1-1-2011

Reliability generalization: lapsus linguae

Julie Marilyn Smith
Wayne State University,

RELIABILITY GENERALIZATION: *LAPSUS LINGUAE*

by

JULIE M. SMITH

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

2011

MAJOR: EDUCATIONAL EVALUATION
AND RESEARCH

Approved by:

_____

Advisor Date

_____

_____

_____

_____

DEDICATION

To my parents, Paul and Marilyn Scheer and to my husband Mark Smith.

ACKNOWLEDGEMENTS

During the past six years as I worked through the various requirements necessary to earn my degree, I have been blessed with many people who have helped me along the way in a variety of ways. I could not have accomplished this work without their guidance, encouragement and support.

First and foremost, I wish to thank Dr. Shlomo Sawilowsky, my major advisor. I truly do not know the words to express my appreciation and gratitude for the many hours he has spent working with me over the course of my program. His patience in explaining concepts and helping me place them in proper context and his willingness to share his vast knowledge of statistics, measurement and research have allowed me to learn, to challenge myself and to grow throughout my tenure as his student. I consider myself very fortunate to have him as my advisor and mentor.

I would also like to thank my other committee members; each of them has contributed to my education and has helped to make my experience at Wayne State rewarding and worthwhile. Dr. Gail Fahoome has taught me much about research, evaluation and statistics – and in the process I not only learned but I had a lot of fun investigating topics to which many people have an aversion! Dr. Lynda Baker guided me through research in Library and Information Science and gave me opportunities to put my new knowledge to use working on real-world projects. When one of my committee members suddenly passed away, Dr. Stephen Hillman graciously offered to join my committee and to support me through the dissertation process. I am very grateful to him for agreeing to work with me and am thankful for his collaboration. Dr. Michael Addonizio joined my committee at the "last minute" and I thank him very much for his time, his help and his willingness to work with me on short notice.

I have had the distinct pleasure to not only study with and learn from these individuals but also to work with them in a variety of different capacities at the university. As such, I have been able to get to know them both as individuals as well as learned professors and scholars. I am grateful to be able to say that I worked with them and also that I have the utmost respect and admiration for all, both personally and professionally.

A very special thanks goes to my cousin, Howard Scheer, who worked tirelessly on my FORTRAN, who patiently explained each step of program coding (along with various idiosyncrasies associated with logic languages) and who made the programming portion of this project a lot less intimidating. Without his expertise, assistance and humor this project would have been much more stressful and would have taken a lot longer to complete.

I would also like to recognize: My friends at the Cranbrook Institute of Science and in the Educational Administration Department at WSU, particularly Geari Viney and Dr. William Hill, for giving me much-needed laughter and for ensuring that I kept everything in perspective. My parents who gave me – among many others – the gifts of education, independence, encouragement to try new things and who taught me the value of having a strong work-ethic. Finally, to my wonderful husband who patiently supported me throughout my studies and who inspired me to keep going when the going got tough. I treasure you all and thank you.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

During the 20th century researchers such as Cronbach, Lord, Novick, Nunnally, and Spearman contributed to the development of an extensive body of knowledge in psychological test theory. Psychological test theory – also known as psychometrics – is the branch of psychology dealing with methods for the design, administration, and interpretation of quantitative tests to measure psychological constructs such as intelligence or personality traits (Anastasi, 1976; Nunnally, 1978; Allen & Yen, 1979; Cronbach, 1990). Test theory is primarily concerned with methods for estimating the extent to which error influences measurements in a given testing situation and devising methods to overcome or minimize error so that test results are accurate and dependable (Crocker & Algina, 1986). As Qualls and Moss (1996) described, tests in the behavioral and social sciences "are employed as informational tools in a variety of contexts such as educational planning, career development, clinical treatment plans, counseling interventions, and a multitude of research investigations" (pp. 209-210). Two examples of tests are the *Stanford-Binet Intelligence Quotient (IQ)* and the *Minnesota Multiphasic Personality Inventory (MMPI)*. As the psychometrics field has grown and expanded, researchers have developed a plethora of new tests and have sought new ways to think about how test results should be used – as well as the importance of test accuracy, reliability and validity.

According to Allen and Yen (1979), measurement theory is a branch of applied statistics, the aims of which are to describe, categorize and evaluate the quality of measurements, and to develop methods for constructing new and better measurement instruments. Stevens (1946) defined measurement as "the assignment of numerals to objects or events according to rules" (p. 677). Thye (2000) explained that there are three components in measurement: (1) the *object* that

is measured, such as Joe Brown; (2) the *instrument* used to assign numerals to the object, such as a 20-item IQ examination; and (3) the *occasion* on which the measurement is taken (p.1279). The process of measurement must be structured and carried out according to a set plan in order to obtain accurate results. Nunnally (1978) provided the following examples:

> Scientists measuring the surface temperature of planets in our solar system should achieve very similar results if they follow the same measurement procedure and use the same instruments. If two different examiners administer the same intelligence test to the same person at two different times, the test is reliable if it results in approximately the same scores on both occasions. (pp. 3-4)

Thus, different individuals using the same measuring instrument (e.g., ruler, thermometer, test) and the same measurement protocol should obtain similar results if they are measuring the same attribute of objects or persons (e.g., length, temperature, intelligence, aptitude). Nunnally (1978) explained why it is critical for measurements from the same measuring devices to yield similar results, by stating: "To the extent to which an approach to measurement provides very much the same result regardless of opportunities for variations to occur, then it is reliable and one can generalize from any particular use of the measurement method to a wide variety of other circumstances in which it might be employed" (p. 191). Measurement instruments which are able to be used in multiple situations by a variety of researchers are a requirement in scientific investigation: without them conclusions based on research would be less useful because it would not be possible to ascertain whether the measurements taken were accurate and replicable.

Researchers in the natural or physical sciences have an advantage over those in the social sciences. The former typically work with visible, or directly measureable, variables such as weight, concentration or density – most of which have specific, standardized instruments from which to obtain a measurement. The latter are many times concerned with latent variables, or

traits not directly measurable, that must be indirectly assessed by the process of measuring related attributes. Knapp (1977) explained:

> One thing that differentiates measurement in the social sciences from measurement in the physical sciences is that most of the instruments used in the social sciences consist of "items" which are gathered together to form a "test." A measurement of a person's height is a single number which can be read off a scale, but a measurement of a person's intelligence is arrived at by combining scores obtained on various test items. Therefore, the determination of the reliability of a social measurement usually poses quite a different problem than the determination of the reliability of a physical measurement… (p. 237)

Whether attempting to measure planetary surface temperature variation or the spatial-relations aptitude of an individual, a common requirement is present: the need for repeatability, or consistency, in measurement. "Inconsistent measurements are a bane to persons engaged in research. Scientists have learned to repeat their measures several times when it is important to obtain results in which they can be confident. The average of a set of repeated measurements provides a more precise estimate of what is being measured than does a single measurement" (Traub & Rowley, 1991, p. 171). In most psychological measurement, however, it is difficult to obtain multiple measurements from the same subject, as noted by Traub and Rowley (1991), "unfortunately, the measuring procedures we use in education usually cannot be repeated as easily as can some of the measuring procedures used in the physical sciences" (p. 171). Ascertaining and understanding the measurement reliability of tests in the social sciences is critical because, as Romano and Kromrey (2009) stated, "Ideally, social science research is conducted using measurement instruments that produce valid and reliable information" (p. 404).

Study Rationale

Vacha-Haase (1998) proposed the application of meta-analytic techniques similar to those used in validity generalization studies to the study of reliability, calling the method

Reliability Generalization (RG). The following paragraphs summarize the various facets underlying RG, as well as its background; in-depth descriptions of all are provided in Chapter 2.

Because RG is presented as a meta-analytic technique, it is important to understand the basics of meta-analysis. As defined by Egger, Smith and Phillips (1997), meta-analysis "is a statistical method of combining data from several studies to more precisely analyze the results and explain differences in research conclusions" (p. 1533). Meta-analysis integrates the results of several independent research studies which are considered to be "combinable." Although the concept of meta-analysis predated Glass by a half century, he spearheaded its application in the social and behavioral sciences in his 1976 Presidential Speech to the American Educational Research Association. Rodriguez and Maeda (2006) described his approach:

> Glassian meta-analysis begins with the collection of all relevant studies with liberal inclusion criteria. The empirical outcomes of these studies are transformed into a common effect size metric. The distribution of these outcomes is described, and study level characteristics are used to explain variation in outcomes. Related methods start with this general approach, using various modifications. (p. 306)

As conceptualized by Glass, meta-analysis presented an opportunity for psychological researchers to avoid conducting new research by synthesizing the data collected in previous studies, combining outcomes and drawing conclusions based on examining the collective results. Glass (1976) described meta-analysis as "the analysis of analyses…statistical studies of a large collection of analysis results from individual studies for the purpose of integrating the findings" (p. 3). As explained by Yin and Fan (2000), meta-analysis provides a systematic approach to "make sense out of a large amount of seemingly inconsistent findings from many primary research studies" (p. 221). Since its introduction, "meta-analysis has experienced an exponential growth in its application to disciplines across a wide spectrum of social and behavioral sciences" (Yin & Fan, 2000, p. 220).

With respect to psychometrics, meta-analytic techniques have recently been employed to study validity, which is generally defined as "the appropriateness, meaningfulness, and usefulness of the specific inferences from the test" (Sartori & Pasini, 2007, p. 361). It is not an *instrument itself* that is valid or invalid: validity is concerned with establishing evidence that the *use* of a specific instrument is congruent with the intent or purpose of the test. "Thus, when we address the issue of validity with respect to a particular test, we are addressing the issue of the validity of the scores on that test for a *particular purpose,* and not the validity of the test or instrument per se: A given test might be used for a number of purposes" (Morgan, Gliner & Harmon, 2001, p. 731). Validity is a theoretical concept that is not directly measured, it is established by drawing inferences; evidence is gathered to support the use of test results for making decisions based on the instrument's intended measurement(s). The use of meta-analysis to examine validity is referred to as *validity generalization*. "Validity generalization studies have been conducted to describe the extent to which validity evidence for scores are generalizable across research contexts" (Romano & Kromrey, 2009, p. 405).

