ROBUSTNESS TO NON-INDEPENDENCE AND POWER OF SAWILOWSKY'S I TEST FOR TREND IN CONSTRUCT VALIDITY

by

JOHN L. CUZZOCREA

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2007

MAJOR: EDUCATIONAL EVALUATION
AND RESEARCH

Approved by:

_____
Advisor Date

_____

_____

_____

_____

UMI Number: 3264012

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

ACKNOWLEDGMENTS

My experience at Wayne State has enriched my life, not just from the perspective of the knowledge that I have attained, but also from the people that I have had the privilege of meeting. First and foremost, I would like to take this opportunity to extend a special thank you to my advisor, Dr. Shlomo Sawilowsky. His encouragement and support throughout was indispensable. He has routinely gone out of his way to assist me in navigating through the various pitfalls I have encountered throughout the program. Dr. Sawilowsky is a tremendous advocate for his students and I have a great deal of respect for him in equal parts as both a person and an academic. I have been fortunate enough to get to know Dr. Sawilowsky on a personal level as the editorial assistant of his journal and the humor, candor, benevolence, wisdom, and expertise he embodied will forever serve as a model for me in my professional life.

Additionally, I need to recognize the contributions of both Dr. Gail Fahoome and Dr. Donald Marcottte in shaping my educational experience. Both individuals embody the characteristics indicative of a superior educator and I have enjoyed being able to get to know each of them on a more personal level outside of the classroom. Dr. Sandra Williams also merits recognition for agreeing to participate on my dissertation committee. My educational experience has benefited from their participation and input, for which I am grateful.

This list of acknowledgments would not be complete without special consideration to my family that has shouldered much of the sacrifice involved in completing the doctoral program. To my wife, Melissa, whom I have grown up with from the time that we dated in high school to the joy we experienced together in the birth of our children, Alexa and Cole. She is truly the love of my life and I am forever grateful for the patience and support she provided, without which I would never have been able to realize my educational aspirations. She is intelligent and compassionate and has been a source of wisdom, stability, and guidance. I am a better and happier person because of her and I am forever thankful that she is a part of my life.

To my children, Alexa and Cole, their exuberance and enthusiasm for the world around them served as a source of inspiration for me. I am driven by the desire to serve as a role model for my children and they have kept me grounded and focused on my most important role in this life, being a father.

Finally, I am grateful to my parents, Leo and Angela Cuzzocrea, having instilled in me a desire to pursue my dreams and encouraged me in the pursuit of those dreams along the way. They have continually sacrificed their own dreams and aspirations to help ensure my success. Their sacrifices and support will forever be remembered and have been a major factor in my desire to complete this program. As well, I would be remiss if I did not mention my in-laws, Michael and Diane Dwyer, whom I regard as my second set of parents. They have been there through my successes and my failures and have always been a source of support and guidance in my life.

TABLE OF CONTENTS

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Prior to the 1950s, the literature in the field of behavioral and social science measurement dealt almost exclusively with predictive and content validity at the exclusion of the third and perhaps most important type, construct validity (Cronbach, 1971; Nunnally, 1978). The concept of construct validity was born out of the recommendations put forth by the joint committees of the American Psychological Association and the American Educational Research Association (Subsequently, they were joined by the National Council on Measurement in Education). The committee distinguished between three types of measurement validity: a) content, b) criterion (predictive and concurrent), and c) construct (Cronbach, 1971). Content validity is a measure of the degree to which the content of a test reflects the material involved in a particular task from which the conclusions will be drawn. Predictive validity is defined as the degree to which a test is able to predict future performance of some behavior or characteristic by correlating the test score with the criterion variable. Finally, construct validity is loosely defined as the degree to which a test is able to measure a particular construct variable.

Interestingly, the majority of research in the realm of psychology or education is based in large part on the study of constructs such as self-esteem, anxiety, intelligence, and self-determination. These are variables that cannot be measured directly because they are abstract, subtle, and often latent. Instead,

they are at best inferred from a collection of observable measurements. As an example, intelligence is gauged through a variety of observable measurements that may include problem-solving, reading comprehension, and spatial reasoning.

Due to its importance in behavioral and social science research, tests aimed at measuring constructs require evidence of validity to provide users with a degree of confidence that the tests are measuring what they purport to measure. Nearly half a century ago, Campbell and Fiske (1959) developed the Multitrait-Multimethod Matrix as a means of analyzing convergent and divergent validity. Analysis of the matrix is hinged on the concept that the greater the degree of convergent and discriminant validity; the greater the evidence of construct validity.

Their approach has had a tremendous impact in the realm of behavioral research, as evidenced by over 2,000 citations over the next 33 years (Sternberg, 1992). Despite the attention that it has garnered over the years, the matrix remains troubled by the same issues that plagued it when Campbell and Fiske developed it. According to Sawilowsky (2002), the “interpretation of the matrix is subjective … (and) not amenable to straightforward interpretation” (p.78).

Over the years, there have been numerous attempts to apply various statistical analyses to the matrix as a means of removing subjectivity (Stanley, 1961; Huber & Baker, 1978; Jöreskog, 1971). However, these approaches are not without their own set of difficulties ranging from the complexity of the procedures to restrictive assumptions that are difficult to satisfy (Schmitt & Stults,

1986; Widaman, 1985). Frustrated by the lack of progress in developing a measure for the Multitrait-Multimethod Matrix, Fiske and Campbell (1992) stated, “Our article had impact because it raised a problem: the links between psychological methods and psychological constructs. The problem is still with us 33 years and more than 2,000 citations later.” (p.394).

In response to the difficulties met in developing a procedure to properly analyze the Multitrait-Multimethod Matrix, Sawilowsky (2002) developed a distribution-free test for trend that contributes evidence of construct validity. The test statistic, I, examines the number of inversions in the Multitrait-Multimethod Matrix; contrary to the desired trend. It is based in part on the requirement made by Jackson (1969) that proper analysis must incorporate the entire correlation matrix. The I statistic revolves around the concept of ascending trend from the Heterotrait-Heteromethod values to the Reliability diagonal. The Heterotrait-Heteromethod values are the correlations between traits that are measured by different methods. The Reliability diagonals are the correlations that measure the internal consistency of the matrix. As a result, if there are relatively few inversions proceeding upward along the matrix, then the null hypothesis would be rejected in favor of the alternative of an increasing trend, which supports construct validity.

According to Sawilowsky (2002), the I statistic combines the logic of Jonckheere’s distribution-free k-sample test against ordered alternatives (Jonckheere, 1954) with the counting function of Mann’s test for randomness in a single sample (Neave & Worthington, 1988). The main advantage of the I statistic

over the Jonckheere test is that it is relatively easy to compute. As well, the Jonckheere test may be computed many different ways, thereby generating different sets of p-values which could produce different results.

The I statistic is a response to the challenge put forth by Fiske and Campbell (1992) that despite the attention their original article has received over the years (see Campbell & Fiske, 1959), the problem of developing a statistical approach to evaluate convergent and discriminant validities remains. Sawilowsky (2002) developed the I statistic as a test of construct validity that incorporates the Multitrait-Multimethod Matrix developed by Campbell and Fiske (1959).

Statement of the Problem

The study was an examination of both the impact of violating the independence assumption in terms of the Type I error rate and a power analysis of the I test. The I statistic is a distribution-free test and does not have the distribution assumptions that plague a parametric approach to evaluating the matrix. However, it is still susceptible to the independence assumption and although the risk of violating this assumption is minimized by using a limited number of the values in the matrix (i.e. minimum, median, and maximum coefficients with a three-point I statistic), the risk of violating this assumption increases as the number of values used in the test increases (i.e. four-point I statistic). Therefore, the concept of violating independence is concerned primarily with the increasing number of data points in the matrix; if they are increased, then the probability of violating independence is also increased. Although the risk

of violating the independence assumption is increased with an increased number of data used in the analysis, it is not known whether a violation of independence will impact the Type I error rate.

As a result, a modified version of the Sawilowsky I test will be created to incorporate more data points. The three-point I statistic is comprised of four groups, representing the different facets of the Multitrait-Multimethod Matrix, with three values in each (i.e., minimum coefficient, median coefficient, and maximum coefficient). A modified four-point version of the I statistic will encompass four data points at each level of the matrix (minimum coefficient, lower quartile, upper quartile, and maximum coefficient). Both versions of the I statistic will be examined to determine the impact upon each when independence has been violated. The study will also examine the power properties of both the three-point and four-point versions of the test to determine if an increasing number of data points will (comparing the three-point version to the four-point version) will lead to greater power.

