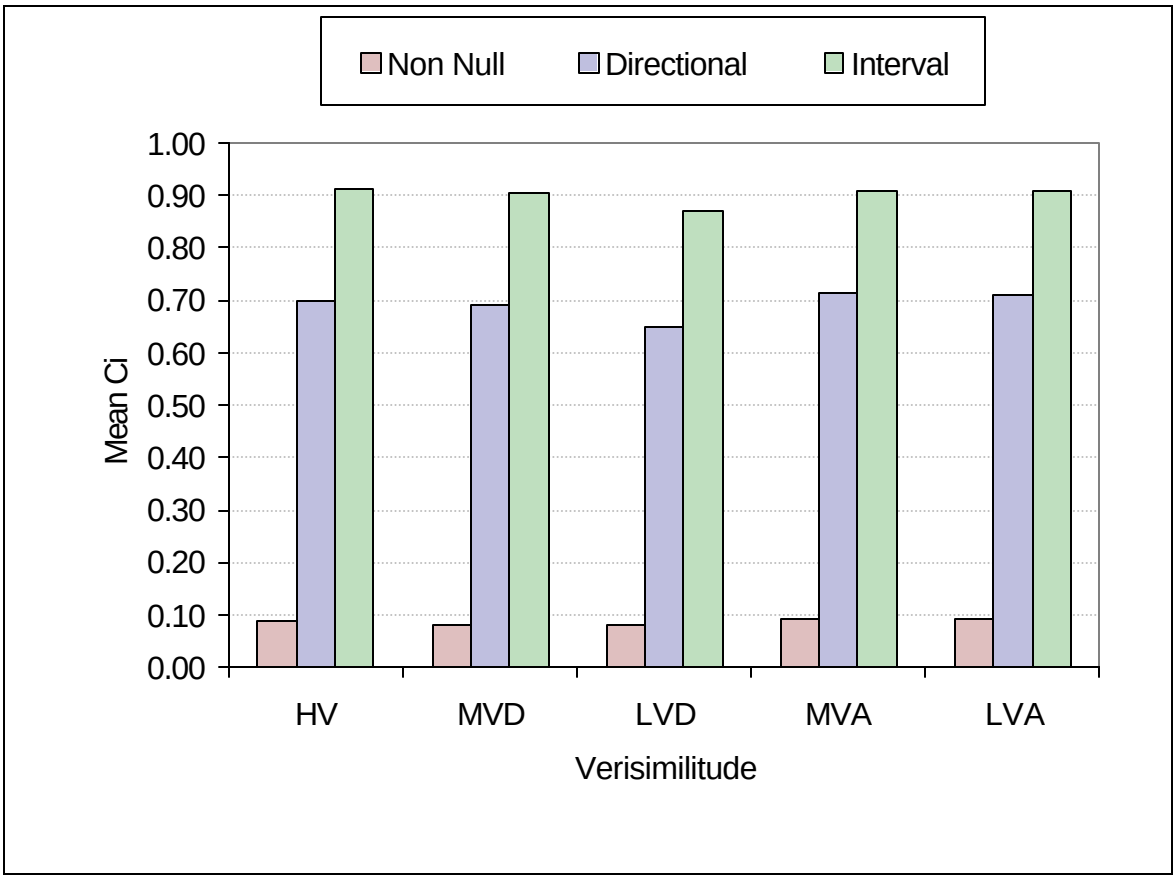

*Figure 37***.** Mean C_i by Level of Intolerance and Verisimilitude.

Relationship between Mean C_i, Precision of Prediction and Model Complexity

To examine the relationship between mean C_i, precision of prediction, and model complexity, it was again necessary to compute marginal values of mean C_i by collapsing across the other central design factors (i.e., verisimilitude, sample size and collinearity). The results of this analysis are displayed in Figure 38. Once again the overwhelming influence of precision of prediction is depicted, while model complexity appears to exert but a slight influence on the magnitude of mean C_i.

For example, for the simplest model containing four variables, mean C_i was estimated to be .08 for the weakest prediction, .66 for the directional prediction and .88 for the interval prediction. However, within each level of intolerance, the mean C_i only evidenced a slight fluctuation with increased model complexity. A modest increase in mean C_i was evidenced with the set of directional predictions, with mean C_i ranging from approximately .66 for the simplest model to approximately .72 for the most complex model.

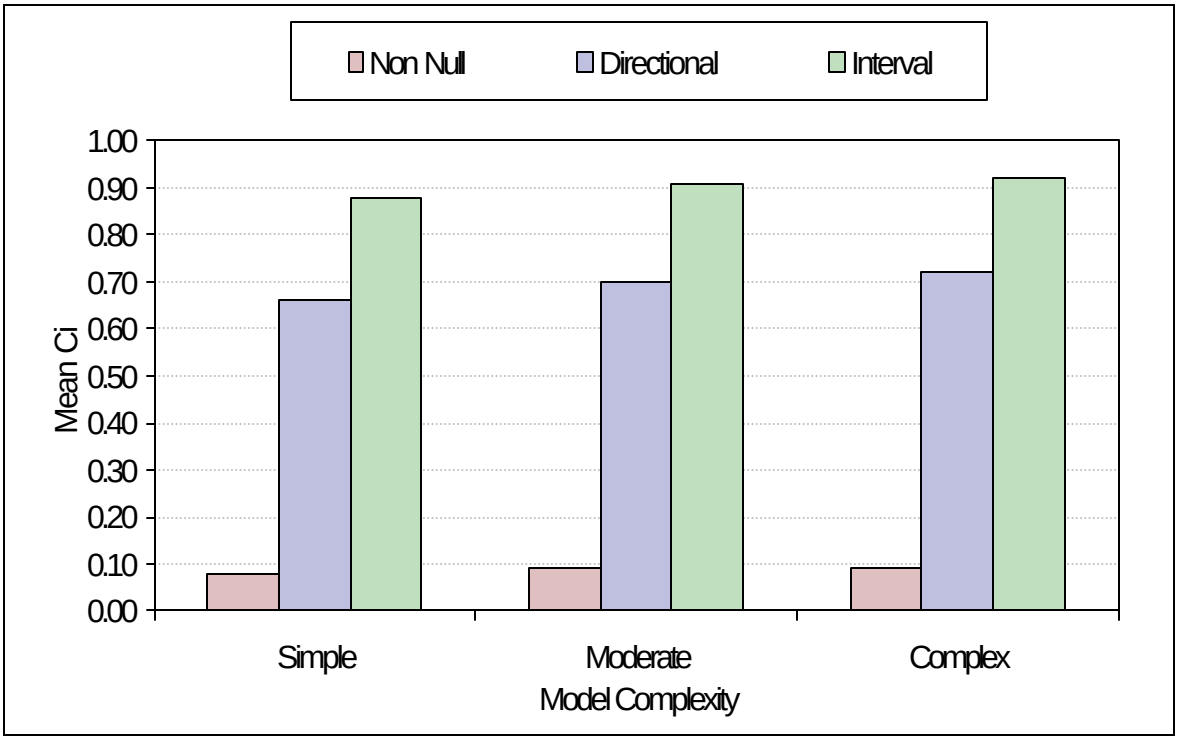

Figure 38. Mean C_i by Level of Intolerance and Model Complexity.