In contrast to validity, reliability – at its most basic level – is the extent to which a test provides consistent information (Nunnally, 1978; Crocker & Algina, 1986; Anastasi, 1976). Feldt and Brennan (1989) described the "essence" of reliability analysis as the "quantification of the consistency and inconsistency in examinee performance" (p. 6). Consistency means that all parts of a test, or different forms of a test that are intended as interchangeable, actually measure the same thing; reliability coefficients quantitatively measure such consistency. Reliability is also concerned with stability: that a test measures the same thing at different times, on different occasions, or in different situations. Lastly, from a statistical standpoint, reliability is concerned with measurement error. Reliability is measured each time a test is administered; it can be

determined via several different available techniques depending on how a test is administered and the sample of examinees who took the test.

Vacha-Haase presented the application of meta-analytic techniques to reliability coefficients as an appropriate way to characterize "(a) the typical reliability of scores for a given test across studies, (b) the amount of variability in reliability coefficients for given measures, and (c) the sources of variability in reliability coefficients across studies" (Rodriguez & Maeda, 2006, p. 306). In presenting this method for analyzing reliability, Vacha-Haase put forth the idea that, similar to validity evidence, reliability coefficients are properties of the scores generated by a test, as opposed to the instrument itself.

It is presumed that Vacha-Haase based the idea for RG on the fact that reliability coefficients differ from study to study and from one test administration to another; this is presumed because (1) the study was based in meta-analytic theory, which seeks to integrate "seemingly inconsistent findings" (Yin & Fan, 2000, p. 221), and (2) no historical research is provided to explain why reliability coefficients vary between studies using the same test (although many reasons were already well known and well documented in the field).

In the study, the *Bem Sex Role Inventory* (BSRI) was the test used to illustrate how RG studies would work. In the discussion, the author noted that "reliability coefficients were fairly variable" (Vacha-Haase, 1998, p. 11), and stated that RG provides a method to help identify the sources of variation in score reliability. For example, it was reported that different methods for computing reliability coefficients result in different reliability coefficient values and that sample size is a predictor for reliability. In addition, it was noted that test form impacted the reliability coefficient calculated. Finally, the author stated that RG indicated which study features do not predict variations in score reliability, for example, variations in response formats (Vacha-Haase,

1998). Most of the findings in the study – for example, different methods for computing reliability result in different values – merely confirmed issues with reliability that have long been known among psychometricians.

Based on the findings, however, the author asserted that the study "…illustrates how important it may be to recognize that reliability does not inure to tests" (Vacha-Haase, 1998, p. 12). Interestingly, the initial RG study meta-analyzed reliability coefficients obtained from a test administered by different researchers to different samples under different conditions, not the scores generated by the test (Sawilowsky, 2000). Also of interest are the facts that no substantive argument was provided to bolster this statement, nor was any scientific evidence or theory-based rationale presented to illustrate how – or to explain why – the psychometric property of reliability would reside within the scores generated by a test and not the instrument itself.

Since the proposed RG method was published, several researchers have identified a variety of problems with the technique. The first to raise substantive questions about RG was Sawilowsky (2000), who presented a number of arguments against the method. In addition to a historical review, Sawilowsky described issues with RG related to several aspects of reliability itself, such as, "statements about the reliability of a certain test must be accompanied by an explanation of what type of reliability was estimated, how it was calculated, and under what conditions or for which sample characteristics the result was obtained" (p. 159). He also explained that Vacha-Haase's approach focused on the *data* obtained from a test administration; this is in direct opposition to traditional psychometric thinking which focuses analyses on *instruments* themselves. To describe this approach, he coined the term *datametrics*, meaning that in RG, "reliability and other desirable characteristics are considered to be concomitant with the

scores or data at hand instead of being a property of the instrument or its use" (Sawilowsky, 2000, p. 160).

Research Question

This study will explain why RG does not provide a proper research method for the study of reliability. This research will focus on sampling error; results will illustrate that the reliability of a test will vary across test administrations based on the size and composition (random vs. non-random selection) of the sample. This study will address the following:

Research Question: Can the fluctuation in estimates of reliability under proper experimental conditions be fully explained via classical measurement theory without resorting to reliability generalization?

Human Participants

Human participants will not be used in this research. This study will employ computer generated (i.e., simulated) data.

Limitations

Classical Test Theory (CTT) will provide the theoretical basis for reliability as studied in this research. Item Response Theory, Cronbach's Generalizability Theory and other frameworks under which reliability may be assessed will not be considered. This study will be restricted to test-retest reliability. The measurement error associated with CTT may differ from that associated with other theoretical frameworks, and measurement error of test-retest reliability does differ from that of other reliability measures. As described by Schumacker and Smith (2007), measurement error in classical measurement models used in testing and assessment differs based on the type of reliability used. In the models, the source of measurement error and the ensuing reliability coefficients will differ on the basis of test design and administration. Test-

retest reliability specifies error primarily due to changes in subjects over time; whereas other forms of reliability attribute error to other causes, for example, internal consistency specifies error primarily due to poor item sampling, and alternate form reliability focuses on variations between item samples on different test forms (Schumacker & Smith, 2007, p. 394). This study will use 2 × 1,000,000 universe subsets (which will be considered to represent a sample from a larger universe and referred to as mini-universes) with Nunnally's domain-sampling approach to calculate test-retest reliability between parallel testlets of varying sizes (n = 5, 10, 15, 25, 50, 75, 87 and 100); results from a population of greater magnitude may provide slightly different results.

CHAPTER 2

LITERATURE REVIEW

Overview

This review examines relevant literature from research in the social sciences related to psychometric theory, test theory and the debate regarding reliability and reliability generalization. Measurements, scales, and the psychometric properties reliability and validity are discussed, followed by detailed descriptions of the various theories and frameworks upon which this research is based and designed.

Elements of Test Theory

In order to understand how an instrument's reliability and validity are assessed and interpreted, a basic understanding of how measurements are obtained within the social and behavioral sciences is needed. One method by which behavioral scientists strive to measure latent (i.e., not directly measureable) variables is via the use of tests, or scales. Scales are questionnaires and other measures designed to quantify constructs such as intelligence, aptitude, or attitude (Crocker & Algina, 1986). In a 1987 issue of the *Journal of Counseling Psychology* specifically focusing on quantitative foundations, Dawis stated "scales are ubiquitous features of counseling psychology research" (p. 481).

The use of scales in research assumes that their measurements possess adequate psychometric properties, that is, that they reliably and validly assess the constructs being studied. As Qualls and Moss (1996) described:

> Assuming an alignment exists between developmental purpose and the intended contextual application, validity evidence and reliability evidence are by far the two most crucial elements that underlie judgments regarding the quality of scores derived from instruments. The effect of these two psychometric properties on resulting inferential decisions must be understood. (p. 211)

Regrettably, misunderstandings persist in the psychological literature related to tests and their psychometric properties and confusion exists regarding the differences between reliability and validity, what each represents and how each should be considered with respect to a particular instrument.

Reliability and validity – though related – provide distinctly different information about a test. Whereas validation is an ongoing process based on the integration of data from many sources, reliability is measured at a point in time. Popham (2009) further expressed their differences by stating, "as far as validity is concerned, the term doesn't refer to the accuracy of a test. Rather, it refers to the accuracy of score-based inferences about test takers. … In the case of reliability, however, it's the test itself that is or isn't reliable. That's a whopping difference." (p. 77) This same concept is reiterated by Weller (2001) who stated "valid tests are accurate assessments of what is taught or what they purport to measure. Reliable tests yield consistent, non-contradictory results" (p. 33) and, as Sartori and Pasini (2007) noted, "from a statistical point of view, reliability refers to the extent that measures are relatively free of random error and are consistent in the numbers assigned to properties of objects and events. … validity is not only a property of measures, but it refers to the truthfulness of the inferences that are drawn from measures" (p. 361). Finally, Sijtsma (2009) aptly portrayed the distinct differences by summarizing:

> Typically, discussions about reliability are more mathematical than discussions about validity. The reason is that reliability formalizes one particular technical property of the test score whereas validity involves assessment of what the test measures, either by means of exploration or the nomological network or by means of theory testing. Unlike reliability estimation, validity research revolves around the use of substantive knowledge about the attribute under consideration and decision making … Reliability is a much narrower concept… (p. 178)

Test Reliability

Reliability, the focus of this research, is concerned with consistency, stability, and measurement error, and its assessment plays a critical role in interpreting quantitative test results. This study follows the traditional psychometric approach to the study of reliability which considers it to be a metric property of a test itself, as opposed to the "datametric" approach used by Vacha-Haase (1998) and described by Sawilowsky (2000). However, it is not possible to discuss test reliability without mentioning test scores: it is the scores obtained from various test administrations that are used to calculate reliability coefficients. This is not a contradiction. Consider the following physical measurement examples from Sawilowsky (2000):

> If three independent measures of an object yield values of 120, 770, and 18, there is no point in discussing whether the scale is measuring in pounds or kilograms. The values are not even similar to the same power of 10. Because there is no consistency, there is no evidence of reliability. Therefore, the question of what the scale is measuring, or the purpose for using it, is moot. In contradistinction, suppose another scale yields values of 12.99, 13.01, and 13.00. These values are more consistent in comparison with the previous scale. The consistency via repeated measures is one type of evidence that indicates that the *scale* is reliable in measuring whatever it measures. (p. 197, emphasis added)

Thus, it is not the *measurements or scores themselves* that are reliable; it is the *device or instrument* used to take the measurements that is reliable. Using a reliable instrument to take the same measurement multiple times according to a set protocol should result in consistent scores that differ only slightly due to the effect of measurement or sampling error. As Ponterotto and Ruckdeschel (2007) explained, "estimates of reliability inform researchers as to what proportion of total score variance is due to true variance versus error variance" (p. 997). Measurement error diminishes the reliability of a score obtained for an individual from a single administration of a test. Revelle and Zinbarg (2009) defined this aspect of reliability as "the fraction of test variance that is true score variance" (p. 145); and, as Moss (1994) clarified, "theoretically, reliability is

defined as the degree to which test scores are free from errors of measurement" (p. 6). Thus, to the extent that measurement error is slight, a measurement may be said to be reliable (Nunnally, 1978).