Significance of the Problem

Human beings are complex organisms and meaningful research into their behavior thereby involves the study of construct variables that, although not directly observed or measured, account for a large part of our understanding. It is for this reason that the vast majority of the literature in behavioral research, namely the fields of psychology and education, has been preoccupied with the study of construct variables. In light of this fact, research involving construct

variables require the use of tests that can be used to provide an accurate measure of such variables. Thus, a test constructed as a measure of a specific construct variable must exhibit a high degree of construct validity if it is to be used.

It is of no wonder that the Campbell and Fiske (1959) article has been cited so often through the years, given the importance placed upon the study of construct variables. In fact, the Multitrait-Multimethod Matrix revealed in this article has become the current gold standard in interpreting construct validity. The Multitrait-Multimethod Matrix provided an approach to determining convergent or discriminant validity, which in turn may be used to either support or refute a claim for construct validity. Unfortunately, a statistical approach to evaluating the matrix was not provided and interpretation was limited to a subjective examination requiring users to eyeball the matrix. Fiske and Campbell (1992) were cognizant of this weakness and expressed their frustration that very little had been accomplished in terms of resolving this issue.

The I statistic was developed as a means of interpreting the Multitrait-Multimethod Matrix from a statistical perspective. Other statistical approaches have been developed to interpret the matrix, however the I statistic does not share the same restrictive underlying assumptions that are difficult to meet in an applied setting. Although further study is required with regard to the robustness and power properties of this statistic, it does hold promise as an alternative approach to interpreting the Multitrait-Multimethod Matrix.

Definition of Terms

Concurrent Validity

A measure of the degree to which a test is able to determine current performance of some behavior or characteristic by correlating the test score with the criterion variable. This type of validity would be useful in determining if a particular test could be substituted for another test that employed a different approach (i.e. substituting a multiple choice test for a fill in the blank test).

Construct

According to Nunnally (1978) it is defined as “the extent to which a variable is abstract rather than concrete…Such a variable is literally a construct in that it is something that scientists put together from their own imaginations, something that does not exist as an isolated, observable dimension of behavior“(p.96). Examples of constructs include variables such as anxiety, self-esteem, intelligence, etc. These variables cannot be directly measured or observed; instead they are manifested in a number of observables. The number of observables increases in relation to the complexity of the construct.

Construct Validity

The degree to which a test is able to measure a particular construct variable. Because constructs are manifested in any number of observables, construct validity is measured in degrees dependent upon the number of

observables that account for the overall domain of the construct that are reflected in the test. Also referred to as trait validity and factorial validity (Nunnally, 1978).

Convergent Validity

The degree to which a measure correlates with other observables that are assumed to measure the same construct. A high correlation is an indicant of construct validity.

Criterion

The term is used in relation to predictive and concurrent validity, whereby the performance on a test is correlated to the criterion variable to determine the degree of validity. According to Cronbach (1971), the criterion variable is “an external variable considered to provide a direct measure of the characteristic or behavior in question” (p.444).

Discriminant Validity

The degree to which a measure fails to correlate with other observables that are assumed to measure a different construct. A low correlation is an indicant of construct validity.

Heterotrait-Heteromethod Values

A component of the Multitrait-Multimethod Matrix that examines correlations between a number of traits and a number of methods of

measurement. They are the sections of the matrix that border the validity diagonal.

Heterotrait-Monomethod Values

A component of the Multitrait-Multimethod Matrix that examines correlations between a number of traits and one method of measurement. They are the sections of the matrix that border the reliability diagonal.

Independence

An assumption of the I statistic, stipulating that the scores used in testing the Multitrait-Multimethod Matrix are independent of one another.

Method

The process employed in obtaining scores on a particular trait.

Multitrait-Multimethod Matrix

A method for determining convergent and discriminant validity as a means of evaluating construct validity. The Multitrait-Multimethod Matrix consists of intercorrelations of more than one trait measured by more than one method and it is comprised of the following components: a) reliability diagonal, b) validity diagonal, c) heterotrait-monotrait block, and d) hetertrait-heteromethod block.

Observables

The term refers to the underlying traits that make up the overall domain of the construct being evaluated.

Power

An indicator used to determine the ability of a test to detect a false null hypothesis, assuming that the null hypothesis is in fact false

Reliability

The degree to which an instrument provides consistent measurements

Reliability Diagonal

A component of the Multitrait-Multimethod Matrix which may also be referred to as the monotrait-monomethod values. The values along the diagonal are measures of internal reliability.

Validity

The process of examining whether an instrument provides an accurate measure of what it purports to measure. It is not the test itself that is validated; rather, it is the usage of the test.

Validity Diagonal

A component of the Multitrait-Multimethod Matrix which may also be referred to as the monotrait-heteromethod values. The values along the diagonal are correlations between different measures of the same trait.

Trait

A characteristic or attribute of the subject that is measured through the use of an instrument or method.

Type I Error

An error made in interpreting the results of a particular test. The null hypothesis is rejected, when it should have been accepted. This leads the researcher to arrive at the conclusion that a treatment is effective, when that is not the case.

Limitations of the Study

A modified version of the three-point I statistic will be created that is composed of four values at each level of the matrix (minimum coefficient, lower quartile coefficient, upper quartile coefficient, and maximum coefficient). A five-point version of the I statistic (minimum coefficient, lower quartile, median, upper quartile, and maximum coefficient) could also have been incorporated into the study; however, the table of critical values for the five-point test was not available

(Jonckheere, 1954) and it was therefore not included in the analyses. Future research may wish to examine the feasibility of a five-point version of the I statistic.

CHAPTER 2

REVIEW OF LITERATURE

Reliability and validity are both of key concern in discussing the quality of any measuring instrument. Although reliability may be achieved without having validity, it is not possible to achieve validity without first determining reliability. The data collected by way of a test are known as observed scores and they are used as a reflection of the true scores of a subject. Reliability is relatively simple to evaluate by means of a reliability coefficient. A high reliability coefficient is to be interpreted as a reliable instrument.

However, reliability does not ensure that the instrument is measuring what it purports to measure. As a result, it is not adequate to determine, solely, the reliability of a test. Instead, validity must also be demonstrated in order for the test to be interpreted properly. Validity is a key feature of any measuring instrument, in that it is important to assess if the usage of a test is valid before a determination may be made regarding its usefulness in scientific research.

An examination of what is meant by the term validity is a necessary first step in understanding how to determine and improve validity. According to Nunnally (1978), a general definition for validity would be that “a measuring instrument is valid if it does what it is intended to do” (p.86). Cronbach (1971) offered a more specific definition by stating that “validation is the process of examining the accuracy of a specific prediction or inference made from a test score” (p.443). In fact, if one was to peruse through the literature, a multitude of

definitions would be found, each with a slightly different interpretation. Ebel (1961) argued that, “No exact scientist would accept such diverse statements as operationally useful definitions of the same quantitative concept” (p.640). The inability of researchers to agree on a concrete definition for the term is a fundamental problem with the concept of validity (Ebel, 1961).

The issue has been further clouded by the fact that evaluating validity is never as simple as making a final determination of either valid or invalid. Instead, validity is measured in degrees along a spectrum. In fact, there is “no way to prove the validity of an instrument purely by appeal to authority, deduction from a psychological theory, or any type of mathematical proof” (Nunnally, 1978, p.87). Unlike the reliability coefficient, which allows for a straightforward interpretation, there is no singly accepted validity coefficient. Rather, evidence is gained in support of validity through an “accumulation of empirical, statistical, theoretical, and conceptual evidence” (Suen, 1990, p.134).

To add to this dilemma, determining the validity of the use of an instrument is a never ending process whereby a specific usage of a test that is currently deemed to be valid is not guaranteed for some future application of the test. There is always the possibility that new evidence may surface that would render the application of an instrument to a different usage as being invalid, whereby the current instrument would either have to be modified or set aside in favor of a different instrument. The observables that make up the domain of constructs important to behavioral research are continually being challenged and revised as determined through a growing body of empirical evidence. Thus,

evaluating validity has become a science unto itself with significant journal space devoted to the different approaches that have been devised over the years to deal with such problems.

These issues were reflected in the findings of Ebel (1961), who provided evidence of twenty tests independently reviewed for evidence of validity. It was found that only one had adequate evidence of validity; despite the fact that almost every test specialist would agree that validity is the overriding concern of any mental test (Ebel, 1961). Using the definition of validity provided by Nunnally (1978), if an instrument does not measure what it is supposed to measure, then it is not possible to draw conclusions from studies that have employed such tests.

Types of Validity

There are three primary types of validity which are currently recognized in the realm of behavioral research. A joint committee composed of both the American Psychological Association and the American Educational Research Association outlined the three types of validity in the published report entitled, Technical Recommendations (as cited in Cronbach, 1971; Cronbach & Meehl, 1955; Ebel, 1961). The different types of validity included in the report were: a) content validation, b) criterion-related validation, and c) construct validation. The report, initially published in 1954, was significant because it formally recognized the existence of construct validity.