In order to determine if verisimilitude might be a potential moderating variable, an examination of the relationship between mean C_i, precision of prediction and truth-likeness was examined for each level of model complexity. The relationships among these central design factors are illustrated in Figures

39-41. As evidenced in these figures, level of intolerance continued to exert a considerable influence on mean C_i, however, there was a slight increase in mean C_i with an increase in model complexity. For example, for the LVD model, mean C_i increases from .84 for the low complexity model to .90 for the high complexity model. Still, the more dramatic increases were observed for this model across level of intolerance, as mean C_i increased from .07 for the non null prediction to .84 for the interval prediction.

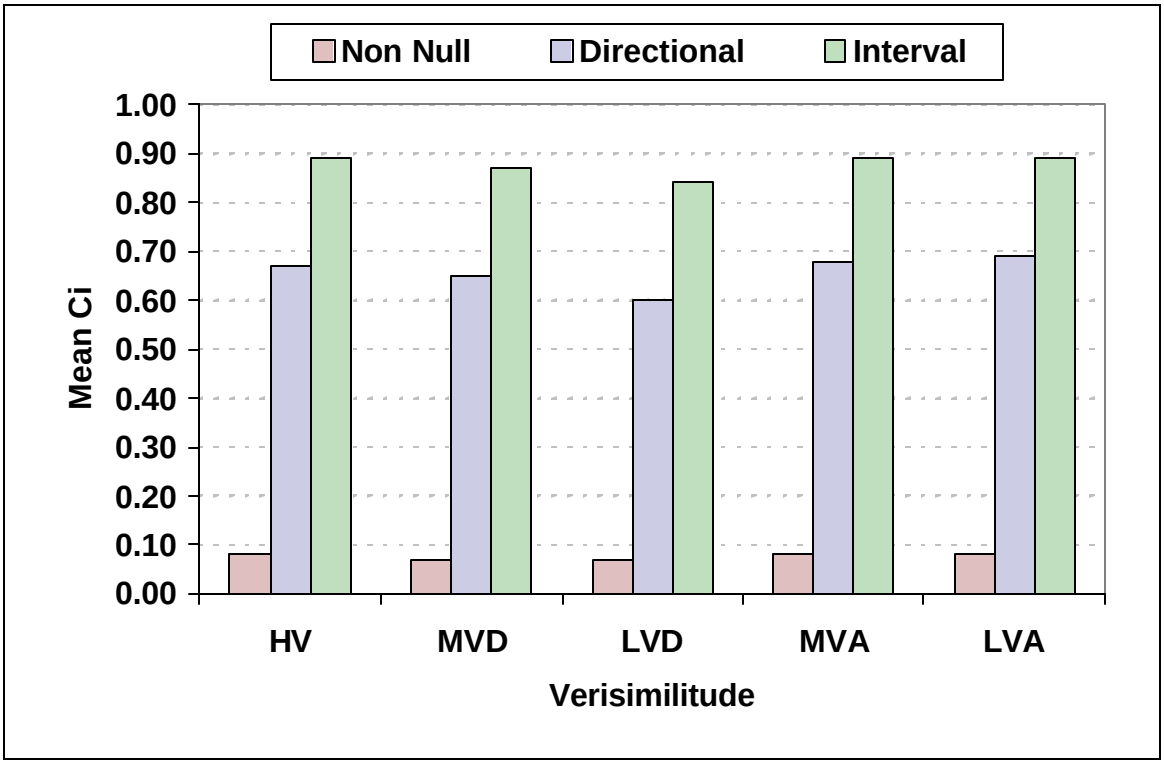

Figure 39. Mean C_i by Level of Intolerance and Verisimilitude, Model Complexity = Low.

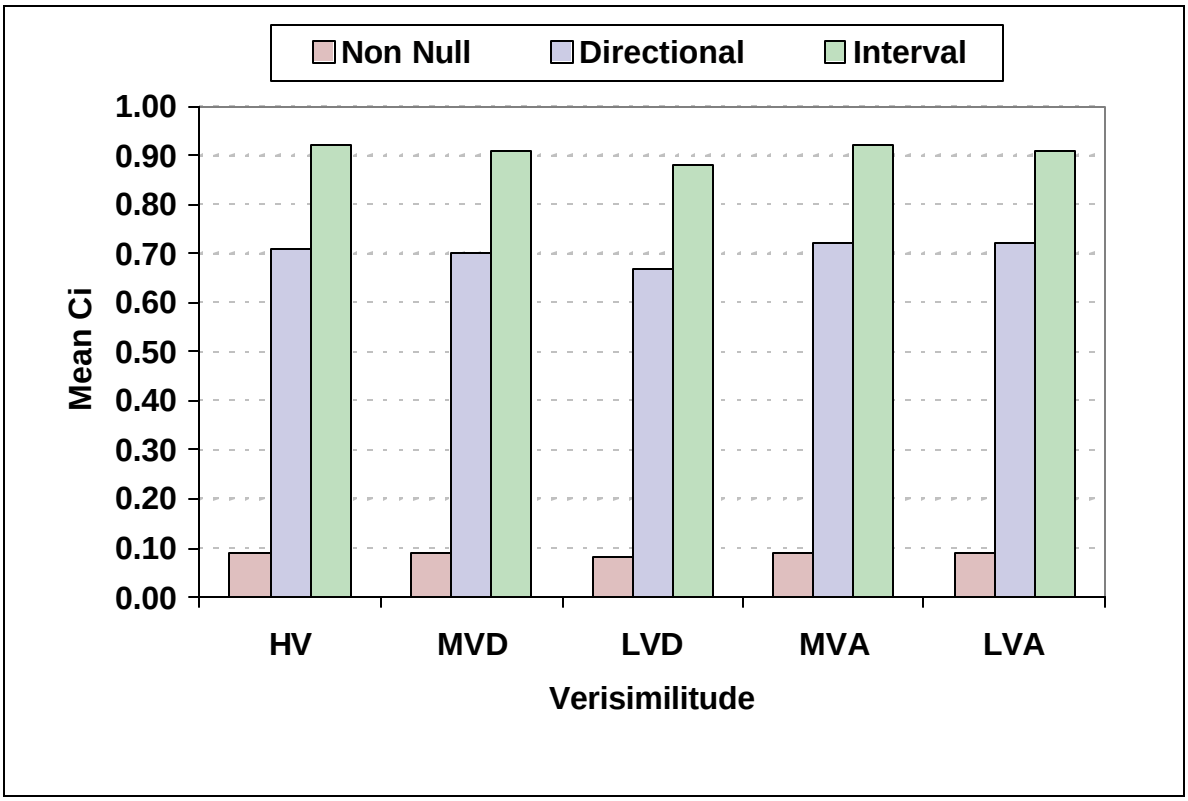

Figure 40. Mean C_i by Level of Intolerance and Verisimilitude, Model Complexity = Moderate.