Although the concepts underlying reliability appear simple, in actuality reliability is a complex concept fraught with subtle difficulties. Much debate continues among researchers with respect to the calculation, interpretation and use of reliability coefficients for tests. Part of the confusion stems from the terminology used in the context of reliability, words such as, consistency, precision, repeatability and agreement have all been used to describe reliability. Weir (2005) commented about this varied jargon, stating that "intuitively, these terms describe the same concept, but in practice some are operationalized differently" (p. 231). Revelle and Zinbarg (2009) noted, "the problem of how to assess reliability has been with us ever since Spearman (1904) introduced the concept of correction for attenuation and that of split half reliability" (p. 145). Another facet of the issue with reliability and tests derives from the fact that several ways exist to assess reliability, including: test-retest, alternate form, and internal consistency. As explained by Gliner, Morgan and Harmon (2001):

> While each method to assess reliability gives some measure of consistency, they are not the same. To say that an instrument is reliable has relatively little meaning because each statement of reliability should specify the type(s) of reliability, the strength of the reliability coefficient, and the types of subjects used. Before using an instrument, an investigator should evaluate reliability and how it was established. (p. 488)

Each approach to reliability is subject to different sources of error and the choice of which to employ is dependent in part on the specific test to be administered, the subjects to whom the test will be given, and the testing conditions. Regardless of which approach is used, however, the same statistic applies: the correlation coefficient is calculated to determine the reliability coefficient for an instrument. The range (–1.00 to +1.00) and interpretation of reliability

coefficients mirror those of a typical correlation: the closer to +1.00 the reliability coefficient, the more reliable the test. In terms of measurement error, Weller described: "Because error contributes to variance in reliability coefficients, the closer reliability coefficients are to +1.00, the more the test is free from error variance" (2001, p. 35).

Reliability Types: Alternate Form, Internal Consistency & Test-Retest

Alternate form reliability is calculated when two different versions of a test are created by writing different – yet comparable – test items, and the two forms are administered twice to the same individuals (usually within a two-week time interval). Thus, "alternate-form reliability describes the consistency of students' performances on two different (hopefully equivalent) versions of the same test" (Popham, 2009, p. 77). The correlation between observed scores on two alternate test forms is usually referred to as the coefficient of equivalence. Developing alternate forms of the same test is time-consuming and difficult to achieve. Although this type of reliability estimation helps correct for memory and practice effects, errors related to unintended differences in content between original and parallel test questions, changes in the trait being measured over time, and the difficulty associated with administering a test to the same subjects under the same conditions at two different times all represent potential sources of error with alternate form reliability.

Internal consistency estimates were developed as a means of estimating reliability without repeated testing. Internal consistency refers to the interrelatedness of a set of test items; the degree to which items in a test are associated due to what they share in common. Because it requires only one test administration to calculate, it is efficient and practical for many studies. Many different approaches to calculating internal consistency reliability are available, examples include: Cronbach's alpha, Spearman-Brown, and Kuder-Richardson 20. Each method uses item

scores from one test administration to estimate the reliability of an instrument by calculating correlations among test items; the goal is to assess content stability by determining if every item on a test correlates with every other item. Popham (2009) explained, internal consistency reliability "describes the consistency with which all the separate items on a test measure whatever they're measuring, such as students' reading comprehension or mathematical ability" (p. 77). Similarly, Sireci, Thissen and Wainer stated:

> Internal consistency estimates of reliability are based on a simple extrapolation from the (average) correlation among the items on one form (that one has) to the (average) correlation between those items and some other forms' items (that one does not have): The extrapolation is that the (average) correlation among the items one has is the same as the correlation of those items with the (hypothetical) second forms' items. (1991, pp. 244-245)

Also, as Green, Lissitz, and Mulaik noted: "If a set of items is measuring the same or similar properties and the property comprises a single continuum or dimension, the items should all covary to some extent. For a fixed number of items, the greater and more consistent the inter-item correlations the more reliable the composite" (1977, p. 828). The primary sources of error with measures of internal consistency are poor test construction and/or test items that do not accurately reflect the trait being measured.

Test-retest reliability, sometimes referred to as a stability coefficient, is a measure of the sustainability of test scores over time. These coefficients refer to the consistency of scores when the same test is administered to the same individuals under the same circumstances on two different occasions. A test-retest reliability coefficient is estimated by correlating the observed scores from two test administrations. This procedure is subject to errors due to carry-over effects of memory and/or practice (Dimitrov, 2002, p. 786). Test-retest reliability estimates are most appropriate for measuring traits that tend to remain stable across the time period between the two test administrations: As Charter (2003) noted, "we expect traits such as intelligence to hold up

well over time and have relatively high [test-]retest coefficients, whereas we expect states such as depression to fluctuate over time and have relatively low [test-]retest coefficients" (p. 290). Test-retest compares measurement stability across the same subjects at different times thus – among the different reliability methods – it is the closest reflection of the proposed reliability generalization method, which compares reliability coefficients across different test administrations. For this reason, it is the method that will be employed to study test reliability in this study.

Classical Test Theory vs. Item Response Theory

Proposed by Charles Spearman in 1904, the true-score model – also known as classical test theory (CTT) – has been a principal theory guiding reliability estimation. CTT postulates a linear model which links an observed test score (X) to the sum of two latent variables, a true (T) and an error score (E). This model is represented as $X = T + E$, and is founded on the proposition that measurement error, a latent variable, is a component of observed scores (Traub, 1997). Most measurements in social and educational research are subject to error in that repetition of a measurement process is unlikely to produce an identical result: it should be noted that error (E) is not simply one term, but a combination of different errors that can vary both between subjects and within subjects across different test administrations. For example, as Woodhouse, et al. (1996), described:

> Most measurements in educational or other social research are subject to error, in the sense that a repetition of a measurement process does not produce an identical result. For example, measurements of cognitive outcomes in schools such as scores on standardized tests can be affected by item inconsistency, by fluctuations within individuals and by differences in the administration of the tests and in the environment of the schools and classes where the tests take place. Measurements of non-cognitive outcomes also, such as children's behavior, self-concept and attitudes to school, can be similarly affected. (p. 201)

Three specific insights formed the backbone of the classic and elegant CTT model: "A recognition of the presence of errors in measurements, a conception of that error as a random variable, and a conception of correlation and how to index it" (Traub, 1997, p. 8). In 1966, Novick stated that "the classical test theory is basically a nonparametric estimation model" (p. 5), and as Hambleton and Slater (1997) later described:

> The true score of an examinee is defined as the examinee's expected score across infinite replications of parallel-forms of the test of interest. An error score is the difference between the construct of interest (i.e., true score) and the observable data (i.e., the test score) and every effort is made to minimize factors contributing to error such as improper sampling of content, poorly constructed items, guessing, cheating, misleading responses (e.g., responses reflecting social desirability), and flaws in the administration process such as test speededness. By reducing both random and systematic errors in the testing process, test score and true score are close and reliability and validity are increased. (pp. 21-22)

Due to the fact that there are two latent variables – T and E – for each examinee, the CTT equation is not solvable unless some assumptions are made. "The assumptions in the classical test model are that (a) true scores and error scores are uncorrelated, (b) the average error score in the population of examinees is zero, and (c) error scores on parallel tests are uncorrelated" (Hambleton & Jones, 1993, p. 255). These assumptions, however, are relatively weak and can be met fairly easily when using real test data in the model. Benefits to using CTT models to study measurement problems include: no large sample size requirement; simpler mathematical analyses compared to item response theory (discussed below); conceptually straightforward model parameter estimation; and CTT analyses do not require strict goodness-of-fit studies to ensure a good fit of model to test data (Hambleton & Jones, 1993). Hambleton and Jones also provided a summarization of the primary limitation of CTT: "one main shortcoming is that they are sample dependent, and this dependency reduces their utility. They are most useful when the examinee sample is similar to the examinee population for whom the test is being developed"

(1993, p. 255). Thus, measurements are partially dependent on the both the test and the examinee sample and this dependence may influence the usefulness of CTT models to some degree.

Along with CTT, other measurement models are used to study the psychometric measures of tests. One of these, item response theory (IRT) is a general statistical theory about examinee performance on items and tests and how test results relate to the abilities measured by the items in a test. IRT suggests that (1) examinee performance on a test relates to a single latent ability or trait underlying responses to items on a particular measurement instrument, and (2) that the relationship between the examinee's ability and his/her probability of providing a correct answer can be described by a monotonically increasing curve. This s-shaped curve is called an item characteristic curve (ICC); an ICC shows the probability of examinees at varying abilities answering an item correctly. ICC's are estimated for each item in a test and it is a person's ability score (denoted θ) that determines the probability of a person to correctly answer any test item (Hambleton & Slater, 1997). The underlying assumption is an expectation that individuals with greater ability have higher probabilities of providing correct answers to test questions compared to those with lower abilities (Dimitrov, 2002).

Although IRT focuses on the accuracy of ability scores, CTT is based upon the accuracy of examinee's observed scores. IRT models offer some benefits to investigating measurement problems such as reliability. Hambleton and Jones (1993) listed the main benefits of IRT as:

1. Item statistics are independent of the groups from which they were estimated.
2. Scores describing examinee proficiency are not dependent on test difficulty.
3. Test models provide a basis for matching test items to ability levels.
4. Test models do not require strictly parallel tests for assessing reliability.

They also stated that "item response theory models tend to be complex and model parameter estimation problems tend to arise in practice. Model fit too can be a problem – it is still not completely clear how problems of model fit should be addressed, especially problems that related to test dimensionality" (Hambleton & Jones, 1993, p. 259). IRT models are also more technically demanding compared to classical models.

With respect to error, neither classical test theory nor item response theory assume that error variance is the same for different individuals. As Sitjsma (2009) explained "applications of tests constructed by means of classical test theory use one standard measurement error for all individuals. Applications of item response theory use a standard deviation of the estimated latent variable conditional on the true value, which varies in magnitude across the scale" (p. 184). However, Brennan (2001) noted "IRT is primarily an elegant scaling model, not a measurement model, because IRT has no explicit role for error of measurement relative to investigator-specified replications" (pp. 304-305).

With respect to CTT models, Novick (1966) stated that classical test theory holds a long and distinguished history of application to the technology of test construction and test utilization (p. 1); although other theories related to testing and measurement have been proposed in recent times, Heiser (2006) affirmed, "classical test theory is not obsolete" (p. 458). CTT offers a straightforward approach in which reliability can be effectively examined and it has much history to support its use in modern measurement problems. Because reliability is in part dependent upon measurement error, CTT is the better option for this study as it provides a simpler approach for a study of this metric; therefore, CTT will serve as the basis upon which test-retest reliability estimates for instruments will be investigated in this study.