Content validity may be illustrated through the example of an examination given to students registered in a course. In order for the examination to be

considered valid, it must encompass the material covered throughout the course, in proportion to the weight given to the different units of study. Nunnally (1978) stated that “the test must stand by itself as an adequate measure of what it is supposed to measure. Validity cannot be determined by correlating the test with a criterion, because the test itself is the criterion of performance.” (p.91). Instead, the overlap between the objectives on the test blueprint and course objectives is determined, and if it reaches an a priori specified level of agreement, the test is pronounced to be content valid for the intended purpose.

Interestingly, test manufacturers concerned with content validity must attempt to build in validity prior to the construction of the measuring instrument, as opposed to criterion-related validity which seeks evidence of validity after the test has been administered. This is done primarily through the use of a plan for constructing the test, called a test blueprint, such that a procedure is outlined for obtaining a representative sample from the domain of content.

Other aspects of the plan may include the ordering of the test items and the instructions provided. Nunally (1978) elaborated further in claiming that a test blueprint and a table of specifications should be incorporated in constructing the test. The test blueprint is an outline of the material covered in the course. A table of specifications pertains to determining the cognitive level of the curriculum objectives that students are expected to achieve. According to Nunnally (1978),

> In spite of some efforts to settle every issue about psychological measurement by a flight into statistics, content validity is mainly settled in other ways. Although helpful hints are obtained from the analyses of statistical findings, content validity primarily rests upon an appeal to the propriety of content and the way that it is presented (p.94).

Test manufacturers thereby have the responsibility of ensuring that the sample of content is representative and the plan for constructing the test is sound.

The next type of validity outlined in the Technical Recommendations report was criterion-related validity. Criterion-related validity is actually comprised of two different types of validity that have been categorized under this heading: a) predictive validity and b) concurrent validity. According to Suen (1990), the only difference between predictive and concurrent validity is the exact time in which the measurements of the criterion variable occurs. As a result, with predictive validity the criterion measurement is taken at some point in the future, whereas with concurrent validity the criterion measurement is taken at the same time as the test is being administered. One benefit to the use of criterion-related validity is that it may be evaluated using a statistical approach, a Pearson Product-Moment Correlation Coefficient.

With regard to predictive validity, it is dependent upon the correlation between the scores obtained on the measuring instrument and those obtained from the criterion variable, which serves as an indicator of the desired characteristic or behavior that the test aims to predict. For example, a widely recognized example of evidence of predictive validity are the studies on the Scholastic Aptitude Test (SAT) and its correlation with final grade point average.

Unlike other forms of validity, interpretation of predictive validity is based solely on the correlation between the scores on the test and the criterion variable. However, common sense must apply and the shrewd researcher must recognize the danger that extraneous variables present to the interpretation of such

correlations. In addition, if the restriction of range assumption has been violated, then it could taint the results by either inflating or deflating the correlation. For this reason, it is important to draw a random sample of subjects when conducting the analysis.

Concurrent validity is closely related to predictive validity; differing in terms of (a) a simulated criterion and (b) shorter timeframe between the test and what constitutes the criterion. Whereas predictive validity involves administering the test prior to obtaining scores on the criterion variable, concurrent validity administers the test concurrent to a simulated criterion variable. A correlation of the scores obtained from the instrument and those taken from the criterion measure are used in evaluating this approach to validity. Concurrent validity is useful in predicting immediate success, such as employment readiness, or for the purpose of substituting one instrument for another (e.g., a less expensive instrument, an instrument more commonly available, an instrument that takes less time or expertise to administer).

The third type of validity as outlined in the Technical Recommendations report was labeled as construct validity. Interestingly, professional literature in the field of measurement dealt almost exclusively with content and predictive validity prior to the decade of the 1950s (Cronbach, 1971; Nunnally, 1978). This was the case, despite the fact that the concept of construct validation had traditionally been a part of behavioral research, without being formally recognized (Cronbach & Meehl, 1955).

According to Cronbach and Meehl (1955), construct validation is “involved whenever a test is to be interpreted as a measure of some attribute or quality which is not operationally defined” (p.282). They are referring to a construct, which is commonly understood to be a fiction that is used to explain reality (Sawilowsky, personal communications). It is a variable that is abstract by nature and cannot be directly measured or observed. The concept of research involving constructs, developed out of personality testing (Cronbach & Meehl, 1955) where it was found that for certain variables, such as self-esteem, there was no appropriate criterion to predict or suitable domain of content to sample (Cronbach, 1971). These variables required an alternative approach to validation.

In its simplest form, construct validity rests with the notion that a particular test actually measures the construct which it is intended to measure. The determination of the validity of a particular test becomes increasingly important as the variables being studied progress along the spectrum from concrete to abstract. Rarely however, are variables of interest in research studies simply defined. Theories espoused in behavioral and social science research are consumed predominantly with the use of construct variables as opposed to specific, observable variables (Nunnally, 1978).

Therefore, constructs are abstract variables that cannot be studied in isolation. A researcher interested in studying intelligence will be interested in employing an instrument that exhibits high construct validity. However, the construct is not measured directly through the use of a test. Instead, the construct is observed indirectly through a collection of indicators known as

observables (Suen, 1990). Therefore, a test aimed at measuring intelligence, must do so indirectly by measuring the observable variables that are manifested within this construct. In terms of intelligence, observables may include reading comprehension, problem-solving skills, and special reasoning. The extent to which the test embodies these manifestations of the construct, will determine whether construct validation has been determined (Suen, 1990).

The uncertainty regarding the domain of the observables is one of the difficulties in determining construct validity. The problem inherent in construct validation, is that constructs can become increasingly complicated based upon the size of the domain of the observables that define it. Nunnally (1978) argued that,

> the larger the domain of observables related to a construct, the more difficult it tends to be to define which variables do or do not belong in the domain…Typically, scientists hold a firm belief about some of the more prominent observables related to a construct, but beyond that they can only hypothesize how far the construct extends (p.97).

Unfortunately, it is impossible from the perspective of a researcher to know if all possible observables have been incorporated. Regardless, defining the domain of the observables is a necessary step in construct validation. A well-specified domain is comprised of a set of observables that are highly correlated with one another and are thereby shown to measure the same construct. Internal consistency in this regard, is an important and necessary step toward determining construct validation, however it is not sufficient.

The Multitrait-Multimethod Matrix

Campbell and Fiske (1959) created the Multitrait-Multimethod Matrix as a means evaluating convergent and discriminant validity, which in turn is viewed as evidence of construct validation. Convergent validity is achieved when there is a high correlation among the various methods specifically designed to measure a particular construct. Conversely, divergent validity is equally important in that a low correlation should be exhibited between the methods that are designed to measure different constructs. Otherwise, it would raise doubt as to whether the constructs are truly distinct.

The matrix has become the accepted approach for construct validation and has received considerable attention amongst researchers since it was first published. According to Sternberg (1992), it has received over 2,000 citations over the years, making it the most cited paper published by Psychological Bulletin. Despite the impact that the matrix has had on the study of construct validity, it is not without its problems. Campbell and Fiske (1959) recognized that further study is required and that “various statistical treatments for Multitrait-Multimethod matrices might be developed…However, the development of such statistical methods is beyond the scope of this paper” (p.103). The development of the Multitrait-Multimethod Matrix was viewed as a necessary first step in determining construct validity, from which it was believed that further research would resolve these issues over time.

The recognized limitations of their study, as presented in their original article, turned to exasperation as little progress had been made in evaluating the

matrix. Fiske and Campbell (1992) expressed their frustration by stating that scholarly journals and researchers alike continue to accept articles that provide no greater evidence of convergent and discriminant validity than from the time their original article was first published. Fiske and Campbell (1992) furthered their discussion by stating that there is still no general consensus of how to statistically evaluate convergent and discriminant validity and have resolved themselves to the fact that “eyeballing an MTMM (Multitrait-Multimethod) matrix in terms of the weak criteria proposed in the original article seems like a sound first step…” (p.394). As a result, the problems discussed by Campbell and Fiske (1959) remain to this day.

Regardless, the Multitrait-Multimethod Matrix developed by Campbell and Fiske (1959) remains the focus of any approach aimed at construct validation. The matrix highlighted the notion that it is not sufficient to demonstrate convergent validity alone and that discriminant validity must also be attained to ensure that indicators of one construct have low correlations with measures aimed at interpreting other constructs (Cronbach, 1971).