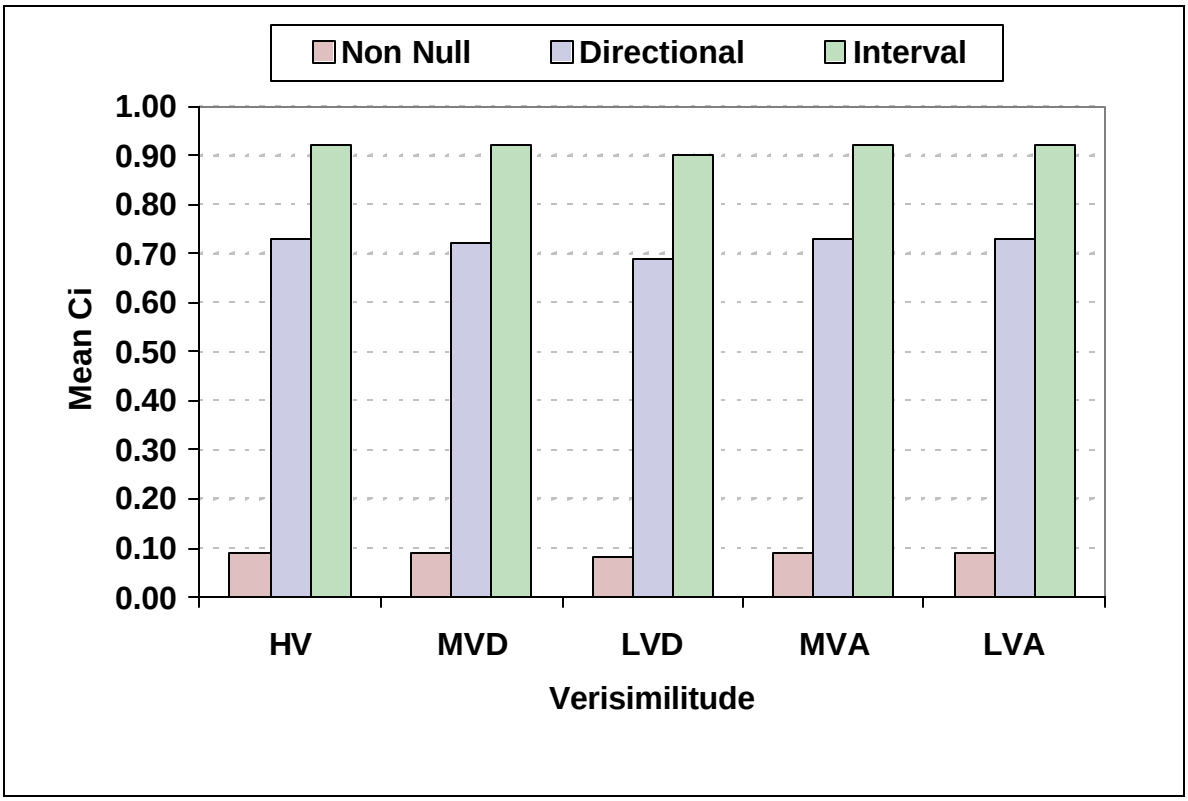

Figure 41 Mean C_i by Level of Intolerance and Verisimilitude, Model Complexity = High.

Probing Deeper: An Examination of Bias Evidenced in the Expected Path Coefficients

As the relationship between mean C_i and the precision of prediction was expected to be moderated by the level of verisimilitude, a series of additional analyses was conducted in an attempt to discern why this relationship was not evidenced in the data. Essentially, an examination of this nature can be considered to be synonymous with assessing the level of bias in the obtained standardized path coefficients. That is, the deviation of the expected sample path coefficients from the population parameters. As closeness is the element of the multivariate corroboration index that captures these deviations in the data, it seemed appropriate to compare the average expected values of multivariate closeness to the estimates obtained from the population. These results were examined by level of model complexity, verisimilitude, collinearity and level of intolerance and are provided in Table 20. As these results suggest, there is very little deviation from "truth" in these data, resulting in multivariate closeness estimates of 1.00 for 80% of the conditions examined. Minor deviations from 1.00 most frequently occurred when making interval predictions, and were relatively consistent across level of model complexity. With a negligible amount of bias, and closeness approximating 1.0, the resultant component of verisimilitude failed to emerge as a salient factor across most of the conditions examined.

Table 20
Expected Multivariate Closeness by Verisimilitude, Intolerance, Model Complexity and Collinearity

		Level of Collinearity					
		Low			Moderate		
Model Complexity	Verisimilitude	Non null	Directional	Interval	Non null	Directional	Interval
4	HV	1.00	1.00	1.00	1.00	1.00	1.00
	MVD	1.00	1.00	1.00	1.00	1.00	1.00
	LVD	1.00	1.00	1.00	1.00	1.00	1.00
	MVA	1.00	1.00	0.99	1.00	1.00	0.99
	LVA	0.99	1.00	0.99	1.00	0.99	0.98
6	HV	1.00	1.00	1.00	1.00	1.00	1.00
	MVD	1.00	1.00	1.00	1.00	1.00	1.00
	LVD	1.00	1.00	1.00	1.00	1.00	0.98
	MVA	1.00	1.00	1.00	1.00	1.00	1.00
	LVA	0.99	1.00	0.99	0.99	1.00	0.99
8	HV	1.00	1.00	1.00	1.00	1.00	1.00
	MVD	1.00	1.00	1.00	1.00	1.00	1.00
	LVD	1.00	1.00	1.00	1.00	1.00	1.00
	MVA	0.99	1.00	0.99	1.00	1.00	0.99
	LVA	0.99	1.00	0.99	0.99	1.00	0.99

Relationship between Mean C_i, Precision of Prediction, Collinearity and Sample Size

The relationship between mean C_i, precision of prediction, collinearity and sample size is illustrated in Figure 42. An examination of this figure reveals the striking lack of variability across the three samples sizes and two levels of collinearity examined in this study. Once again, mean C_i evidence a dramatic increase as the intolerance level of the theory increased. For each of the non null predictions, mean C_i was estimated to be approximately .09, evidencing no variability across the various levels of verisimilitude, model misspecification and sample size, regardless of whether the degree of verisimilitude was a function of adding paths or deleting paths from these models. In the case of directional predictions, the average C_i value was estimated to be approximately .70 for each level of verisimilitude or "truth-likeness" and sample size. Similarly, for the most precise, or interval prediction, the average C_i did not evidence any substantial degree of variability across level of collinearity or sample sizes. Of course, rapid acceleration of the average C_i was evident as predictions became increasingly more precise or risky. While non null predictions resulted in an estimated average C_i of only .09, precise predictions were rewarded with an average C_i of approximately .90, regardless of the level of collinearity or size of the sample.