Test Types: Tau Equivalent, Congeneric Equivalent and Parallel

Lee, Brennan, and Frisbie (2000) advanced the following definition: "A *test* is a generic characterization of the multiple forms that might be created using the test specifications" (p. 12). One example of this is the *Stanford-Binet Intelligence Scale*, a standardized test that assesses intelligence and cognitive abilities in both children and adults. Many different questions are used to comprise one individual test form, but by selecting different questions from a large pool of potential questions, many forms of the same test could be created, all of which could be used interchangeably to measure the same underlying construct: intellectual functioning. This is true of many psychological and educational tests: even if only one particular test currently exists, it may still be considered one form of the test because other, different but equivalent tests could – at least in theory – be designed using the same specifications to measure the same construct (questions within a test may be thought of as indicators for constructs).

Because classical test theory involves measuring latent constructs, the psychometric literature characterizes tests in three ways: congeneric, tau-equivalent, and parallel. Congeneric tests are measures of the same latent trait, that is, they measure a single underlying construct; however, they may have different scale origins and units of measurement and they may vary in precision. For congeneric tests, the correlations between true scores will be unity, but the variances of the true scores may vary. In a test that is tau-equivalent, the test components all measure the same latent trait and it is assumed that true scores have equal variances in the population of respondents. Tau-equivalence implies that the correlations between true scores are all equal to unity and that the variances and covariances of the true scores on the components of the measure are all equal: these tests can be interpreted as measuring a single underlying construct (Dimitrov, 2002; Osburn, 2000). Tests are parallel when "the correlations between true

scores are all unity and the variance and covariance of the component true scores of the measure are all equal. In addition, the components are equally reliable. Parallel components are unidimensional with equal factor loadings and equal error variances" (Osburn, 2000, p. 344). Parallel measurements within the context of CTT are defined as interchangeable or equivalent (Novick, 1966). Although the characteristics of experimental data are seldom precisely known, this study uses Monte Carlo methods which will allow tests to be defined as parallel because their means and variances can be held constant in the calculation of the reliability coefficients.

Instrument Reliability Considerations

A number of elements in combination affect a test's ability to produce reliable results. The nature of reliability – that it is not static – has been studied and described by many different researchers in a variety of disciplines during the past century. According to Sawilowsky (2000), one of the first to systematically expound on this was Symonds (1928), who defined reliability as the correlation between two comparable tests and, in reference to reliability, stated:

> It is customary to group the factors influencing test reliability into: (1) Factors in the construction of the tests themselves and (2) factors in the variability of the individuals taking the tests. For certain factors this is a clear cut distinction; for others both irregularity in test construction and the variability in individuals seems to be operative. (pp. 74-75)

Table 1 lists the factors Symonds identified as influencing test reliability. As Sawilowsky (2000) noted, "it has been acknowledged throughout the literature that tests' 'reliability' estimates are alterable" (p. 198); this is evidenced in Symonds work as well as in the work of many other researchers. For example, as Schumaker and Smith (2007) noted, "In CTT, the…reliability coefficient is affected by several factors: (a) the number of items, (b) group homogeneity, (c) the time limit, (d) reverse scoring, and (e) negative interitem correlation" (p. 401). Thus, it is necessary to consider the test itself, conditions under which a test is

administered, and the sample of individuals taking the test; the interactions among these factors influence test reliability.

Table 1: Factors Influencing Test Reliability (Symonds, 1928)

Factors Related to…	
The Test Itself	***The Individual Taking the Test***
Number of test items	Speed required to take the test
Length of time needed to take a test	Accuracy in taking the test
Range of difficulty of test items	Incentive and effort (e.g., motivation)
Evenness in scaling	Obtrusion of competing ideas (e.g., previous experiences, concentration level, outside influences)
Interdependence of test items (e.g., an answer to one item is dependent on a previously answered item)	Distractions (e.g., temperature of the testing room, noise, etc.)
Scoring (i.e., objective vs. subjective scoring)	Accidents during testing (e.g., breaking a pencil, getting a defective test booklet, etc.)
Scoring inaccuracy (i.e., errors in scoring)	Illness, worry, excitement of examinee (i.e., emotional and physical state)
Chance in answering	Time interval between test repetitions
Position of the correct item in a multiple choice list	Cheating
Homogeneity of test material	Learning curve
Common experiences of test subjects required to complete the test (e.g., using examples or language with which not all students can necessarily identify)	
Time of year in which a test is administered	
Inclusion of extraneous material in a test (e.g., items not covered in class or in the textbook)	
Catch questions (i.e., questions that must be answered due to sudden insight, not learning)	
Emotional tinge of words in test items	
Length of test items	
Choice of words and terms	
Poor sentence structure	
Inadequate or faulty directions	
Test formatting (e.g., printing, spacing, paragraphing, margins, font size, etc.)	

Instrument design.

With respect to an individual test form, a test with few items will generally yield a lower reliability coefficient estimate than a longer test. Reliability coefficients depend on variation among individual responses to items and – because a test represents a sample of items – if the sample is too small, chance in the selection of items will play a large part in determining the scores that examinees obtain (Traub & Rowley, 1991). Increased test reliability obtained from increasing test length, however, follows the law of diminishing returns. For example, doubling the length of a test with reliability of .60 increases the reliability to .75, tripling the length of the same test will increase the reliability to .81, and lengthening the test by five times increases the reliability to .88 (Crocker & Algina, 1986).

When considering item type, objectively scored tests (e.g., multiple-choice tests) tend to be more reliable than tests whose scoring process includes some subjectivity (e.g., essay tests). Traub and Rowley (1991) explained two reasons for this: "…first, they [objectively-scored tests] eliminate scorer inconsistency as a source of measurement error, and, secondly, they are able to cover more content, thus reducing the unreliability that can result from luck in the selection of questions" (p. 177). Well-written items are also integral to obtaining legitimate reliability estimates for a test; items should be crafted using proper language, grammar, vocabulary, and structure. If test items are unclear or vague, error will be introduced by the varying interpretations placed on the items by examinees.

Finally, item difficulty plays a role in reliability. When an item is very difficult for the examinees being tested, many may leave the item unanswered or simply guess. "Guessing adds an element of randomness to scores: some gain a mark through chance; others of equal ability are not so rewarded" (Traub & Rowley, 1991, p. 177). Conversely, if an item is so easy that all

examinees are able to answer it correctly, it does nothing to enhance test reliability (although it does not detract from it either). Items contributing the most to test reliability are those that discriminate, that is, items on which examinees who possess the knowledge and/or skills needed to answer the question correctly have a better chance of successfully responding compared to those who lack the necessary knowledge and/or skills. In order to maximize reliability a test should be designed at a level of difficulty that matches the abilities of the examinees; neither too easy for the group, nor too difficult (Traub & Rowley, 1991).

Test administration.

Test administration refers to the physical conditions under which a test is given. The testing environment itself may cause variations in examinee responses due to noise, temperature, lighting, seating, or other physical aspects. In addition, the directions provided to examinees for responding to questions, any time limits imposed for completing the test and the test administrator are all conditions that can impact the reliability of an instrument. To the extent that these factors vary from one administration of a test to another, and to the extent that the conditions in test environment affect some examinees differently from the way they affect others, "test scores will vary for reasons other than differences among the examinees in the knowledge and skill being tested" (Traub & Rowley, 1991, p. 177). For example, if a test is timed, one of the abilities required by examinees is that of working quickly. Reliability may actually be enhanced in this case because timing adds an attribute on which examinees may differ in a consistent manner: ability to respond speedily; the difficulty is that this ability may not be an intended measurement of the instrument. Also, directions to test-takers are important because they help control the effects of guessing among examinees. Instructing examinees to answer every question, including those that cannot be correctly answered from knowledge, should have

the effect of reducing differences among examinees based on guessing (Schumaker & Smith, 2007; Traub & Rowley, 1991).

Sample.

The range of true differences in abilities within a group of examinees tested influences the size of a test's reliability coefficient. For the same test, a group of examinees having similar abilities will yield a lower index of reliability compared to a group in which the ability range is more widespread. For example, "a statistics anxiety test given to a group of gifted and talented students would yield less variation than among a general group of education students" (Schumaker & Smith, 2007, p. 401); Traub and Rowley (1991) further explained:

> We are looking at the capability of the test to make reliable distinctions among the group of examinees with respect to the ability measured by the test. If there is a great range of ability in a group, a good test should be able to do this very well. But if the examinees differ very little from one another, as they will if the test covers a limited range of tasks in which all examinees are highly skilled, reliable distinctions will be difficult to make, even with a test of high quality. (pp. 177-178)

In addition, when using a standardized test in the field, a researcher typically administers the test to a small, non-random sample of subjects. Test developers, by contrast, "norm" a test by administering it to a very large, random, representative sample of subjects for whom the test was created. The reliability that is calculated based on this normative sample provides test administrators with the "reliability estimate of a test for general purposes" (Sawilowsky, 2000, p. 170). Morrow and Jackson (1993) stated that "depending on sample size and calculated reliability, reported sample reliabilities may deviate greatly from the population parameter" (p. 353), or normed-estimate. Thus, it should be expected that the reliability calculated for a test administered to a small, non-random sample would differ from that provided by the test developer; as Traub (1994) explained:

> The usual reliability experiment provides a sample estimate of the reliability coefficient for the population. Had a different sample of persons participated in the experiment, the reliability coefficient obtained would almost certainly have been a different number. (p. 66)

Measurement Error

Given the vast array of evidence identifying factors which influence test reliability, it seems clear that variance in reliability coefficients will be observed between different test administrations. Not only is this concept clear, it is explainable by another well-known statistical concept: measurement error. The combination of a specific test, its administration, and group of examinees impact the reliability of a test by introducing random variation – measurement error – into test scores: as stated by Nunnally, "some error is involved in any type of measurement" (1978, p. 190). Error must be taken into account when studying the psychometric properties of instruments used in research. Two basic types of measurement error are examined with CTT: systematic and random.

Systematic error impacts each observation in the same way each time a measurement is recorded (Nunnally, 1978). The temperature of the testing room is one example of systematic error; if an examinee is uncomfortably warm or cold every question answered will be influenced to some extent by the environment. Thye (2000) referred to these errors as transient, describing them as "those factors that impact individual responses the same way within a given experimental session, but vary across distinct experimental sessions" (p. 1284). Systematic, or transient, errors create variability that cannot be attributed to the independent variable. These errors are not reproduced exactly at each test administration; they vary between each testing situation (Nunnally, 1978; Thye, 2000). Random errors of measurement, by contrast, affect each observation by chance and they vary within each test administration. Random errors "are caused by momentary fluctuations in the way people feel about, attend to, or behave toward the

experimental treatment" (Thye, 2000, p. 1285). Like systematic errors, random errors are not caused by the independent variable, but they affect the dependent variable; thus why they are measurement errors.