According to Campbell and Fiske (1959), “At the current stage of psychological progress, the crucial requirement is the demonstration of some convergence, not complete congruence, between two distinct sets of operations. With only one method, one has no way of distinguishing trait variance from unwanted method variance.” (p.102). A measuring instrument that displays high convergent validity between traits that are intended to converge and low discriminant validity between traits that are supposed to differ is viewed as

evidence of construct validity. The Multitrait-Multimethod Matrix is thereby necessary in evaluating discriminant validity as the trait and method variance is highlighted through the intercorrelations between several traits and several methods.

The matrix is subdivided into various components that contribute to the analysis which include the: a) reliability diagonal, b) validity diagonal, c) heterotrait-monomethod block, and d) heterotrait-heteromethod block. Campbell and Fiske (1992) provided a guideline for interpreting the matrix and determining the degree of convergent and discriminant validity. Figure 1 provides an illustration of the various components of the Multitrait-Multimethod Matrix.

	Method One			Method Two			Method Three		
	A_1	B_1	C_1	A_2	B_2	C_2	A_3	B_3	C_3
Method One									
A_1	(.89)								
B_1	**.51**	(.89)							
C_1	**.38**	**.37**	(.76)						
Method Two									
A_2	[.57]	.22	.09	(.93)					
B_2	.22	[.57]	.10	**.68**	(.91)				
C_2	.11	.11	[.46]	**.59**	**.58**	(.81)			
Method Three									
A_3	[.56]	.22	.11	[.67]	.42	.33	(.94)		
B_3	.23	[.58]	.12	.43	[.66]	.34	**.67**	(.92)	
C_3	.11	.11	[.45]	.34	.32	[.58]	**.58**	**.60**	(.85)

Figure 1. An Example of a Multitrait-Multimethod Matrix (Campbell & Fiske, 1959, p.82)

Note. A = assertive; B = cheerful; C = serious. Values in parentheses represent the reliability diagonal. Values in the squared brackets represent the validity diagonal. Boldface type represents the heterotrait-monomethod values and regular type represents the heterotrait-heteromethod values.

To evaluate convergent validity, the values found in the validity diagonal "should be significantly different from zero and sufficiently large to encourage further examination of validity" (Campbell & Fiske, 1959, p.82). Conversely, the process in determining discriminant validity is more involved. To begin, the values in the validity diagonal should be higher than the values found in the corresponding heterotrait-monomethod block. Second, the values in the heterotrait-monomethod block should be higher than the values found in the heterotrait-heteromethod block. In applying the rationale outlined by Campbell and Fiske (1959), there should be an ascending trend from the heterotrait-heteromethod values to the reliability diagonal.

A Critique of the Multitrait-Multimethod Matrix

The guidelines provided by Campbell and Fiske (1959) in evaluating convergent and discriminant validity have been debated. According to Jackson (1969), "it is important to recognize these (guidelines) as preliminary recommendations rather than rigorous, objective criteria which take into account the structure implicit in the matrix." (p.31). Specifically, Jackson (1969) argued the following points:

1. The practice of examining the validity correlations to ensure that these values are higher than those found in the appropriate row and column may be laborious if the matrix is large and complex. Further, it is subjective in that there is no stipulated baseline value for a decision on whether convergent or discriminant validity has been achieved.

2. Correlations within the matrix may fluctuate due to sampling error, measurement error, or the respective reliabilities of the variables. This could lead to an incorrect conclusion of either convergent or discriminant validity. According to Jackson (1969), one should consider standardizing the reliabilities, regardless of the fact that it would involve tampering with the data.
3. Emphasis should be placed on evaluating the entire set of relationships between traits found within the matrix, instead of examining single correlation coefficients. A large matrix increases the probability of inconsistent evidence for convergent and discriminant validity. As well, there may be evidence of convergent and discriminant validity apparent through an examination of the whole matrix, that is not revealed through an investigation of individual traits.
4. Method variance within a monomethod unit will distort the correlations between traits. Method variance will challenge the notion that correlations should be higher for heteromethod validities, than correlations with irrelevant traits in the monomethod matrix. Although it may be argued that methods for reducing method variance may be employed in order to satisfy the criteria outlined by Campbell and Fiske (1959), it is not always possible.
5. The larger the Multitrait-Multimethod Matrix, the more difficult it is to reach the conclusion of convergent or discriminant validation. As the

number of traits and methods increase, so do the number of correlations to be evaluated.

As a result of the deficiencies outlined above, Jackson (1969) stated the need for a quantitative approach that will address the problems inherent in the guidelines established by Campbell and Fiske (1959) to evaluate the matrix.

Additionally, Widaman (1985) provided a critique of these guidelines by stating that “the comparison procedures proposed by Campbell and Fiske (1959), though rather straightforward to follow, do have a number of shortcomings…” (p.1). According to Widaman (1985), the following issues need to be addressed:

1. The underlying approach to evaluating the matrix is dependent upon a comparison of the different correlations. The lack of independence of the correlations prevents a statistical analysis of the overall pattern of correlations found within the matrix.
2. It is not possible to acquire precise measures of trait or method variance. This information would be of assistance in determining which measures require modification.
3. Differences in the level of reliability among variables will distort the correlations between measures which serve as the basis for evaluating the matrix.

Although a necessary a first step in determining construct validation, the work of Campbell and Fiske (1959) requires clarification and an extension of these basic underlying principles for evaluation through an appeal to statistical methods.

According to Boruch, Larkin, Wollins, & MacKinney (1970), “Such an examination is rather complicated, and the results are often ambiguous.” (p.835).

Quantitative Approaches to Evaluating the Multitrait-Multimethod Matrix

The Multitrait-Multimethod Matrix devised by Campbell and Fiske (1959) provided a baseline approach for determining construct validity. Due to the importance of constructs in behavioral research, the matrix has received a great deal of attention over the years. As a result, various statistics have been employed as a means of analyzing the matrix.

One approach that has received attention as a means of evaluating the matrix has been the nonparametric procedure that was developed by Hubert and Baker (1978). An advantage of this procedure would be that it does not have the assumptions that hamper parametric statistics, namely normality and homogeneity of variance. The procedure employs three indices that reflect the criteria for evaluating the matrix as outlined in Campbell and Fiske (1959). The indices are as follows:

1. the average of the same-trait correlations (as a reflection of the condition that the correlations within subsets should be large);
2. the difference between the average same-trait correlation and the average for different traits measured under different methods (as a reflection of the condition that the within subset correlations should be larger than the between subset correlations based on different methods);

3. the difference between the average same-trait correlation and the average same-method correlation (as a reflection of the condition that the correlations within subsets should be large)

(Hubert & Baker, 1978)

The first indicant is an approach for determining convergent validity, whereas the second and third indices are to be used as evidence of discriminant validity. The problem with this approach is that a method for joining the three indices has yet to be developed (Sawilowsky, 2002). As a result, one must examine whether each of the indices is significant before a determination of construct validity can be made. In the event that all three indices are significant, then the decision is quite simple; however, if significance is not unanimous among the three, then the procedure becomes subjective.

However, the nonparametric approach is just one of a collection of statistical procedures that have surfaced over the years. Another approach that has received much attention as a means of evaluating the matrix has been the analysis of variance procedure (Stanley, 1961). According to Schmitt and Stults (1986), the most common approach employed in the evaluation of the Multitrait-Multimethod Matrix has been analysis of variance. Stanley (1961) was the first to propose a three way factorial analysis of variance approach for examining the Multitrait-Multimethod Matrix.

However, despite its popularity, the analysis of variance approach is not without its problems. Boruch et al. (1970) argued that this may often be inappropriate due to the following reasons: a) it does not take into account

measurement error, b) there are strict assumptions pertaining to additivity of effects and homogeneity of variance, and c) observations may be a function of the scale used. Schmitt and Stults (1986) argued that the restrictive assumptions of the analysis of variance procedure may limit its practical use in evaluating the matrix: a) independence may be jeopardized by the intercorrelations inherent in the matrix and b) homoscedasticity is jeopardized if the variances of different trait and method factors are equal. Yet another problem with this approach is that an examination of trait-method interaction requires data collection to be repeated (Stanley, 1961).

Finally, confirmatory factor analysis has become the current en vogue approach to analyzing the Multitrait-Multimethod Matrix. In examining congeneric tests, Jöreskog (1971) discussed the use of confirmatory factor analysis as a means of analyzing the Multitrait-Multimethod Matrix. The procedure involves the analysis of a number of different models; each evaluated using the Chi Square Goodness of Fit test. The first step in the analysis is to test the model that all methods are equivalent in measuring each trait. If the model is found to fit the data, then the interrelationships between the trait factors may be further analyzed by using a factor analysis of the correlation matrix. However, if this is not the case, then one must test additional models that take into account trait and method factors. In addition, if any of the methods are found to be exhibit high collinearity then they are to be combined into one. One problem cited by Jöreskog (1971) in using this technique, is that one does not actually test a given hypothesis, rather it is a process of continually shaping the model such that the

differences between subsequent chi square values are of greater concern than the values themselves.