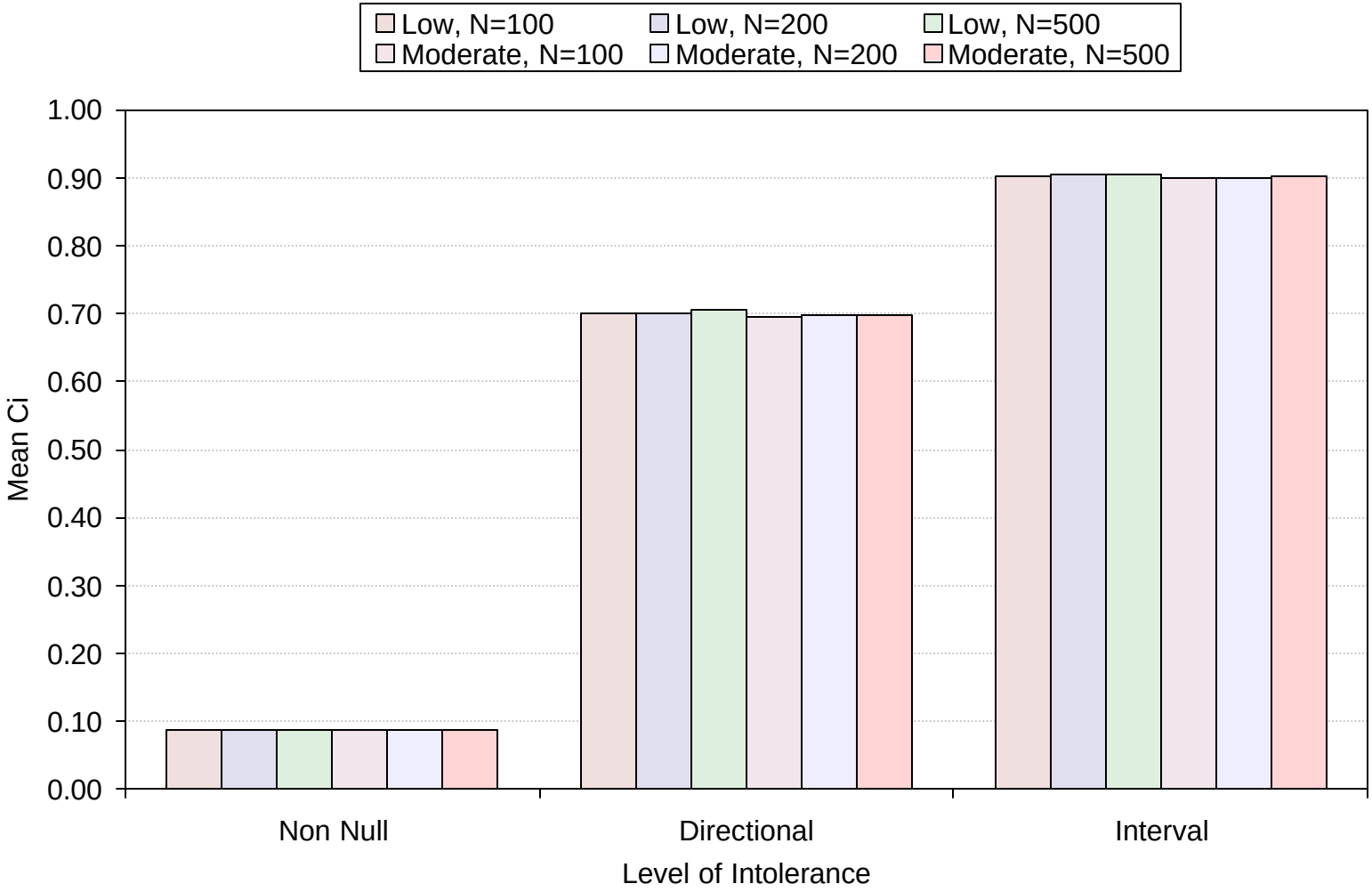

Figure 42. Mean Ci by Level of Intolerance, Collinearity, and Sample Size.

Summary

An examination of the relationship between the estimated mean C_i and the central design factors revealed the overwhelming influence of the precision of prediction. In all cases, the level of intolerance was observed to dramatically influence the magnitude of the estimated mean C_i, regardless of the design factor under consideration. This conclusion is consistent with the findings of the past empirical research when the behavior of the corroboration index was examined in both univariate and bivariate contexts.

Factors that were anticipated to moderate the relationship between mean C_i and the precision of prediction, such as verisimilitude, model complexity and collinearity failed to exert any substantive influence. These unexpected results led to the necessity to probe deeper into the data in an attempt to better understand these disconcerting results. These additional analyses revealed the negligible influence of sampling error and an inherent lack of statistical bias in the models under investigation.

Chapter Five

Conclusions, Implications, and Recommendations

Organization

The purpose of this chapter is to provide a set of sound conclusions that are firmly grounded in the results of this study, the findings of past empirical research and the body of literature that coheres around the central theme of theory testing in the social sciences. Readers are reminded about the controversy surrounding theory appraisal in general and reacquainted with the research problem and the primary purpose of the study. A brief synopsis of the method is also provided. The results are then discussed with respect to each research question and hypothesis. Following this recapitulation of the major findings of the study, important implications for practice and theory are advanced. The chapter concludes with recommendations for future research.

There is little argument that the elucidation and empirical testing of theories are important components of research in any field. Yet despite the long history of science, the extent to which theories are supported or contradicted by the results of empirical research remains ill defined. Quite commonly, support or contradiction is based solely on the "reject" or "fail to reject" decisions that result from tests of null hypotheses that are derived from aspects of theory. Decisions and recommendations based on this forced and often artificial dichotomy have been scrutinized in the past. In recent years, such an overly simplified approach to theory testing has been challenged on logical grounds (Meehl, 1997, 1990, 1978; Serlin & Lapsley, 1985). Theories differ in the extent to which they provide

precise predictions about observations. The precision of predictions derived from theories is proportional to the strength of support that may be provided by empirical evidence congruent with the prediction. However, the notion of precision linked to strength of support is surprisingly absent from many discussions regarding the appraisal of theories.