Although both types of error are important to consider, random errors are commonly considered the more troublesome of the two (Nunnally, 1978; Nunnally & Bernstein, 1994; Thye, 2000). As Nunnally (1978) described:

> Systematic biases [errors] contribute to the mean score of all subjects being studied…the mean score of all subjects is not very important in studies of individual differences and in most psychological experiments. Random errors are important in all studies, because to the extent they are present, limits are placed on the degree of lawfulness that can be found in nature. (p. 190)

Thye also noted that, "a constant error (e.g., adding 6 units to each score) will not reduce the correlation between two measures, [but] random measurement error will always attenuate measures of association" (2000, p. 1279). Because reliability is determined using correlation coefficients, this attenuation can impact the analyses resulting from the test data. If a test result is influenced by a large amount of measurement error, an improper inference may be made with respect to the reliability of the instrument (Woodhouse, et al., 1996). Nunnally (1978) related measurement error and test reliability in the following manner:

> To the extent to which measurement error is slight, a measure is said to be *reliable*. Reliability concerns the extent to which measurements are *repeatable* – when different persons make the measurements, on different occasions, with supposedly alternative instruments for measuring the same thing and when there are small variations in circumstances for making measurements that are not intended to influence results. In other words, measurements are intended to be *stable* over a variety of conditions in which essentially the same results should be obtained. … If the data obtained from experiments are influenced by random errors of measurement, then the results are not exactly repeatable. Thus, science is limited by the reliability of measuring instruments and/or the reliability with which scientists use them. (p. 191, emphasis in original)

Classical Test Theory: Reliability and Measurement Error

In CTT, reliability is defined in terms of variations of an examinee's responses from his or her true score. Stated another way, "in classical reliability theory, the candidate's 'true score' is the mean of a hypothetical distribution of scores the candidate would earn over many replications of the measurement process" (Livingston, 2004, p. 334). Sijtsma (2009) referred to this as an individual's *propensity distribution*, explaining that:

> This is an individual's distribution of test scores that would result from the endless administration of the same test under the same circumstances, such that different test scores are the result of independent repetitions of the administration procedure. The correlation between only two of these repetitions in the population of interest is the test-score reliability, ρ_{xx}'. The variation in an individual's propensity distribution is interpreted as measurement error variance. (p. 179)

A perfectly reliable instrument would always yield the same score on each administration to a particular individual: That score would be the person's true score for the instrument, a theoretically pure representation of a person not influenced by the particulars of any single measurement, i.e., measurement error. No measuring instrument is perfect however; variability or error is associated with all tests and, on any test occasion, random errors cause a person's observed score to differ from his or her true score. "These random deviations from true score are caused by measurement error. The distribution of observed scores around true score defines the distribution of measurement error" (Traub & Rowley, 1991, p. 174). Nunnally (1978) described measurement error as follows:

> The wider the spread of obtained scores about true scores, the more error there is in employing the type of instrument. The standard deviation of the distribution of errors for each person would be an index of the amount of error. If the standard deviation of errors were much the same for all persons, which usually is assumed to be the case, one standard deviation of errors could typify the amount of error to be expected. This typical standard deviation of errors is called the *standard error of measurement*, σ_{meas}. The size of σ_{meas} is a direct indication of the amount of error involved in using a particular type of instrument. (p. 193, emphasis in original)

Thus, "reliability (symbolized r_{kk}) is the proportion of total instrument variance that is true score variance. This relationship is illustrated by the following equation,

$$r_{kk} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2}$$

The theory of reliability is the theory of random measurement error" (Traub & Rowley, 1991, p. 174).

In Classical Test Theory (CTT), the following statements are made with respect to error (E in X = T + E): error is assumed to be uncorrelated from one measurement to another, error is assumed to be unrelated to a person's true score, errors have a mean of zero over repeated measurements, and it is only scatter around a true score – not bias in the true score itself – that is presumed due to errors of measurement (Cronbach, 2004; Knapp, 1977; Lord, 1955; Traub & Rowley, 1991). Error can be quantified by a reliability coefficient and, because error can cause observed scores to differ from true scores in either direction, 68% of the measurements can be expected to fall within one standard error of measurement (± 1 SEM) of a person's true score and 95% fall within two SEM (± 2 SEM). The SEM is a determination of the amount of variation or spread in the measurement errors for a test, in other words, it is the difference between an obtained score and its theoretical true score counterpart (Harvill, 2005). A reliability estimate is used in the SEM calculation as follows:

$$SEM = s\sqrt{\left(1 - r_{xy}{}'\right)},$$

Where s is the standard deviation of the observed scores and $r_{xy}{}'$ is the reliability estimate (Harvill, 2005). Morrow and Jackson (1993) stated:

> The potential variability of the reliability estimate has a predictable influence on the SEM. If the population reliability is .80 and the population standard deviation is 10, the SEM would be 4.47 in the population. If one considers the potential

> variability of the reliability estimate due to sample size, the standard error of measurement would also show variability and could be considerably larger than reported. For example, if N = 5, the estimated SEM could range from 1.183 to 9.997 based on a 95% CI with r_{xy}' limits of .986 and .001. (p. 354)

Thus, the SEM quantifies the precision of scores on a test and it is sometimes referred to as typical error. The SEM provides an "absolute index of reliability" (Weir, 2005, p. 237). With respect to SEM, in 2004 Cronbach stated: "Here we have a direct report on the degree of uncertainty about the person's true level of performance" (p. 410).

Reliabilty Calculation

In all cases, the reliability of a test is estimated from the obtained test scores of a group of examinees (Harvill, 2005). The statistic used to assess reliability is the correlation; the average size of the correlations among items is directly related to the variance of total scores: high positive correlations among items make for a large variance among test scores (and vice versa for low correlations). Test reliability depends upon positive average correlations among items; in addition, a highly reliable test has a larger variance than a less reliable test (Nunnally, 1978). Thye (2000) described how, by using indicators (i.e., test questions), the amount of variance caused by true score variance – the reliability of a test – may be estimated. He stated:

> When a study is replicated, neither transient nor random response error will reproduce in precisely the same way. Yet, true score tendencies of the trait will reproduce because the trait is stable. Thus, by correlating the scores from [questions within the instrument], the researcher can estimate true score variance net of transient and random response error. … The goal here is to assess consistency across parallel instances of the same theoretical construct. When scores from two different measures are correlated, only the common sources of variance add to the correlation coefficient. If these measures are perfectly identical in all regards, then the correlation will register 1.0 for the two indicators. (p. 1287)

Error contributes to variance in reliability coefficients (correlations), "the closer reliability coefficients are to 1.00, the more the test is free from error variance" (Weller, 2001, p. 35); a

reliability coefficient of 0.80, indicates that the scores from one form of a test could be expected to correlate 0.80 with the scores from another equivalent form of the test. (Note that reliability can, in theory, be negative. This does not typically occur, however, because a group of examinees is usually working toward a similar goal of obtaining high scores on a test, which causes a positive correlated relationship between correct responses to test items and total test scores.) The size of the estimated reliability coefficient for a test will depend on the specific sources of errors that may potentially affect the test results. Specific error types depend in part on the way the test is administered, and in part to which form of reliability estimation is used. "In studying the information in test manuals, it is important to note not only the size of the reliability coefficients reported, but also the type of estimate reported, the kinds of error that it acknowledges, and the population of examinees that was sampled" (Traub & Rowley, 1991, p. 176). For example, three potential sources of error, which in turn are associated with different reliability types are: trait stability over time (test–retest reliability), domain or content sampling (alternate form reliability), and item variability or interrelatedness of items (internal consistency reliability) (Rodriguez & Maeda, 2006).

The correlation approach may be employed to study test-retest, parallel or alternate forms reliability (situations in which two test administrations are conducted) or internal consistency reliability (in which a test is administered only one time). Tzeng and Welch (1999) stated that this approach "is also used to index the reliability of a test having multiple (k) items by computing intercorrelations among the items of the test" (p. 119). In this study test-retest is the reliability method employed, thus, the Pearson Product-Moment Coefficient of Correlation (Pearson's r) will be used to calculate the correlation coefficients for testlets according to its formula:

$$r_{XY} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{\left[n\Sigma X^2 - \left(\Sigma X\right)^2\right]}\sqrt{\left[n\Sigma Y^2 - \left(\Sigma Y\right)^2\right]}}$$

Overview: Reliability Generalization

In 1998, a method for studying the reliability of a measuring instrument across multiple administrations was introduced (Vacha-Haase, 1998). It was coined reliability generalization (RG) and was intended to generalize reliability results from different administrations of the same test. As noted in Chapter 1, RG was fashioned after a meta-analytical methodology employed to study test validity: validity generalization. However, as also described in Chapter 1, validity is a completely different concept than reliability. Validity relates neither to an instrument itself nor to obtained test results; it is instead concerned with the inferences, conclusions or propositions that are made based upon the intended purpose of a test. Validity is not determined by a single statistic or study, but by a body of research that demonstrates the relationship between the test and its success at measuring the behavior, concept, construct or trait it was intended to measure. Reliability, by contrast, estimates the consistency of a measurement – or more simply – the degree to which an instrument measures the same way each time it is used under the same conditions. Reliability (estimated at a point in time) is concerned with the accuracy of the actual measuring instrument or procedure, validity (assessed over time) has to do with the degree to which a test accurately reflects whatever the test was designed to measure. Because (1) the same instrument (i.e., test) may be used to measure different aspects of the same construct, or (2) multiple different instruments may be available to measure the same construct (for example, Kramer and Conoley (1992) listed 14 measures of general anxiety and 15 measures of depression with high construct validity), meta-analytic generalization of the use of a test – validity – is appropriate; the study of reliability, however, is not amenable to such a procedure.

RG and reliability: The proposition.

Vacha-Haase (1998) proposed that reliability generalization (RG) could be used to characterize the mean measurement error variance and sources of variability of variances across studies. Presumably based on observed variance among reliability coefficients calculated for the same test across different studies – in this case the *Bem Sex Role Inventory* (BSRI) – the author asserted that reliability was a metric property of test scores (or measurements resulting from using an instrument), not a property of the test (or measuring instrument) itself. It was argued that "reliability does not inure to tests, but rather to scores" (Vacha-Haase, 1998, p. 213). The support for this assertion, however, is not described as being based on any statistical or theoretical underpinning, but instead comes from the editorial policies of the journal *Educational and Psychological Measurement*, in which the editor proscribed contributing authors from making statements such as "the reliability of the test" or "the test is reliable" (Thompson, 1991, p. 843). This directive was set forth based in part on the variety of ways in which reliability can be estimated for an instrument, and also in part due to years of poor reporting practices among researchers.