Kalleberg and Kluegel (1975) argued in favor of confirmatory factor analysis; citing the real value of this approach is that it allows researchers to decompose and examine the different components of the matrix correlations. According to Kalleburg and Kluegel (1975), a severe limitation of the Campbell and Fiske (1959) approach was the assumption that traits and methods are uncorrelated and that methods are minimally correlated with each other; an assumption that was deemed questionable and one that may lead to an incorrect assessment of validity if violated. In contrast, confirmatory factory analysis has been found to have the following advantages: a) the researcher may estimate values of the correlations for methods, traits, or among methods and traits and b) the researcher may estimate values of the effects of each method and trait factor on the given measure (Kalleburg & Kluegel, 1975). Moreover, Bagozzi (1978) argued that confirmatory factor analysis “allows the researcher to explicitly partition the variance due to construct, method, and error; it provides insights as to why convergent and discriminant validity has or has not been attained…” (p.12).

Pedhazur & Schmelkin (1991) discussed the use of confirmatory factor analysis; they argued that the procedure was very complicated in terms of an overwhelming number of models involved in the analysis and they believed that the complexity of the procedure may lead different researchers, analyzing the same data, to arrive at different conclusions. Furthermore, confirmatory factor

analysis is based on the assumption that the relationship between the variables is linear and additive (Kalleburg & Kluegel, 1975). Schmitt and Stults (1986) recommended this approach despite the deficiencies of the procedure that are outlined in their article. Perhaps this may be construed as evidence of a sense of complacency that has begun to take root, as researchers become willing to accept flawed approaches due to a lack of suitable alternatives.

Despite the attention that the matrix has received over the years, an appropriate approach to evaluating the matrix has yet to be found. Each of the attempts outlined above have deficiencies that discredit their use in applied settings. Many of the previous attempts at developing a statistical approach to analyzing the Multitrait-Multimethod Matrix have been plagued in part by assumptions that can only be categorized as restrictive and difficult to satisfy.

As a result, Sawilowsky (2002) created a quick, distribution-free test that does not suffer the same pitfalls of its predecessors. It was called the I statistic because it focuses on the number of inversions found within the matrix. The I statistic is relatively simple to compute, it incorporates the entire matrix, and it does not have the restrictive assumptions that have hampered previous efforts.

The I statistic is a combination of the Jonckheere's distribution-free k-sample test against ordered alternatives (Jonckheere, 1954) and Mann's test for randomness in a single sample (Neave & Worthington, 1988). According to Sawilowsky (2002), "The I statistic combines the counting function of the Mann's test with the logic of Jonckheere's statistic." (p.85). Whereas Jonckheere's test uses all of the values within the matrix, which increases the power of the test, but

also increases the probability of violating the independence assumption; the I statistic is limited to three values at each level of the matrix: a) minimum coefficient, b) median coefficient, and c) maximum coefficient. As a result, a minimum, median, and maximum value is derived from the each of the following components of the Multitrait-Multimethod Matrix: a) reliability diagonals, b) validity diagonals, c) heterotrait-monomethod block, and d) heterotrait-heteromethod block.

The basic premise of the I statistic is an examination of the upward trend of values, from the heterotrait-heteromethod values to the reliability diagonal, as evidence of construct validity. This approach incorporates the criteria outlined by Campbell and Fiske (1959), in that the values in the heterotrait-heteromethod block should be lower than the values found in the heterotrait-monomethod block, which in turn should be lower than those found in the validity diagonals, and so forth. Therefore, construct validity is supported through fewer inversions. A nominal number of inversions are easily regarded as evidence of construct validity; however, the decision becomes more difficult and subjective as the number of inversions increase.

As previously discussed, the I statistic requires that each of the four levels of the matrix is organized into three values. As a result, the total number of combinations of the twelve values is 369,600. In addition, due to the fact that the order of the values within each level of the matrix is fixed, there are a maximum number of 54 inversions as opposed to 66. A permutation of the data provided the number of times that each inversion (1 to 54) would occur; these values were

in turn divided by 369,600 and the resulting probabilities were summed, thereby creating the cumulative distribution function. In examining the table of critical values, ten inversions would be significant at the 0.01 alpha level, 14 inversions would be significant at the 0.05 alpha level, and 17 inversions would be significant at the 0.10 alpha level. This in turn would eliminate the subjectivity that results from “eyeballing” the matrix as suggested by Campbell and Fiske (1959).

In comparing the I statistic to the guidelines provided by Campbell and Fiske (1959), it was found that the I statistic provided comparable results. Two examples of real data sets from the Campbell and Fiske (1959) article were used as a means for comparison. In their analysis of the first data set presented in Figure 2, Campbell and Fiske (1959) determined that “the evidence of test validity to be presented here is probably poorer than most psychologists would have expected” (p.85). In their assessment of the data, only one of the traits had a value in the validity diagonal that was higher than the all of the values in the heterotrait-heteromethod block. Sawilowsky (2002) analyzed the same data set using the I statistic as a means of analysis and found that there were a total of 15 inversions. Using the table of critical values discussed earlier, the matrix was not significant at the 0.05 alpha level.

	Peer Ratings				Association Test			
	A_1	B_1	C_1	D_1	A_2	B_2	C_2	D_2
Peer Ratings								
A_1	(.82)							
B_1	**.74**	(.80)						
C_1	**.63**	**.65**	(.74)					
D_1	**.76**	**.78**	**.65**	(.89)				
Association Test								
A_2	[.13]	.14	.10	.14	(.28)			
B_2	.06	[.12]	.16	.08	**.27**	(.38)		
C_2	.01	.08	[.10]	.02	**.19**	**.37**	(.42)	
D_2	.12	.15	.14	[.16]	**.27**	**.32**	**.18**	(.36)

Figure 2. An Example of a Multitrait-Multimethod Matrix Exhibiting Poor Construct Validity (Campbell & Fiske, 1959, p.86)

Note. A = courtesy; B = honesty; C = poise; D = school drive. Values in parentheses represent the reliability diagonal. Values in the squared brackets represent the validity diagonal. Boldface type represents the heterotrait-monomethod values and regular type represents the heterotrait-heteromethod values.

In contrast, Figure 3 provides an example of a real data set that Campbell and Fiske (1959) have determined is "…typical of the best validity…" (p.97). Their evaluation is based on the fact that all but one of the traits have validities that exceed the values found in both the heterotrait-heteromethod and heterotrait-monomethod blocks. In comparison, the I statistic also provided strong evidence of construct validity. It should be noted that the data set was modified from the one used by Campbell and Fiske (1959) in that the self ratings method was eliminated from the analysis. Nonetheless, an examination of the data uncovered a total of 7 inversions which is significant at the 0.01 alpha level.

	Staff Ratings					Teammate Ratings				
	A_1	B_1	C_1	D_1	E_1	A_2	B_2	C_2	D_2	E_2
Staff Ratings										
A_1	(.89)									
B_1	**.37**	(.85)								
C_1	**-.24**	**-.14**	(.81)							
D_1	**.25**	**.46**	**.08**	(.84)						
A_1	**.35**	**.19**	**.09**	**.31**	(.92)					
Teammate Ratings										
A_2	[.71]	.35	-.18	.26	.41	(.82)				
B_2	.39	[.53]	-.15	.38	.29	**.37**	(.76)			
C_2	-.27	-.31	[.43]	-.06	-.03	**-.15**	**-.19**	(.70)		
D_2	.03	-.05	.03	[.20]	.07	**.11**	**.23**	**.19**	(.74)	
E_2	.19	.05	.04	.29	[.47]	**.33**	**.22**	**.19**	**.29**	(.76)

Figure 3. An Example of a Multitrait-Multimethod Matrix Exhibiting High Construct Validity (Campbell & Fiske, 1959, p.96)

Note. A = assertive; B = cheerful; C = serious; D = unshakeable poise; E = broad interests. The Self Rating Method was not included in the analysis, but does appear in the article. Values in parentheses represent the reliability diagonal. Values in the squared brackets represent the validity diagonal. Boldface type represents the heterotrait-monomethod values and regular type represents the heterotrait-heteromethod values.

Although Campbell and Fiske (1959) have provided a heuristic approach for evaluating construct validity, a statistical approach that incorporates these guidelines is necessary in order to eliminate the subjectivity involved in this process. Fiske and Campbell (1992) argued that “editors and readers are accepting matrices showing limited convergence or discrimination, or both, perhaps because these are so typical, so common in the published literature.” (p.393). Sawilowsky (2002) has shown that the I statistic provided comparable results to those achieved by Campbell and Fiske (1959) using a quick test that eliminates the subjectivity that has plagued this process in the past.