Statement of the Problem

Meehl (1997,1990a) has presented a logically sound index of corroboration to summarize the extent to which empirical tests of theories provide support or contradiction of those theories. However, the numerical properties of this index have not been investigated beyond some of the most basic predictions about population mean differences, zero order, and first-order partial correlations (Hogarty & Kromrey, 2002, 2001, 2000).

Purpose

The purpose of this study was to build upon the previous research by advancing to the next logical step through the evaluation of the utility of the corroboration index and its behavior when appraising theories employing path analytic methods in the context of social science research. Many researchers approach path analysis by beginning with a model in which there is substantial confidence. This confidence may stem from either theoretical or substantive reasoning about the linkages between the variables under investigation. Less attention, however, is typically given to estimating the magnitude of such

linkages. Most areas of psychology do not permit a high degree of precision. According to Blaich (1998), quasi-quantitative predictions of rough magnitudes of effects could help advance the field. Therefore, the primary focus of this investigation was on the precision in the prediction of the magnitude of effects and an examination of factors that moderate the relationship between corroboration and precision.

Method

A Monte Carlo study was conducted to investigate the utility of a multivariate corroboration index in the appraisal of theories employing path analytic methods. A series of simulations was conducted that related theoretical predictions to empirical results. The study was designed in the context of hypothetical theories, the cores of which predict a single outcome from various configurations of exogenous and endogenous variables. The resulting path coefficients were the parameter estimates of primary interest.

Five factors were manipulated in these simulations: level of verisimilitude (low, moderate, high); level of intolerance (non null, directional, and interval predictions); model complexity (low, moderate, and high); level of collinearity (low and moderate); and sample size (100, 200, and 500). The level of verisimilitude was determined by the proportion of estimated paths that were either added to or deleted from a 'true' model. Level of intolerance, or precision of prediction, was determined by the width of the tolerance interval (i.e., 95%, 50%, or 25% of the Spielraum). The levels of collinearity were selected to reflect a condition with a low level of collinearity, that would not tend to indicate a deleterious influence

with respect to the analyses and results, and a moderate level of collinearity (VIF=1.5 and VIF=3.0, respectively). The sample sizes selected for this study represent values that range from those that might be considered insufficient (e.g., N=100 for the high complexity models) to those that might be considered more than adequate (e.g., N=500 for both the low and moderate complexity models).

For this study, six population correlation matrices were constructed based on a specified number of variables, level of collinearity, and true path model. 10,000 multivariate normal samples were generated from each population correlation matrix. Sample correlation matrices were then constructed and analyzed using a series of regression equations simulating path analysis. Standard deviations were then calculated for each of the 270 conditions. Lastly, the obtained values of C_i resulting from each path analysis were pooled, and the average value of C_i was evaluated in the context of the central design factors.

Relationship between Mean C_i, Verisimilitude, Intolerance, Model Complexity, Collinearity, and Sample Size

For the analysis of the relationship between mean C_i and the central design factors in the study, a factorial ANOVA was conducted, treating the average C_i as a dependent variable. The independent variables in this analysis were the five central design factors. Omega-squared was used to estimate the proportion of variance accounted for in the population by each effect, as well as each of the interaction effects. Somewhat surprisingly, only a single factor, level of intolerance, emerged to explain more than half of the variance in C_i. The lack

of influence of any of the other factors led to the decision to conduct another set of analyses, this time, controlling for the level of intolerance. Two factors that were responsible for explaining a portion of the variance in C_i emerged only for the lowest level of intolerance (i.e., the non null condition). With the exception of the influence of the number of variables and level of verisimilitude for these non null predictions, the results suggest that after such a large portion of the variance in C_i was attributed to the level of intolerance, the other central design factors were unable to account for a noticeable amount of the variance.

Relationship between the Standard Deviation of C_i, Model Complexity, Collinearity, and Sample Size

An examination of the standard deviation of C_i revealed a striking lack of variability across all of the conditions examined. These results are relatively consistent with previous empirical findings. Despite the lack of relationship between the standard deviation and model complexity, it was expected that the standard deviation would be influenced by the level of collinearity and sample size. That is, we would expect to see more stability, and hence less variation, in C_i as sample sized increased. Further, the level of collinearity, determined by the variance inflation factor (VIF) was expected to influence the variability of the obtained standardized path coefficients. This influence would also be expected to translate into less stable estimates of C_i.

An examination of the magnitude of the standard errors of the path coefficients helped to shed some light on this puzzling finding. As expected, the magnitude of the standard errors, which represents the typical difference

between $\hat{\beta}$ and β, was observed to decrease as sample size increased. Further, the more complex models evidence more sampling error than the moderate or simple models. However, in many cases, the magnitude of the standard error was not large enough to cause the estimated path coefficients to fall outside of a given tolerance interval. This lack of deviation from the tolerance interval resulted in multivariate closeness estimates of approximately one across most of the conditions examined. Because the value of intolerance is constant for any given condition, the within-cell variability of C_i was primarily dependent on the variability in closeness, and hence the finding that C_i did not vary appreciably across samples.

Relationship Between Mean C_i, Precision of Prediction, Model Complexity, and Level of Collinearity

When the relationship between mean C_i, precision of prediction, model complexity and level of collinearity was explored, once again the overwhelming influence of the precision of prediction was noted. Mean C_i evidenced a dramatic increase in magnitude as the precision of the prediction increased. Within each level of collinearity and sample size, the magnitude of mean C_i remained stable. Although the relationship between mean C_i and precision of prediction was not anticipated to be moderated by sample size, the absence of the influence of collinearity was somewhat surprising.

Relationship Between Mean Ci, Precision of Prediction, and Verisimilitude

The relationship between mean C_i and the precision of prediction was expected to be moderated by the closeness of the data to the theory (i.e., verisimilitude). As this expected relationship was not evidence in the data, the level of bias in the obtained standardized path coefficients was investigated. These analyses were conducted by level of model complexity, verisimilitude, collinearity and level of intolerance.

Initially, the deviation of the expected sample path coefficients from the population parameters was estimated. As stated earlier, it is the closeness element of the multivariate corroboration index that captures these deviations in the data; therefore it seemed prudent to investigate the average expected values of multivariate closeness. As the resultant bias was negligible, multivariate closeness estimates approached the upper limit of 1.00 consistently across the conditions examined. These results help to elucidate the finding that verisimilitude failed to play much of a role in these results.