The *Bem Sex Role Inventory* RG study.

According to Vacha-Haase (1998), the BSRI is a widely used instrument in the area of gender orientation research; for this reason it was chosen as the example for conducting a RG study. Published in 1974 by Bem, the BSRI provides an assessment of gender identity by examining how males and females describe themselves and how these self-descriptions fit with various attributes that are typically recognized as being masculine or feminine. Vacha-Haase (1998) stated, "This popular inventory has been translated into many languages and used with

various populations. The measure has been the focus of numerous measurement studies and has also been employed in a wide range of substantive studies" (p. 9).

To conduct the RG study, Vacha-Haase (1998) searched one database (PsycINFO) for articles published between 1984 and 1997; the search resulted in 628 relevant articles (p. 10). Of these, the majority were not included in the RG research for several reasons: (1) the researchers did not report any reliability; (2) reliability was only reported from the testing manual or from other studies; and/or (3) because reliability was not reported in a meaningful manner. Only 57 of the articles reported reliability coefficients for the data at hand and were included in the RG study, however, among these, some articles contained more than one reliability coefficient. "For example, several articles provided reliability coefficients separately for the two genders, various ethnic backgrounds, ages of participants, or population settings (e.g., clinical vs. non-institutionalized)" (Vacha-Haase, 1998, p. 10); due to this, 87 pairs of reliability coefficients were analyzed in the study.

The method employed in conducting the RG analysis was to dummy code each of the variables, such as reliability coefficient type (e.g., test-retest, Kuder-Richardson, etc.), BSRI test form (long or short version), study participant genders, and so on. According to Vacha-Haase, "the first task in the reliability generalization meta-analysis was to characterize both typical reliability and the variability of M and F score reliability coefficients, each expressed in squared metrics" (1998, p. 11). Regression analyses were used to examine how coded study features predicted reliability coefficients and results were reported both as beta weights and coefficients. In addition, multivariate reliability generalization was also conducted due to the fact that the instrument used to exemplify the RG method has two different scales (i.e., test forms).

In the discussion, it was reported that different methods for computing reliability coefficients result in different reliability coefficient values and that sample size is a predictor for reliability. In addition, it was noted that test form – long vs. short version – impacted the calculated reliability coefficient. Finally, the author stated that the "results also indicate which study features do not predict variations in score reliability. Such features in the present study included the origins of the sample (e.g., students or not) and variations in the Likert-type response format" (Vacha-Haase, 1998, p. 12).

Issues with Reliability Generalization

After Vacha-Haase's article was published, Sawilowsky (2000), and Knapp and Sawilowsky (2001) published several articles detailing many different issues – statistical and otherwise – with RG studies that were neither acknowledged nor addressed by Vacha-Haase (1998). Since Sawilowsky's first publication in 2000, several other researchers have reiterated and commented on the many problems he identified regarding the proposed RG method to study reliability. The potential problems with using the RG method to study reliability across studies include the:

- comparison of reliability coefficients that were obtained via different reliability calculation methods;
- failure to account for the different errors associated with each type of reliability;
- consideration of different test forms and/or formats as the same test and comparing their reliabilities;
- ignoring differences in test administration conditions (including failing to determine if proper test protocol was followed);
- inappropriate coding of groups (confounding independent variables); and

- misspecification of samples in the RG analysis.

In addition, many issues with reliability reporting exist in the literature. As described in the following paragraphs, these many problems – both individually and in combination – provide evidence that brings the scientific integrity of the proposed RG method into question.

Comparing different reliabilities.

Incorporating different types of reliability coefficients in a single RG analysis violates a basic meta-analytic principle: studies must be combinable in order to be included in a meta-analysis. For example, in the BSRI-RG study Vacha-Haase used multiple regression with the dependent variable being either test-retest correlation coefficients, Cronbach's α, or KR-20 coefficients. As Sawilowsky (2000) clearly pointed out "reliability can be estimated from different perspectives…and can be affected by different conditions" (p. 159). Dimitrov (2002) followed, explaining why combining reliability coefficients calculated via different methods is problematic for RG:

> Cronbach's α can either underestimate the reliability when the measures are not at least essentially tau-equivalent or overestimate the reliability when correlations among errors occur. Also, the test-retest correlation coefficient can be a reasonable estimate of reliability only if the measures are essentially tau-equivalent and have equal error variances. Thus, the dependent variable used by Vacha-Haase (1998) might be a mixture of apples and oranges, as implied also by her own conclusion that "the results…indicate that internal consistency and test-retest reliability coefficients seem to present considerably different pictures of score quality." (p. 792)

In fact, the problem with comparing different reliability coefficients has been known for a very long time. In 1951, Cronbach responded to a criticism that split-half coefficients fail to provide the same information as internal consistency coefficients, by stating:

> The two coefficients are measures of different qualities and should not be identified by the same unqualified appellation 'reliability.' A retest after an interval, using the identical test, indicates how stable scores are and therefore can be called a coefficient of *stability*. The correlation between two forms given

> virtually at the same time, is a coefficient of *equivalence*, showing how nearly two measures of the same general trait agree. (p. 298, emphasis in original)

More recently, Rodriguez and Maeda (2006) commented that "some meta-analysts may argue that a function of meta-analysis is to explain variation in effects, and in the case of reliability coefficients, all types of reliability should be included while coding each type of reliability uniquely to assess the degree to which type of coefficient explains variation in reliability" (p. 309). They go on to explain how difficulties arise when combining different reliability coefficient types: "for instance, coefficient alpha varies with the number of items used. The magnitude of test-retest correlations is largely a function of the time interval between testing and relevant intervening experiences. The magnitude of split-half reliability estimates varies as a function of the method of splitting the form in half" (Rodriguez & Maeda, 2006, p. 309).

The various approaches to calculating reliability coefficients provide different types of information about a test. This has been discussed repeatedly throughout the literature, Sawilowsky (2000) wrote "many authors have noted, reliability paradigms and their coefficients are simply not interchangeable" (p. 159), reaffirming Pedhazur and Schmelkin (1991) who asserted that "reliability coefficients obtained via one paradigm constitutes different information than that from a reliability coefficient obtained via a different paradigm of measurement" (Sawilowsky, 2000, p. 161). In 2009, Popham reiterated, saying "because these three incarnations of reliability [split-half, alternate form and internal consistency] constitute meaningfully different ways of thinking about a test's consistency…*approaches to reliability are not interchangeable*" (p. 77, emphasis added).

Different test forms and lengths.

Another problem with RG results from analyzing tests that have different forms and/or lengths. It is known that reliability is related to test length: in general, the shorter the test the

lower its reliability. It seems clear that – if a RG study is to be conducted – the researcher should only use tests that were exactly the same in the comparisons. However, "in RG studies, it is typical to find instruments altered or modified and administered to a wide variety of individuals under various conditions. Even after adjusting alpha for the number of items used in a given study, changes in content can result in changes in the nature of the construct (independent of the number of items altered), making RG less meaningful" (Rodriguez & Maeda, 2006, p. 319).

Along with length is the issue of test format, for example, Caruso's (2000) study of the *NEO Personality Scales* revealed that the test has both long and short forms as well as multiple versions which contain different questions. Similarly, Vassar and Crosby's (2008) study of the *UCLA Loneliness Scale* documented three different versions of the test. The BSRI itself has two different versions, one long and one short. As Vacha-Haase (1998) described, the long version consists of 60 adjectives or short phrases which are split into three sets of 20 items that are considered either masculine, feminine or neutral; the short version consists of only 10 items for each of the traits. It would seem reasonable then, that although the two versions purport to measure the same construct, their reliabilities should be considered separately. Roth and Sackett (1991) commented on this issue stating: "Note that if the different tests are perfectly parallel measures of the same psychometric construct they are completely interchangeable and no problems arise. However, approximately or nominally parallel or congeneric measures are not the same as parallel measures and therefore are not interchangeable" (p. 325). As Dimitrov (2002) summarized, "using different test forms in a single RG analysis can also cause problems because, as indicated by previous research, the reliability depends on factors such as item response format, (positively/negatively) wording of stems, and type of scales" (p. 793).

Test administration conditions.

Because reliability concerns measuring the degree of consistency or stability of an instrument over repeated administrations, then "it follows that an investigator must be able to specify what would constitute a replication of a measurement procedure in order to provide any meaningful statement about the degree to which a measurement procedure is reliable" (Brennan, 2001, pp. 295-296). In other words, to compare reliabilities calculated for the same test on different occasions, it is important to ensure that the test was administered in the same manner on each occasion. Feldt, Woodruff and Salih (1987) provided an example related to time allotted for a verbal fluency test:

> Reliability comparisons…among measures which require different amounts of testing time would be reasonable only in those situations in which existing standardized instruments are compared. In such circumstances the researcher must administer the instruments at the lengths and within the time limits for which the norms apply. (p. 99)

Thye (2000) stated "reliability coefficients can be estimated (1) within each condition of an experiment, and/or (2) across the same conditions of multiple experiments" (p. 1288). He explained that, under such conditions, reliability coefficients are not sensitive to experimental manipulation and do exactly what they should; gauge the amount of total variance caused by true score variance within a specific experimental treatment. This leads to a very important principle: "measurement reliability is only meaningful when calculated within each condition of an experiment, or across the same conditions of distinct experiments" (p. 1288). He then went on to state that reliability coefficients can be interpreted as the potential for data collected during one test administration to correlate with other data collected under "identical conditions" (p. 1296). Therefore, if one cannot be certain that the protocol developed by the test developer is followed properly within different testing administrations it is not possible to have confidence that the

reliabilities calculated from various replications of a test are comparable. Sawilowsky (2000) summarized, stating:

> In the absence of considerations of designed experiments, or even in violation of the test manufacturer's standardized administration and scoring procedures, then it would not be surprising to find some dispersion of the reliability estimates in these studies. (p. 198)

Sample issues.