CHAPTER 3

METHODOLOGY

This research study will not involve the use of human or animal test subjects; nevertheless, permission was sought from the Human Investigation Committee at Wayne State University. Instead, the study involves a Monte Carlo simulation whereby data will be obtained through repeated sampling from the uniform distribution, as opposed to collecting data from a group of test subjects.

A program will be written in Compaq Visual Fortran (Version 6), to compute the three-point and four-point versions of the I test. Specifically, the programs will be written with the intent of examining the robustness of each test with regard to the internal correlation structure and to examine the power properties of each version of the test. The design layouts used in the analysis will be modeled on the matrices provided in Campbell and Fiske (1959). As a result, both the three-point and four-point versions of the I test will be computed using a 2x3, 3x2, 3x3, 2x4, and 3x5 matrix.

The data used in the analysis will be obtained using a random number generator. The number of values obtained will be dependent upon the design layout being modeled. As an example, with a 2x3 matrix, the total number of values obtained from the random number generator will be 21. These values will then be placed into one of four groups corresponding to the different levels of the Multitrait-Multimethod matrix. Therefore, in a 2x3 matrix, there are 6 heterotrait-heteromethod values, 6 heterotrait-monomethod values, 3 validity diagonal

values, and 6 reliability diagonal values. The three-point version of the I test requires three data points at each level: a) minimum, b) median, and c) maximum values. The four-point version of the I test requires four data points at each level: a) minimum, b) lower quartile, c) upper quartile, and d) maximum values. These data points will be obtained by sorting the data placed within each level to determine the minimum and maximum values and then computing the median for the three-point I test and the lower and upper quartiles for the four-point version of the I test.

In analyzing the robustness of I test, separate subroutines will be programmed to calculate both the three-point and the four-point versions of the test. A counter will be written into the program to check for the number of significant results at the 0.05 alpha level. This process will be repeated for 1,000,000 repetitions and the number of times that the null hypothesis is rejected will then be divided by 1,000,000; thereby providing the Type I error rate. This process will in turn be repeated for the 0.01 alpha level.

These results will be compared to those obtained by computing the I test using random, as opposed to sorted values. Specifically, a program will be written to compute both the three-point and four-point versions of the I test, whereby values will be placed within each level at random. Therefore, there is no internal correlation structure within each level. As a result, the program to be used to calculate the three-point I test using random data, will only obtain 12 random values from the uniform distribution, as opposed to 21 (assuming a 2x3 matrix). The first three values will be placed in the heterotrait-heteromethod level;

the next three values will be placed in the heterotrait-monomethod level, and so forth. The four-point I test program using random data will obtain 16 random values from the uniform distribution, as opposed to 21. The first four values will be placed in the heterotrait-heteromethod level; the next four values will be placed in the heterotrait-monomethod level, and so forth. As a result, the values will not be sorted and the minimum, median, and maximum values will not be calculated for the three-point I test, nor the minimum, lower quartile, upper quartile, and maximum values for the four-point I test. This process will in turn be repeated for the 0.01 alpha level.

Despite the fact that the values are not ascending within each level of the randomized version of the I test, the number of comparisons will remain constant for both the randomized and sorted versions of the I test. As a result, there will still be 54 comparisons made for the three-point version and 96 comparisons made for the four-point version. There will be no comparisons made within each level in determining the number of inversions. By maintaining the same number of comparisons, the critical values will remain the same and thus a comparison may be made for the random and sorted versions of both the three-point and four-point I tests regarding the Type I error rate.

The next phase of the study will be an examination of the power properties of both the three-point and four-point versions of the I test. The first phase of the study will focus on Type I error rate, whereby significance is based solely on the number of inversions, without regard for the types of values comprised within each of the levels. In an applied setting, an analysis of the Multitrait-Multimethod

matrix may be found to be significant; however, the results would be valid only if the reliability diagonal values were greater than or equal to 0.8. As a result, in determining the power properties of the I test, the reliability diagonal values will be kept above a predetermined standard. Specifically, a series of programs will be written for both the three-point and four-point versions of the I test that will ensure that the reliability diagonal values used in the analysis are greater than or equal to 0.7, 0.8, and 0.9 respectively. For each program, the number of significant results will be divided by the total number of repetitions to determine the power of the test. This process will be completed for both the 0.05 and 0.01 alpha levels.

Unlike the first phase of the study which will have 1,000,000 repetitions, this part of the research will have 2000 repetitions. The reason for the dramatic decrease in the number of repetitions is related to the time involved in processing 1,000,000 repetitions when the values are required to be above a predetermined standard. As a result, if the random number generator returns values that are below this predetermined standard, then the program will be prompted to loop back to the beginning to find a new random set of values from the distribution. As an example, if the reliability diagonal values are required to be greater than or equal to 0.9, then the program will be required to cycle through numerous times before it will return values that conform to this requirement.

As with the first phase of the research which focused on the robustness of the I test, the results obtained in the power analysis were compared to those obtained by computing the I test using random, as opposed to sorted values.

Once again, a program will be written to compute both the three-point and four-point versions of the I test, whereby values will be placed within each level at random. As a result, there will be no internal correlation structure within each level. The program will be set to 2000 repetitions and the number of significant results will be divided the number of repetitions to determine the power of the test. This process will be completed for both the 0.05 and 0.01 alpha levels.

In order to establish a baseline for comparison, the relative efficiency will be calculated to quantify and thereby allow for a comparison between the power of the four-point I test and the three-point I test. The relative efficiency will be calculated by dividing the three-point randomized values by the three-point sorted values. As well, the four-point randomized values will be divided by the four-point sorted values. The next step will be to divide the quotient from the four-point calculation by the quotient from the three-point calculation. This will provide the relative efficiency of the four-point I test versus the three-point I test and this calculation will be repeated for the 0.7, 0.8, and 0.9 thresholds for each of the experimental design layouts at both the 0.05 and 0.01 alpha levels.

The critical values to be used for the analysis of the three point I statistic were obtained from Sawilowsky (2002). It was found that the critical values for the three-point I statistic at the 0.05 and 0.01 alpha levels were 14 and 10, respectively. In contrast, the critical values for the four-point I statistic were obtained from Jonckheere (1954). Critical values for the 0.05 and 0.01 alpha levels were obtained by counting the number of inversions starting from the bottom of the table (refer to his Table 3, p.145). This is due to the fact that the

Jonckheere test works in reverse order to the Sawilowsky I statistic. It was found that the critical values for the four-point I statistic at the 0.05 and 0.01 alpha levels were 29 and 23 respectively.

CHAPTER 4

RESULTS

Type I Error Results

The first phase of the study involved an analysis of the robustness of the I test with regard to violations of independence. Specifically, the I test is in violation of the independence assumption because of the internal correlation structure inherent within each of the levels (i.e. minimum, median, and maximum values for each level for the three-point version). It was predicted by Sawilowsky (2002), that the Type I error rate would increase with an increasing number of data points (i.e. the three-point versus the four-point versions of the test). Although it was predicted that the Type I error rate would be adversely affected, the severity in violating this assumption remained unknown.

As a result, the Type I error rate for both the three-point and four-point versions of the I test were examined at both the 0.05 and 0.01 alpha levels. The three-point and four-point sorted versions of the I test were compared to the three-point and four-point randomized versions of the I test for various experimental design layouts (i.e. 2x3, 2x4, 3x2, 3x3, and 3x5 matrices).

In Table 1, it is shown that the randomized versions of both the three-point and four-point versions of the test performed as expected, with a Type I error rate that was close to 0.05; specifically, 0.042514 for the three-point randomized version and 0.042045 for the four-point randomized version. In examining the three-point and four-point sorted versions of the I test, it was found that the Type I error rate did increase with an increasing number of data points. Using the 2x3

matrix as an example, the Type I error rate for the three-point sorted version of the I test was 0.002193 and the Type I error rate for the four-point sorted version of the I test was 0.007527. This result was consistent across each of the experimental design layouts tested.

Table 2 examined the robustness of both the three-point and four-point versions of the I test at the 0.01 alpha level. Once again, it was found that the randomized versions of the test performed as expected, with a Type I error rate that was close to 0.01 (i.e. 0.009254 for the three-point randomized version and 0.009789 for the four-point randomized version). As well, it was found that

Table 1. Type I Error Rate for both the Three-Point and Four-Point I Test at the 0.05 Alpha Level

Matrix	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-point Sorted Values
2x3	0.042514	0.002193	0.042045	0.007527
3x2	0.042514	0.001807	0.042045	0.006161
2x4	0.042514	0.000039	0.042045	0.000285
3x3	0.042514	0.000001	0.042045	0.000036
3x5	0.042514	0.000000	0.042045	0.000000

Note: Values obtained using 1,000,000 repetitions

Table 2. Type I Error Rate for both the Three-Point and Four-Point I Test at the 0.01 Alpha Level

Matrix	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-point Sorted Values
2x3	0.009254	0.000106	0.009789	0.000842
3x2	0.009254	0.000081	0.009789	0.000585
2x4	0.009254	0.000001	0.009789	0.000006
3x3	0.009254	0.000000	0.009789	0.000000
3x5	0.009254	0.000000	0.009789	0.000000

Note: Values obtained using 1,000,000 repetitions

the Type I error rate increased with an increasing number of data points. Using the 2x3 experimental design layout, it was found that the Type I error rate for the three-point sorted version of the I test was 0.000106 and the Type I error rate for the four-point sorted version of the I test was 0.000842. Once again, the result was consistent across each of the experimental design layouts tested.