Relationship between Mean C_i and Precision of Prediction

Based upon previous empirical research, the precision of prediction was expected to exert a considerable influence on the magnitude of mean C_i. It was anticipated that this relationship would be substantively stronger than the relationship between mean C_i and verisimilitude, model complexity, collinearity, and sample size. In a typical condition in which a weak or non null prediction was made, very little corroborative evidence was observed, however, the

advancement of a directional prediction offered considerable improvement. As expected, the most precise or interval predictions yielded the greatest amount of corroboration. Consistent with past empirical findings (Hogarty & Kromrey, 2002, 2001, 2000), the results of this study revealed the profound influence of intolerance, which provided a ceiling for the magnitude of C_i, regardless of the other design factors examined in the study.

Implications for Theory and Practice

The introduction of a corroboration index was not intended to supplant the use of significance tests in general. Surely, null hypothesis testing has its place. In many of the situations that confront applied researchers, it is vital to distinguish findings that are likely due to chance and those that are not. Often, tests of statistical significance are employed as a starting or entry point in an investigation, prior to embarking on further analyses. In this vein, the use of null hypothesis testing is used as a type of screen, providing insight regarding how to proceed with additional analyses. And, in many cases, tests of statistical significance are used simply because potentially viable alternatives are not superior or available.

Still, there is little doubt that abuses of statistical significance testing are abundant. The results of tests of statistical significance provide limited information that is often misused and misinterpreted. In certain disciplines, for example advertising or marketing, the use of statistical testing is misleading given the inherent nature of the sampling methods employing such as quota,

convenience, or large but perhaps not truly representative mail samples. Another commonly held notion is that studies that do not include asterisks are flawed. Perhaps the most compelling reason to avoid such over reliance on tests of statistical significance is that it precludes us from thinking about solving problems and addressing research questions in a different way. Further, there are certainly a large array of methodological tools that do not rely on experimental designs. The use of mixed methods is but one of the emerging viable alternatives to strict adherence to the null hypothesis way of knowing.

As a complement to tests of statistical significance there has been a renewed emphasis toward requiring the reporting of effect sizes along with results of hypothesis tests. Further, more attention is now being given not only to point estimates of effects, but also the degree of confidence that we can place on these estimates and hence an emphasis on the reporting of confidence bands. We as researchers should always remain mindful of the arsenal of tools at our disposal as we search for answers to important questions and seek to discover the nature of the relationships that exist within the complex social systems that we investigate. The methods that might drive educational leaders in their effort to uncover the antecedents to high turnover among teachers might be the very same methods that business leaders apply to the study of factors related to satisfaction in the workplace.

An important shift in the business of theory appraisal should involve the comparison of alternative theories and models rather than comparisons of outcomes to the null hypothesis. In many disciplines there exists a complex and

overlapping array of social systems that beg for methods and tools that can serve as an enhancement to traditional analytic methods, rather than alternatives or preferred substitutes.

The sheer logic of appraising a scientific theory is often more complicated than some would believe (Meehl, 1997). In addition to the aforementioned argument regarding the precision of prediction (that is, a precise prediction that is supported by the data warrants more logical evidence of support than does a weak prediction supported by the data), the movement from theory into an empirical test necessitates the incorporation of many logical components besides the theory itself. Meehl (1997) presents these components as elements of an equation:

$$\left(T \bullet A_x \bullet C_p \bullet A_i \bullet C_n\right) \rightarrow \left(O_1 \supset O_2\right)$$

Where T = the theory being "tested,"

A_x = Auxiliary theories relied upon during the conduct of the research.

C_p = *Ceteris paribus* (all other things being equal),

A_i = Instrumental theories related to measures and controls employed,

C_n = Realized particulars (the extent to which the research was actually conducted as we think it was), and

$O_1 \supset O_2$ = the material conditional "if you observe O_1, you will observe O_2."

That which is subject to empirical test is not the theory alone, but the amalgam of these elements. Data that appear to contradict a "theory" may arise because of errors anywhere in this combination of elements. Emphasis should be given and attention focused on the influences of the other factors in this amalgam.

Recommendations for Future Research

Path analysis continues to enjoy widespread use in the appraisal of theories in many disciplines. Although an arsenal of fit indices are available to aid researchers in assessing the tenability of an estimated model, these indices lack a critical components that gives consideration to the precision of the prediction under investigation. Meehl (1990a) contends the way in which a theory accumulates "money in the bank" is by passing several stiff tests; claiming that "the main way a theory gets money in the bank is by predicting facts that, absent the theory, would be antecedently improbable" (p 115). A theory's merit is a matter of degree, rather than a yes or no question, as it is treated in null hypothesis testing (Meehl, 1990a). Theoretical support depends on a variety of factors, including the relative uniqueness of the prediction, how surprising the prediction is, the precision of prediction, and degree of correspondence between the prediction and the observed data (Nickerson, 2000).

The conditions examined in this study were chosen based on the types of situations that applied researchers would be expected to encounter in the conduct of a traditional path analysis. The inclusion of three levels of model complexity and three sample sizes seemed to be reasonable representations of situations that are commonly confronted. The three tolerance intervals, or levels of intolerance were fairly representative of the strength of predictions that are typically advanced in the literature. That is, we would not be surprised to observe researchers making non null or directional predictions, even though we would hope to see even more precise predictions. For this reason, a rather

liberal tolerance interval was chosen for the most precise predictions, rather than choosing a narrower interval or point prediction that would not be an reasonable reproduction of reality. The correlation matrices that were created for these analyses were also chosen with care. Collinearity is a major threat in this type of correlational analysis and hence was featured as one of the central design factors in this study. Larger values of collinearity were explored but not included due to the deleterious influence that more redundancy in the data would be expected to exert. The level of verisimilitude, or model misspecification, seemed to be in line with what an applied researcher might be expected to come across in the investigation and/or comparison of a number of competing models. Given the inherent nature of these models, and their seemingly reasonable conditions, it is discomforting that the multivariate corroboration index as currently formalized was not often successful in distinguishing between misspecified models, and models varying in complexity and collinearity.