Another difficulty with RG across studies results from the inappropriate coding of groups (by gender, ethnicity, age, etc.) or, even worse, from biasedness of the instrument on groups coded with independent variables. Sawilowsky (2000) explained that using dummy-coded regression is appropriate when working with unequal sample sizes; however, this is not the case when non-random samples are used. He stated:

> It is a statistical artifact of measures of internal consistency that their correlational engine is less stable when the range is restricted. The most commonly cited distortion is in terms of traits or characteristics of the sample, which is referred to the problem of group homogeneity. (p. 167)

When randomization is violated, reliability may increase or decrease; in addition, Vacha-Haase's failure to consider this issue resulted in misspecified dummy coded regression. For example, Sawilowsky (2000) noted that Vacha-Haase's coding design for gender in her 1998 BSRI study led to the confounding of independent variables because the gender of study participants was coded twice. Despite the fact that the misspecification of variables by a meta-analyst is problematic, there is an even more fundamental issue with RG: reliability statistics are sample dependent. As Rodriguez and Maeda (2006) stated "if for no other reason than sampling error, study effects will always vary at some level, and meta-analysts should recognize this condition" (p. 310). The instrument norming process uses large random samples, but tests used in the field are typically administered to small, non-random groups; thus, it seems reasonable to

expect that reliability coefficients calculated at each test administration would vary from those published by test developers.

In Vacha-Haase's (1998) RG-BSRI study, 89% of the studies included in the meta-analysis for which sampling plans were known used non-random samples (as estimated by Sawilowsky, 2000, p. 168), a point which is not stated in the original research. Sawilowsky (2000) commented that random sampling violations causes reliabilities calculated on a sample to differ from those computed by test developers; he cited a collection of literature supporting this and also quoted Wood (1991) who said "'That is why there is no point in talking about the reliability of a test or examination, unless it is in terms of a strictly defined population'" (Sawilowsky, 2000, p. 170). Rodriguez and Maeda (2006) explained that various estimates of reliability capture specific sources of measurement error, thus resulting in different sampling distributions. They summarized the issue as follows:

> Instrument developers should have reported psychometric properties of their instrument, including item statistics and an assessment of dimensionality. These results can often be found in technical manuals or articles reporting the development of the instrument. At the same time, caution is always needed, because these statistics are typically sample dependent (as is coefficient alpha) and may not hold in all populations or in all samples from any given population, particularly when those samples are nonrandom. (p. 308)

Based on the literature, sufficient evidence exists to support the tenet that different estimates of reliability should not be combined.

Reporting issues.

Although there is much debate regarding reliability reporting practices among researchers, the dispute stems not from concerns related to scores vs. test, but instead from reporting failures in published research. Many researchers do not report differences between specific study sample calculated reliability and that provided by the test developers for the norm-

group. In addition, many studies fail to report the method used to calculate reliability, which makes a difference in terms of the types of errors that are considered in the reliability coefficient. Finally, there is a tendency for researchers to fail to report any reliability coefficients at all: It is well documented that many studies neither cite norm-group nor sample specific reliability. In 2000, Sawilowsky stated "in my view, authors ought always to report these reliability coefficients from manuals and other sources along with the reliability estimate obtained from, and a description of, the researcher's own sample" (p. 170). Recently, Romano and Kromrey (2009) echoed his thought and made a statement that, because reliability can fluctuate across studies, it is recommended that researchers evaluate the reliability of their studies and report their results (p. 405). However, inconsistencies with respect to reporting reliability coefficients issues continue to be prevalent in the literature: Yin and Fan (2000) described the issues when they reported:

> Of the articles reviewed for the present study, only 7.5% [90 out of 1,200] of the articles reported meaningful reliability coefficients for the data used in the studies. The overwhelming majority of the articles (80.1%) reviewed in this study did not even mention the reliability issue in their reports, let alone provide reliability estimates for their data. (p. 216)

More recently, when conducting a study on the reliability of the *Ways of Coping Scale* (WOCS), Rexrode, Petersen and O'Toole (2008) noted: "of the 661 citations found during the data collection process, 92 were usable" (p. 267). Vassar's (2008) study of the *Satisfaction with Life Scale* (SWLS) discovered similar issues with non-reporting; he stated "of the 196 articles, 62 (31.6%) reported internal consistency estimates for the data at hand, yielding a total of 77 reliability coefficients" (2008, p. 49). Similarly, Beretvas, et al. (2008), reported in a study of Nowicki and Strickland's *Internality-Externality Scale* (NSIE), "a total of 166 articles were identified and evaluated, although only 19 (11.45%) of the studies reported estimates for the

samples involved" (p. 102). Finally, in their study of the *UCLA Loneliness Scale*, Vassar and Crosby (2008) related: "of the 213 articles remaining, 80 reported internal consistency estimates for the data at hand" (p. 602). Such inconsistent practices have, in part, contributed to the confusion surrounding psychometric test properties with respect to both their uses and interpretations.

In 2002, Baugh commented that "unfortunately, disregard for the central role of measurement in empirical investigations has been and remains widespread in the social sciences" (p. 256). His statement relates to the poor reporting practices noted above, and reflects the opinions of many social scientists with respect to psychometrics. He goes on to state: "because so few researchers report reliability [for their own data], and even fewer interpret effects in light of reliability, the practical impact of this affect attenuation is largely unknown" (Baugh, 2002, p. 260). Sijtsma recently reiterated "test construction and test practice are plagued by bad habits…and [are] in need of more direction" (2009, p. 169). These are not novel observations, in 2000, Sawilowsky argued: "statements about the reliability of a certain test must be accompanied by an explanation of what type of reliability was estimated, how it was calculated, and under what conditions or for which sample characteristics the result was obtained" (p. 159). According to Vacha-Haase (1998), this non-reporting issue provides a strong argument for RG. It is, however, a completely separate and distinct matter from that of variation among reliability coefficients. Although measurement integrity is crucial to scientific research and tests that yield consistent results are of paramount importance to social scientists, it is important to address the non-reporting issues plaguing social and behavioral research; this issue however fails to provide a basis for RG.

Measurement Error

Since Vacha-Haase (1998), several researchers have conducted and published studies that purported to examine reliability generalization. However, they tended to interpret RG study results by indicating conditions under which instruments yield scores with greater or lesser measurement error. A typical example of an RG interpretation was provided by Capraro, Capraro and Henson (2001): "in sum, *measurement error* [emphasis added] in MARS [*Mathematics Anxiety Rating Scale*] scores appears to increase in adult samples and perhaps in other homogenous age groups" (p. 384). Vacha-Haase's 1998 article title itself *Reliability generalization: Exploring variance in measurement error affecting score reliability across studies*, suggests that the criterion of interest was actually measurement error, as opposed to test reliability.

In their critique, Rodriguez and Maeda (2006) presented various examples of RG studies and concurred that: "the titles and interpretations suggest the value of interest is measurement error; however, the values studied are reliability coefficients" (p. 319). It would seem that analyzing sources of variation is a more reasonable approach to study differences in reliability coefficients between different test administrations, because error contributes to the magnitude of the reliability coefficient calculated. As Sawilowsky (2000) pointed out by quoting Nunnally and Bernstein (1994), "it is meaningful to think of a test as having a number of different reliability coefficients, depending on which sources of measurement error are considered" (p. 170). Dimitrov (2002) further described:

> The accuracy of measurement for a study relates to information about the study-specific population reliability and its sample estimates. Within the classical framework, such information is provided, for example, by confidence intervals for alphas or reliability for congeneric measures but not by reliability box-plots in RG across studies. (p. 795)

Additionally, he argued that RG across studies does not provide adequate information about the accuracy of measurement with a specific study. Such information is instead provided by the standard error of measurement (SEM) and the reliability coefficients for the study sample and their population parameter estimates (Dimitrov, 2002). Hence, it may be stated that reliability coefficients do not reflect variations in measurement error – instead, measurement error contributes to differences between the reliability calculated based on a norm-group and the reliability a researcher calculates from a small, non-random sample of examinees.

RG Proponents and Studies

Since Vacha-Haase (1998), other researchers have conducted RG studies to determine why calculated reliability coefficients may vary for the same test when it is administered to different samples. Following the same procedure and method described by Vacha-Haase, RG has been conducted on several tests including the: *Beck Depression Inventory* (Yin & Fan, 2000); *Ways of Coping Scale* (Rexrode, Petersen & O'Toole, 2008); *Internality-Externality Scales* (Beretvas, et al., 2008); *Minnesota Multiphasic Personality Inventory* (Vacha-Haase, et al., 2001); *UCLA Lonliness Scale* (Vassar & Crosby, 2008); *Big Five Factors Personality Assessment* (Viswesvaran & Ones, 2000); *NEO Personality Scales* (Caruso, 2000); and *MacAndrew Alcoholism Scale* (Hakim & Viswesvaran, 2002).

Throughout these various studies several common themes are evident: lack of reliability reporting by researchers in general, inconsistent reliability methods, test variations, and differences in examinee samples. As documented herein, many problems with RG exist, not the least of which is that RG studies published to date rely on one article as the basis for employing the method: this is disconcerting because, as previously noted, other than stating a conviction that reliability is a metric property of scores, no supporting scientific theory or evidence is put

forth to give credence to the statements made. Also as described here, several aspects among RG studies have been examined and nearly all have been shown to be incompatible to studying in combination (e.g., it is inappropriate to compare reliability coefficients calculated using different reliability methods).

A few investigators have put forth an effort to expand upon RG as it was originally proposed. Baugh (2002) proposed a method for correcting effect sizes for reliability. Baugh's study was based upon the notion that reliability is a property of test scores and that score reliability tends to attenuate effect size estimates. Baugh did not provide any background for his assertion that reliability inures to scores; discussions in the article simply state this as fact. He introduced a technique to correct for study "artifacts" (p. 258), and stated that these artifacts may be systematic or unsystematic; it seems clear that the artifacts to which he referred are measurement error because he went on to state:

> The majority of systematic artifacts attenuate the population correlation ρ. Specific artifacts attenuate ρ differently, and knowledge of the artifact allows for the quantification of influence on the effect estimate. For example, reliability coefficients for scores on dependent variables provide an estimation of the amount of random measurement error present in the dependent variable scores. (2002, p. 259)

Baugh presented formulas to correct for the reduction in effect size that may be caused by score unreliability, and concluded that greater measurement error leads to smaller values in reliability coefficients which in turn reduces effect size (p. 259). He ends with a caution that researchers should report both reliability and effect size for their data because reporting both acknowledges "the presence of measurement error in all analyses and calls attention to its impact" (p. 260). As noted with respect to Vacha-Haase (1998), it would seem that the metric of interest in Baugh's study was measurement error, not reliability.