Power results

The second phase of the research examined the power of the I test by maintaining a predetermined threshold for the reliability diagonal values used in the analysis. The I test was computed with minimum reliability diagonal values set at 0.7, 0.8, and 0.9. It was expected that the power of both the three-point

and four-point versions of the test would increase as the predetermined threshold for the reliability diagonal values increased, because it was logical to assume that there would be fewer inversions. As a result, focus was instead placed upon the examination of the three-point versus the four-point I test in terms of power.

The relative efficiency was calculated as a means of quantifying the performance of the four-point I test in relation to the three-point I test at each of the thresholds set for the reliability diagonal values. The relative efficiency was calculated by dividing the three-point randomized values by the three-point sorted values. As well, the four-point randomized values were divided by the four-point sorted values. The next step was to divide the quotient from the three-point calculation by the quotient from the four-point calculation. This provided the relative efficiency of the four-point I test versus the three-point I test and this calculation was repeated for the 0.7, 0.8, and 0.9 thresholds for each of the experimental design layouts at both the 0.05 and 0.01 alpha levels.

Tables 3, 4, and 5 illustrate the comparative power of both the three-point and four-point versions of the I test at the 0.05 alpha level, using various experimental design layouts (i.e. 2x3, 2x4, and 3x2 matrices respectively). Programs were written to compute the three-point and four-point versions of the I test using a 3x3 and 3x5 matrix; however, due to limitations in the processing speed of the computer used, the programs did not resolve values for these design layouts. However, it must be noted that these power equations are in closed form; therefore, a lack of resolution only indicates a limitation of

resources. These values would compute given the proper time and resources to complete the analysis.

Tables 3, 4, and 5 each display an increased efficiency of the four-point over the three-point versions of the I test. In Table 3, the relative efficiency of the four-point test is nearly double (1.88) in comparison to the three-point test with a minimum reliability diagonal value of 0.7. In Table 4, the relative efficiency is more than four times greater (4.16) with a minimum reliability diagonal value of 0.7. A higher relative efficiency was displayed in Table 5 as well with a value that is double that of the three-point version with a minimum reliability diagonal value of 0.7. The gains in relative efficiency do tend to decrease as the minimum reliability diagonal values increase. Despite this fact, the four-point I test was proven to be a more powerful test because it draws on a greater number of data points.

Table 3. Comparative Power Between the Three-point and Four-point Versions of the I Test Using a 2x3 Matrix Design Layout at the 0.05 Alpha Level

Reliability Diagonal Values	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-Point Sorted Values	Relative Efficiency
≥ 0.7	0.3040	0.1430	0.3920	0.3460	1.88
≥ 0.8	0.3980	0.2305	0.5305	0.5020	1.63
≥ 0.9	0.5405	0.3600	0.6510	0.6830	1.57

Note: Values obtained using 2000 repetitions

Table 4. Comparative Power Between the Three-point and Four-point Versions of the I Test Using a 2x4 Matrix Design Layout at the 0.05 Alpha Level

Reliability Diagonal Values	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-Point Sorted Values	Relative Efficiency
≥ 0.7	0.3040	0.0390	0.3920	0.2090	4.16
≥ 0.8	0.3980	0.0625	0.5305	0.3365	4.03
≥ 0.9	0.5405	Did not resolve	0.6510	Did not resolve	n/a

Note: Values obtained using 2000 repetitions. N/A = not applicable

Table 5. Comparative Power Between the Three-point and Four-point Versions of the I Test Using a 3x2 Matrix Design Layout at the 0.05 Alpha Level

Reliability Diagonal Values	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-Point Sorted Values	Relative Efficiency
≥ 0.7	0.3040	0.1315	0.3920	0.3490	2.06
≥ 0.8	0.3980	0.2165	0.5305	0.5120	1.77
≥ 0.9	0.5405	Did not resolve	0.6510	Did not resolve	n/a

Note: Values obtained using 2000 repetitions. N/A = not applicable

Tables 6, 7, and 8 illustrate the comparative power of both the three-point and four-point versions of the I test at the 0.01 alpha level, using various experimental design layouts (i.e. 2x3, 2x4, and 3x2 matrices respectively). Once again, programs were written to compute the three-point and four-point versions of the I test using a 3x3 and 3x5 matrix; however, due to limitations in the processing speed of the computer used, the programs did not resolve values for these design layouts.

Table 6. Comparative Power Between the Three-point and Four-point Versions of the I Test Using a 2x3 Matrix Design Layout at the 0.01 Alpha Level

Reliability Diagonal Values	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-Point Sorted Values	Relative Efficiency
≥ 0.7	0.0995	0.0205	0.1525	0.0949	3.02
≥ 0.8	0.1410	0.0435	0.2435	0.1755	2.34
≥ 0.9	0.2280	0.0839	0.3395	Did not resolve	n/a

Note: Values obtained using 2000 repetitions. N/A = not applicable

Table 7. Comparative Power Between the Three-point and Four-point Versions of the I Test Using a 2x4 Matrix Design Layout at the 0.01 Alpha Level

Reliability Diagonal Values	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-Point Sorted Values	Relative Efficiency
≥ 0.7	0.0995	0.0005	0.1525	0.0130	16.96
≥ 0.8	0.1410	0.0025	0.2435	Did not resolve	n/a
≥ 0.9	0.2280	Did not resolve	0.3395	Did not resolve	n/a

Note: Values obtained using 2000 repetitions. N/A = not applicable

Table 8. Comparative Power Between the Three-point and Four-point Versions of the I Test Using a 3x2 Matrix Design Layout at the 0.01 Alpha Level

Reliability Diagonal Values	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-Point Sorted Values	Relative Efficiency
≥ 0.7	0.0995	0.0160	0.1525	0.0845	3.45
≥ 0.8	0.1410	0.0299	0.2435	0.1535	2.97
≥ 0.9	0.2280	Did not resolve	0.3395	Did not resolve	n/a

Note: Values obtained using 2000 repetitions. N/A = not applicable

The trend regarding the increased efficiency of the four-point I test versus the three-point I test is again displayed in Tables 6, 7, and 8. In Table 6, the relative efficiency of the four-point test is more than three times greater (3.02) in comparison to the three-point test with a minimum reliability diagonal value of 0.7. In Table 7, the relative efficiency is nearly seventeen times greater (16.96) with a minimum reliability diagonal value of 0.7. A higher relative efficiency was displayed in Table 8 as well with a relative efficiency nearly three and half times greater with a minimum reliability diagonal value. Once again, the difference in relative efficiency did decrease as the minimum reliability diagonal values increased; however, the fact remained that the four-point I test is more powerful than its three-point counterpart.

One of the problems uncovered in the research was the fact that the I test proved to be extremely conservative. As a result, although the critical values used in the analysis were mathematically correct based on elementary combinatorial analysis (i.e. 14 and 10 for the three-point I test at the 0.05 and 0.01 alpha levels respectively and 29 and 23 for the four-point I test at the 0.05 and 0.01 alpha levels respectively), it was found that as a hypothesis test, the lack of independence within each level of the I test resulted in depressed Type I errors for these critical values.

As a result, ad hoc critical values were tested to determine the critical values that should be used in an applied setting to optimize the power of the test. They were obtained for both the three-point and four-point versions of the I test at

both the 0.05 and 0.01 alpha for the following experimental design layouts: a) 2x3 matrix, b) 2x4 matrix, c) 3x2 matrix, d) 3x3 matrix, and e) 3x5 matrix.

The ad hoc critical values for both the three-point and four-point versions of the I at the 0.05 alpha level are presented in Table 9. It was found that the ad hoc values were quite different from those taken from the cumulative distribution function. As an example, the optimal critical value for a 2x3 matrix at the 0.05 alpha level is 19 for the three-point I test and 35 for the four-point I test. These values are different from those taken from the suggested values of 14 and 29 respectively. The difference is greater as the matrix becomes larger. In analyzing a 3x5 matrix, it was found that the optimal critical values were 22 for the three-point I test and 41 for the four-point I test.