There is considerable evidence that suggests that the current formulation of this multivariate index of corroboration needs to be reexamined. The overwhelming influence of the precision of prediction suggests that alternative representations of multivariate intolerance should be considered in order to a ensure a more appropriate balance between the two components that combine to measure corroboration. Further, the inability of this index of corroboration to distinguish between situations in which weak predictions that are correct warrant the same degree of corroboration gleaned from precise predictions that are not correct is troublesome. To illustrate these discrepant findings let us consider

both ends of the spectrum, that is, just how false things could possibly get, or the worst case scenario for C_i, and the best case scenario for a given condition. Consider a low complexity supplementary model, with the lowest level of verisimilitude, and a precise prediction (i.e., 4 variables, 2 supplementary paths, interval prediction). The obtained C_i for this model would be .95, whereas the obtained value of C_i for the comparable model with high verisimilitude would be .99. As this example illustrates, the current formulation of the multivariate index of corroboration does not do a good job of detecting model misspecification. Further, if we consider conditions in which a greater degree of deviation from truth is possible, across all three levels of intolerance (that is, obtained path coefficients of –1.0 when truth was estimated to be .30) the resultant indices of corroboration would range from .08 for the non null prediction to .35 for the directional prediction to .43 for the interval prediction. Of course, deviations this large while not evidenced in the data for this study, might be expected if this work was to be replicated with a different set of correlation matrices and more extreme conditions.

Although the shortcomings of the index in its current form cannot be disputed, the importance of theoretical intolerance as a determinant of degree of corroboration was once again brought to light, underscoring the need for the development of precise theories in the social sciences. The results from this study suggest that efforts to develop theories in the social sciences and related disciplines that enjoy greater precision of prediction may concomitantly provide critical tests with greater potential for corroboration.

In this study, the closeness component, or measure of verisimilitude appeared to behave badly, and hence did not serve to inform the multivariate corroboration index of the true nature of the data across many of the conditions. It would seem prudent to investigate the behavior of this index with a modified index of closeness. A different formulization of this component might include a different conceptualization of the relationship between the intolerance interval and the Spielraum. One modification of this component is to reflect the distance of an obtained estimate with respect to how large the deviation really is versus relative to the maximum possible distance. That is, $Cl = 1 - \left(\frac{D}{S - I} \right)$ rather than the original univariate conceptualization of $Cl = 1 - \left(\frac{D}{S} \right)$. This modification should reduce the tendency for closeness to be overstated and result in a more accurate reflection of multivariate corroboration. Future work on the conceptualization of both the components of closeness and intolerance, as well as an adjustment for sample size, is currently under consideration. Interpretations of “risk”, utility and the performance of the index under varied conditions will be sought from philosophers of science and applied researches to aid in the reconceptualization of the index. It is anticipated that this inquiry will lead to a set of recommendations regarding the use of the index, the degree of risk that represents a risky prediction given the context of the research being conducted and other potential uses of the index.

Given the shortcoming of the index and the need for further investigation, applied researchers are cautioned against using this index of corroboration in its

current form. However, an appropriately modified index of corroboration may serve a variety of different functions across a variety of disciplines and contexts. The index might serve as one indicator among a host of properties or indices that are predictors of the success of a theory's long-term fate (Faust & Meehl, 2002). Given the presence or absence of other properties, an index that examines predictive accuracy in relation to risk might be given more or less weight. If we consider a collection of desirable traits that might include parsimony, novelty, risk, qualitative diversity or breadth and elegance of mathematical beauty, an index of corroboration might be considered a minor player. However, absent some of these more desirable properties, precision and 'truth-likeness' may carry a more formidable amount of weight.

A reformulated multivariate corroboration index may be applied in the planning of empirical studies as well as for the interpretation of research results. Its utility may extend beyond univariate, bivariate and traditional path analysis as more sophisticated methods such as Hierarchical Linear Modeling (HLM) and Structural Equation Modeling (SEM) enjoy more widespread attention and use. Ideally, in addition to the precision of prediction (i.e., intolerance), the index would be sensitive to factors such as sample size, model misspecification or verisimilitude, and model complexity. Its use should serve to move the arguments surrounding theory testing away from the testing of null hypotheses into a consideration of the complexity of the research context, the degree of "risk" entailed by the theory's predictions, and the extent to which the obtained data (absent the theory) represent a "damn strange coincidence."

References

Alvin, D.F., & Hauser, R.M. (1975). The decomposition of effects in path analysis. *American Sociological Review, 30*, 37-47.

American Psychological Association. (1991). Publication Manual (5th ed.), Washington, DC.

Asher, H. B. (1983*). Causal Modeling (2nd ed.).* Beverly Hills: Sage.

Belsley, D. A. (1984). Reply. *The American Statistician*, *3*8, 90-93.

Bentler, P.M., & Bonnet, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588-606.

Bentler, P.M. (1983). Some contributions to efficient statistics for structural models: Specification and estimation of moment structures. *Psychometrika*, *48*, 493-571.

Bentler, P.M. (1989). *EQS structural equations program manual*. Los Angeles: BMDP Statistical Software.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238-246.

Bentler, P.M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

Blaich, C. F. (1998). The null-hypothesis significance test procedure: Can't live with it, can't live without it. *Behavioral and Brain Sciences, 21*, 194-195.

Bollen, K. A. (1986). Sample size and Bentler's and Bonett's nonnormed fit index. *Psychometrika*, *51*, 375-377.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research, 17,* 303-316.

Boomsma, A. (1982). The robustness of LISREL against small sample size in factor analysis models. In K. G. Joreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction* (Part 1, pp. 149-173). Amsterdam: North-Holland.

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

Campbell, D.T. (1990). The Meehlian corroboration-verisimilitude theory of science. *Psychological Inquiry*, *1*, 142-147.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378-399.

Chow, S. L. (1990). In defense of Popperian falsification. *Psychological Inquiry*, *1*, 147-149.

Cudeck, R., & Browne, M.W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research*, *18*, 147-167.

Dar, R. (1987). Another look at Meehl, Lakatos, and the scientific practices of psychologists. *American Psychologist, 47*, 145-151.

Duncan, O. D. (1975). *Introduction to Structural Equation Models*. New York: Academic Press.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling, 6*, 56-83.

Faust, D., & Meehl, P.E. (2002). Using meta-scientific studies to clarify or resolve questions in the philosophy and history of science. *Philosophy of Science*, *69*, S185-S196.

Finch, S., Cummings, G., & Thomason, N. (2001). Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement, 61*, 181-210.

Finney, J. M. (1972). Indirect effects in path analysis. *Sociological Methods and Research, 1*, 175-186.

Fraas, J. W., & Newman, I. (1994). A binomial test of model fit. *Structural Equation Modeling, 3*, 268-273.

Gerbing, D. W., & Anderson, J.C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models in educational research. In K.A. Bollen & J.S. Long (Eds.), Testing structural equation models (pp.40-65). Newbury Park, CA: Sage.

Gholson, B. & Barker, P. (1985). Kuhn, Lakatos, and Laudan: Applications in the history of physics and psychology. *American Psychologist*, *40*, 755-769.

Gigerenzer, G. (1998). Surrogates for theories. *Theory and Psychology*, *8*, 195-204.

Hanson, N.R. (1958). *Patterns of discovery; an inquiry into the conceptual foundations of science.* Cambridge, England: University Press.

Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Hoelter, J.W. (1983). The analysis of covariance structures: Goodness-of-fit indices. *Sociological Methods and Research*, *11*, 325-344.

Hogarty, K. Y. & Kromrey, J. D. (2000, April). *Risky predictions and damn strange coincidences: An initial consideration of Meehl's Index of Corroboration*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Hogarty, K. Y. & Kromrey, J. D. (2001, April). *Corroboration and Coincidence: The Moderating Effect of Statistical Complexity on the Relationship Between Research Design Factors and Meehl's* $C_{i.}$ Paper presented at the annual meeting of the American Educational Research, Seattle.

Hogarty, K.Y. & Kromrey, J.D. (2002, February). *What's N got to do with it? A modification of Meehl's Index of Corroboration.* Paper presented at the annual meeting of the Eastern Educational Research Association, Sarasota.

Hu, L., & Bentler, P. M. (1999). Cut off criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-54.

Humphreys, L.G. (1990). View of a supportive empiricist. *Psychological Inquiry*, *1*, 153-155,

Joreskog, K.G., & Sorbom, D. (1981). *LISREL V: Analysis of linear structural relationships by the method of maximum likelihood. Chicago: National Educational Resources.*

Joreskog, K.G., & Sorbom, D. (1984). *LISREL VI user's guide (3rd ed.)*. Mooresville, IL: Scientific Software.

Kerlinger, F. N. (1964). *Foundations of behavioral research*. New York: Holt, Rinehart, and Winston.

Kirk, R. E. (Ed.). (1972). *Statistical Issues*. Monterey, CA: Brooks/Cole.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Lakatos, I. (1970). Falsification and the methodology of scientific research programs. In I. Lakatos & A. Musgrave (Eds.) *Criticism and the growth of knowledge* (pp. 91-196). Cambridge, England: Cambridge University Press.

Lauden, L. (1977). *Progress and its problems.* Berkeley: University of California Press.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151-159.

MacCallum, R. C., Browne, M.W., & Preacher, K. J. (2002). Comments on the Meehl-Waller (2002) procedure for appraisal of path analysis models. *Psychological Methods*, *7*, 301-306.

MacCallum, R. C. Wegener, B. N., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin, 114*, 185-199.

Markus, K. A. (2002). The converse inequality argument against tests of statistical significance. *Psychological Methods*, *7*, 147-160.

Marsh, H.W., Ball, J.R., & McDonald, R.P. (1988). Goodness-of-fit indices in confirmatory factor analysis: Effects of sample size. *Psychological Bulletin*, *103*, 391-411.

Maxwell, S. E. & Delaney, H. D. (1990). *Designing experiments and analyzing data*. Belmont, CA: Wadsworth.

McDonald, R.P., & Marsh, H.W. (1989). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin*, *107*, 247-255.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806-834.

Meehl, P. E., (1990a). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1,* 108-141.

Meehl, P. E., (1990b). Author's response. *Psychological Inquiry, 1,* 173-180.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numeral predictions. Harlow, Lisa, Ed; Mulaik, Stanley, Ed; and Steiger, James, ED. *What if there were no significance tests?* p. 393 – 425.

Meehl, P.E., & Waller, N.G. (2002). The path analysis controversy: A new statistical approach to strong appraisal of verisimilitude. *Psychological Methods, 7*, 283-300.

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy: A reader.* Chicago, Aldine.

Mulaik, S. A. (2002). Commentary on Meehl and Waller's (2002) path analysis and verisimilitude. *Psychological Methods*, *7*, 316-322.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241-301.

Olsson, U., Troye, S. V., & Howell, R. D. (1999). Theoretical fit and empirical fit: The performance of maximum likelihood versus generalized least squares estimation in structural equation models. *Multivariate Behavioral Research, 34*, 31-58.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction (3rd ed.).* Fort Worth: Harcourt Brace College Publishers.

Popper, K. R. (1959). *The logic of scientific discovery*. New York: Basic.

Reichardt, C. S. (2002). The priority of just-identified recursive models. *Psychological Methods*, *7*, 307-315.

Roberts, S., & Pashler, H. (2002). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, 358-367.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton, N.J.: Princeton University Press.

Serlin, R. C. & Lapsley, D. K. (1985). Rationality in psychological research: The good enough principle. *American Psychologist, 40*, 73-83.

Steiger, J. H., & Lind, J.C. (1980, May). *Statistically based tests for the number of common factors.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Tanaka, J.S. (1993). Multifaceted conceptions of fit in structure equation models with latent variables. In K.A. Bollen & J.S. Lond (Eds.), *Testing structural equation models (*pp. 136-162). Newbury Park, CA: Sage.

Tanaka, J.S., & Huba, G.J. (1985). A fit index for covariance structural models under arbitrary GLS estimation. *British Journal of Mathematics and Statistical Psychology*, *42*, 233-239.

Thompson, B. (2002, April). *What is AERA, anyway?* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Thompson, B., & Daniel, L.G. (1996). Factor analytic evidence for the construct validity of scores: An historical overview and some guidelines. *Educational and Psychological Measurement*, *56*, 213-224.

Tucker, L.R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1-10.

Waller, N. G., & Meehl, P.E. (2002). Risky tests, verisimilitude, and path analysis. *Psychological Methods*, *7*, 323-337.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.

About the Author

Kristine Y. Hogarty is a native New Yorker who moved to Florida to begin her career at the University of South Florida after serving four years in the United States Marine Corps. She was awarded a Bachelor of Science degree in Business Management in 1991, and a Master of Arts degree in Criminology in 1994. During her tenure at USF she has served in many roles, including undergraduate advisor, teaching assistant, research associate and research consultant. Kris is currently the Coordinator of Research and Assessment Systems for the College of Education.

Kris's primary research interests include applied statistics and data analysis. For the past eight years, Kris has had the good fortune to collaborate with faculty members and fellow doctoral students across a variety of disciplines on a host of research projects and grants. These opportunities and experiences have frequently resulted in presentations at regional, national, and international conferences. A number of these collaborative works have also been published in peer-reviewed journals such as the *Journal of Research on Computing in Education*; *Behavior Research Methods, Instruments and Computers*; *Multiple Linear Regression Viewpoints*; *Psychometrika;* and *Educational and Psychological Measurement*.

Last but not least, Kris is an ardent animal lover and a huge fan of palm trees, pelicans, sunsets and sandy beaches.