These findings were consistent with ad hoc critical values tested at the 0.01 alpha level. The ad hoc critical values for both the three-point and four-point versions of the I at the 0.01 alpha level are presented in Table 10. Once again, these values were quite different from those taken from the suggested values of 10 for the three-point I test and 23 for the four-point I test. Using a 2x3 matrix as an example, the optimal critical value at the 0.01 alpha level is 16 for the three-point I test and 29 for the four-point I test. Once again, these differences grew larger as the matrix grew more complex.

Table 9. Ad Hoc Critical Values for both the Three-Point and Four-Point I Test at the 0.05 Alpha Level

Matrix	Ad Hoc Critical Values	Three-Point Sorted Values	Ad Hoc Critical Values	Four-point Sorted Values
2x3	19	0.0418	35	0.0491
3x2	21	0.0426	38	0.0389
2x4	19	0.0389	35	0.0445
3x3	21	0.0285	39	0.0387
3x5	22	0.0343	41	0.0405

Note: Values obtained using 1,000,000 repetitions

Table 10. Ad Hoc Critical Values for both the Three-Point and Four-Point I Test at the 0.01 Alpha Level

Matrix	Ad Hoc Critical Values	Three-Point Sorted Values	Ad Hoc Critical Values	Four-point Sorted Values
2x3	16	0.0080	29	0.0075
3x2	19	0.0083	35	0.0094
2x4	16	0.0069	30	0.0088
3x3	19	0.0037	36	0.0069
3x5	20	0.0033	39	0.0089

Note: Values obtained using 1,000,000 repetitions

CHAPTER 5

CONCLUSION

According to Sawilowsky (2002), the problem with the Jonckheere test was that it used all of the values in the matrix and thereby greatly increased the risk of violating the assumption of independence and would thereby lead to an inflation in the Type I error rate. By using only three data points within each level, the three-point I test was conceived as an alternative test of trend that would limit the severity of violating this assumption.

However, the question of Type I errors due to lack of independence remained a source of concern and warranted further examination (Sawilowsky, 2002). Upon completing the analysis, it was found that the Type I errors did indeed increase. The four-point I test was found to have a higher Type I error rate than the three-point I test across all experimental design layouts tested. This result was not surprising and confirmed the prediction made by Sawilowsky (2002), who also noted that “The violation of independence is a recipe for disaster in terms of Type I errors. There is no statistic that can overcome a true lack of independence, either within or between scores” (Sawilowsky, 2007, p. 208).

Although the Type I error rate did increase with an increasing number of data points, it was found that the test was extremely conservative to begin with. As a result, there was no harm with regard to an increasing Type I error rate because of the conservative nature of the test. Unfortunately, the test proved to be so conservative that its power would be low in an applied setting. In analyzing

the power of the both the three-point I test and the four-point I test, there was a gain in relative efficiency in using the four-point I test; however, the power still proved to be weak as a practical alternative.

Although the research found that the power of both the three-point and four-point versions of the I test were weak; it must be noted that the I test is still a better alternative to evaluating the Multitrait-Multimethod Matrix than an eyeballing of the matrix, using the guidelines established by Campbell and Fiske (1959). The I test was found to be very conservative in terms of its Type I error rate; however, it does not suffer from the pitfalls that have hampered other approaches that have been developed over the years. Specifically, it does not have the restrictive assumptions of the ANOVA approach or the complexity of the Confirmatory Factor Analysis approach. As a result, the I test should still be considered as an acceptable approach to evaluating the matrix until a superior alternative is developed.

In recognition of the conservative nature of the I test, a set of ad hoc critical values were obtained to increase the power of both the three-point and four-point versions of the I test. The hope was that a trend would emerge from the set of ad hoc critical values, whereby a general rule could be developed to determine the appropriate critical values for any experimental design layout. Unfortunately, a trend could not be noticed that would provide a reliable and accurate rule that researchers could employ in an applied setting. By implementing the ad hoc critical values, there is a noticeable gain in power; however, values were only provided for a limited array of matrices. Without a

general rule in determining the appropriate critical values, a researcher would have to independently test the optimal critical values for matrices that extend beyond the context of this research.

REFERENCES

Bagozzi, R. P. (1978). The construct validity of the affective, behavioral, and cognitive components of attitude by analysis of covariance structures. *Multivariate Behavioral Research, 13*, 9-31.

Boruch, R. F., Larkin, J. D., Wolins, L., & MacKinney, A. C. (1970). Alternative methods of analysis: Multitrait-multimethod data. *Educational and Psychological Measurement, 30*, 833-853.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp.443-507). Washington, D.C.: American Council on Education.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist, 16*, 640-647.

Fiske, D. W. & Campbell, D. T. (1992). Citations do not solve problems. *Psychological Bulletin, 112*(3), 393-395.

Hubert, L. J. & Baker, F. B. (1978). Analyzing the multitrait-multimethod matrix. *Multivariate Behavioral Research, 13*, 163-179.

Jackson, D. N. (1969). Multimethod factor analysis in the evaluation of convergent and discriminant validity. *Psychological Bulletin, 72*(1), 30-49.

Jonckheere, A. R. (1954). A distribution-free *k*-sample test against ordered alternatives. *Biometrika, 41*, 133-143.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*(2), 109-133.

Kalleburg, A. L. & Kluegel, J. R. (1975). Analysis of the multitrait-multimethod matrix: Some limitations and an alternative. *Journal of Applied Psychology, 60*(1), 1-9.

Neave, H. R. & Worthington, P. L. (1988). *Distribution-free tests*. London: Unwin Hyman.

Pedhazur, E. J. & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, N.J.: Erlbaum.

Sawilowsky, S. S. (2002). A quick distribution-free test for trend that contributes evidence of construct validity. *Measurement and Evaluation in Counseling and Development*, 35, 78-88.

Sawilowsky, S. S. (Ed.) (2007). *Real data analysis*. Charlotte, N.C.: Information Age Publishing.

Schmitt, N. & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 10*(1), 1-22.

Stanley, J. C. (1961). Analysis of unreplicated three-way classifications, with applications to rater bias and trait independence. *Psychometrika, 26*(2), 205-219.

Sternberg, R. J. (1992). Psychological Bulletin's top 10 "hit parade". *Psychological Bulletin, 112*(3), 387-388.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Erlbaum.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*(1), 1-26.

ABSTRACT

ROBUSTNESS TO NON-INDEPENDENCE AND POWER OF SAWILOWSKY'S I TEST FOR TREND IN CONSTRUCT VALIDITY

by

JOHN L. CUZZOCREA

May 2007

Advisor: Dr. Shlomo Sawilowsky

Major: Educational Evaluation and Research

Degree: Doctor of Philosophy

The Multitrait-Multimethod Matrix by Campbell and Fiske (1959) provided a foundation for evaluating construct validity. Unfortunately, efforts to apply statistical techniques to analyze the matrix have been unsuccessful. As a result, Sawilowsky (2002) created a quick, distribution-free statistic as a solution to analyzing the matrix. However, a question regarding the robustness of the statistic to departures from independence as well as the power properties of this test required further examination. The purpose of this study was to examine the Type I error rate when independence is violated and the power properties of the Sawilowsky I test. In addition, a modified version of the Sawilowsky I test was developed and examined in terms of its robustness when independence has been violated and its power. Ad hoc critical values were determined to improve the statistical power of this approach to analyzing the Multitrait-Multimethod Matrix.

AUTOBIOGRAPHICAL STATEMENT

JOHN L. CUZZOCREA

EDUCATION:
January 2003 – April 2007
Wayne State University, College of Education.
Ph.D. in Educational Evaluation and Research, Theoretical and Behavioral Foundations.

September 2000 – April 2002
University of Windsor, Faculty of Education
Master's in Administrative Education.

WORK EXPERIENCE:
February 1999 - Present
Teacher
Windsor-Essex Catholic District School Board

FACULTY APPOINTMENTS:
May 2006 – Present
Adjunct Instructor
Wayne State University, College of Education, Department of Educational Evaluation and Research, Theoretical and Behavioral Foundations.

PUBLICATIONS:

Cuzzocrea, J. L. & Sawilowsky, S. S. (2007). Pietro Paoli, Italian algebraist. *Journal of Modern Applied Statistical Methods, 6*(1).

Sawilowsky, S. S. & Cuzzocrea, J. L. (2007). Joseph Liouville's 'Mathematical Works of Évariste Galois. *Journal of Modern Applied Statistical Methods, 6*(1).

HONORS AND AWARDS:
2004-2006 Wayne State University: Graduate Professional Scholarship
1996-1998 University of Windsor: In-Course Scholarship

JOURNAL/EDITORIAL ACTIVITY:
2005 – Present
Editorial Assistant
Journal of Modern Applied Statistical Methods