Charter (2003) wrote an article discussing formulas for combining different types of reliability coefficients (test-retest, alternate forms, split-half, coefficient alpha, etc.) from several samples to "produce the exact reliability one would compute if one had raw data from the samples" (p. 643). The formulas, which require sample means, standard deviations, sample sizes and reliability coefficients, were proposed to be used specifically with meta-analytic RG studies as described by Vacha-Haase (1998). Charter stated that use of the formulas "would allow the research to examine the average shrinkage of sample coefficients relative to the best-guess population coefficient generated from these formulas. The investigator would then point out that the population coefficient is most likely higher" (p. 644). Citing Thompson (2002) Charter concluded "reliability generalization studies appear to be limited by two things, (a) the creativity and insightfulness of the researchers and (b) the information reported in the prior studies examined in the reliability-generalization meta-analysis…these formulae should provide creative opportunities for researchers" (p. 646). Unfortunately, Charter failed to address known problems with RG, to further explicate the reasons why different reliability coefficients should not be combined, or to explain the effect of non-random sampling on reliability estimates.

Henson (2004) followed Charter's formulas to "characterize the reliability of a scale's scores across studies" (p. 818). He stated "too few researchers realize that reliability is a function of the obtained scores from a scale and is not a property of the scale itself (see Thompson & Vacha-Haase, 2000; Henson, 2002). Reliability can and does fluctuate from sample to sample when using the same measure" (p. 818). He explained that, because it varies from sample to sample, reliability generalization can be employed to examine the amount of variability in reliability estimates. Henson asserted that combining reliability estimates "provides a more accurate estimate of the population reliability parameter" and that it gives insight into the

precision of the estimate (p. 819). He used fourteen studies of the *Coopersmith Self-esteem Inventory* – all of which reported Cronbach's alpha – to calculate a combined reliability estimate and confidence interval (CI); the combined reliability coefficient was computed to be 0.90 with a CI (.893, .901). From this data he concluded that the combined coefficient was greater than any individual reliability estimate "due to increased variability of the combined data" and that "the CI width generally decreases as the obtained reliability and sample size increase" (p. 819). Interestingly, among the fourteen studies, all had reliability coefficients between .78 (CI: .716, .835) and .89 (CI: .838, .932) with one exception, a study that used only 29 participants reported reliability as .68 (CI: .488, .826). These findings are very similar to what Sawilowsky (2000) identified in Vacha-Haase's (1998) study: that descriptive statistics for the BSRI showed remarkably narrow confidence intervals for applied research (p. 169). This leads once again to the question of what scientific evidence is driving the notion that reliability is a metric property of scores as opposed to tests when it seems clear that it is the proper and systematic use of an instrument that affords consistency to measurements.

Fan and Yin (2003) published an empirical study on examinee characteristics and reliability. Citing Thompson and Vacha-Haase (2000) and Yin and Fan (2000), they asserted that "it is generally recognized that measurement reliability should be considered as the characteristic of test scores rather than the test itself" (p. 357). The study they reported was designed to assess the extent to which group heterogeneity and group performance levels affect reliability, specifically Cronbach's alpha. Fan and Yin used data from two tests one criterion-referenced test used in Texas high schools, the *Texas Assessment of Academic Skills* (TAAS), and one norm-referenced test, the *Iowa Test of Basic Skills* (ITBS). They tested for the effect of group heterogeneity by drawing random samples from a large pool (TAAS, n = 50,000; ITBS, n =

10,000) of test scores and systematically restricting the range of scores on which reliability was calculated. They first drew random samples (TAAS, n = 500; ITBS, n = 100) from the entire pool, then ordered the data and drew random samples from only the middle 90%, 80% and 50%, thus restricting the range and lowering the variance with each successive sampling. To examine the effect of group performance on reliability, they drew the same sized random samples for both tests but for this metric they drew first from only the upper and lower 75% and then from only the upper and lower 50% of the data.

Fan and Yin noted that the range restrictions were implemented symmetrically from the tails, thus all samples had similar mean scores (performance levels) but different variances; however, because the overall test score distributions were negatively skewed, the variability in scores was not held constant as performance level varied. To address the group heterogeneity question, they calculated alpha for each sample pool (e.g., middle 90%) and compared this reliability coefficient to that of the smaller samples used. Their results suggest that "when performance levels are comparable, the assumption of invariant measurement error is empirically tenable" and "measurement reliability largely depends on group variability, as the classical reliability theory predicts" (p. 364).

With respect to group performance, they followed a similar procedure but, due to the skewed distribution, group performance and group variability were confounded so they could not effectively assess the effect of performance on reliability because "the difference in reliability estimates between the high- and low-performance groups may be due to group variability differences, or to performance differences, or to the combination of the two" (p. 365). To help correct for the differences in standard deviations, they adjusted the group variability to conduct their analyses, this resulted in a reported finding that the low-performance group tended to have

more measurement error. Based on results, they concluded that measurement reliability will be reduced by group variability restriction and that it will also be affected by group performance. The authors mentioned two limitations to the study. First, the use of knowledge-based data (correct vs. incorrect scores) may have provided different results compared to typical-performance measurement data (attitude or interest scores); second, the lack of experimental control with respect to controlling the effects of different score characteristics. Interestingly, they suggested that a Monte Carlo study be conducted using simulated data with defined characteristics to prevent confounding of variables.

RG Counter Arguments

Sawilowsky (2000) was the first to issue a retort providing many substantial arguments against the RG method and refuting the proposition that reliability inures to test scores. He framed the context of the issue by explaining that it is not appropriate to make definitive statements, such as, "the reliability of test X is .90" (p. 158), because there are several ways to estimate reliability and the various methods account for measurement error in different ways. He argued that "statements about the reliability of a certain test must be accompanied by an explanation of what type of reliability was estimated, how it was calculated, and under what conditions or for which sample characteristics the result was obtained" (Sawilowsky, 2000, p. 159). Researchers must provide context when reporting reliability coefficients; a given test may have several different reliability estimates depending on how much the above mentioned variables (e.g., sample composition, physical conditions, reliability type, etc.) differ among test administrations. Sawilowsky (2000) described the manner in which reliability was presented in 63 different educational and psychological measurement and/or evaluation textbooks between the years 1986-2000. He showed, by quoting a variety of sources, that reliability is considered to

be a property of a test or measuring instrument itself among researchers, but that it is agreed its estimation will vary – it is not an unchanging constant (p.110).

In addition, Sawilowsky (2000) reiterated the long known reality that test scores – in fact, measurements in general – are not static; they fluctuate based on a number of different factors which affect their consistency. He explained:

> Because even the person taking a test may affect its reliability, social and behavioral science scales are not entirely analogous to the weight scale from the physical sciences. Factors such as temperature, air pressure, humidity, spring tension, battery condition and the levelness of the floor may affect the weight scale, but are readily discernable and compensated. There are, however, an infinite number of disconcerting factors that may arise in measuring straightforward variables such as achievement and how much more so for complex educational and psychological constructs such as aptitude. Even those that are known cannot be easily compensated for, potentially mitigating the consistency evidence. (pp. 197-198)

These issues have long been known and studied as previously evidenced by the various impacts to reliability which Symonds (1928) identified more than 80 years ago.

Rejoinder to the RG counter arguments.

In a retort to Sawilowsky (2000), Thompson and Vacha-Haase (2000) declared that they would not address the issues brought forth by Sawilowsky (2000), they stated:

> In this response to Sawilowsky's comments our primary focus will not be defending ourselves or our editorial policies and our analytic proposals from all criticisms or critics. Instead as in our original work, here our clear and present purpose remains moving the field toward more reflective practices as regards measurement.
>
> Therefore, we focus on those criticisms of Professor Sawilowsky that we believe are immediately relevant to our determined (albeit ambitious) focus. We especially hope that our major emphases will not be lost within a litany of minor concerns; we avoid this possibility by addressing here only the issues most relevant to our objectives, and demur from commenting otherwise. (Thompson & Vacha-Haase, 2000, p. 124)

As such, the article redirected the context of the argument away from the many problems with RG. It instead focused on the semantics of the word reliability and the interpretation of reliability coefficients as they relate to scores, not tests. Regrettably, Thompson and Vacha-Haase failed to present any substantive arguments or scientific evidence to support either the use of the RG method or the assertion that reliability inures to scores. They also neglected to address specific issues with respect to sampling error, measurement error, differences in reliability estimations based on reliability type calculated, testing conditions, reporting practices that cause reliability coefficients to vary between different examinee samples and/or testing administrations or any of the many other reasons provided to explain how and why reliability coefficients for the same instrument will vary across test administrations.

Study Rationale

This study will illustrate how the same instrument, administered under varying conditions, can result in different reliability coefficients due to sampling error, thus obviating the concept of RG. Showing that reliability is a psychometric property of a test is critical; since Vacha-Haase (1998) and Thompson and Vacha-Haase (2000) published their articles, the use and reporting of reliability has been confounded for researchers in many fields. The need to address the issue was evidenced when Thomas and Truax (2008) wrote a book chapter in the text *Handbook of Research Methods in Abnormal and Clinical Psychology* entitled *Assessment and Analysis of Clinically Significant Change*. In the chapter they stated:

> There is some controversy over the terms "score reliability" and "test reliability" with one faction insisting that the concept of reliability only applies to scores (Thompson and Vacha-Haase, 2003) and another holding that reliability is a property of a test under particular conditions (Sawilowsky, 2003a, 2003b). Since both factions wish to control word usage, it is not possible to employ the term or the usage in a way that satisfies both. We have not yet decided if this is an important distinction or just sophistry and will probably violate the standards of

> both factions in this brief exposition. In general, we lean toward the Sawilowsky position. (pp. 232-233)

Because test or instrument reliability is fundamental to research in behavioral, educational, clinical and many other scientific disciplines, it is critically important to clarify this question and to support its answer with sound, scientific evidence. The following paragraphs describe the background and components on which the methodology was based and designed.

Methodological Components

Various potential methodological problems with proposed RG studies, the confusion surrounding the reliability of tests and the notion that RG studies are concerned with measurement error, not test reliability have been highlighted. In order to study the reliability of a test, to illustrate that it is a stable metric of the instrument itself (not the test scores generated by examinees), and to show how reliability indices will vary, but only within the expected limits of standard errors of measurement, Monte Carlo methods will be used with the domain-sampling model as described by Nunnally (1978) to study the test-retest reliability of a matrix of correlated tests.

Domain sampling model.

In order to design and conduct a study regarding the reliability of a test, a sample of test items is needed. Nunnally (1978) provided a description of such a model: "the most useful model for the discussion of measurement error is that which considers any particular measure as being composed of *a random sample of items from a hypothetical domain of items*" (p. 193, emphasis in original). What Nunnally described is a test composed of randomly sampled items from an infinite number of potential items that could be chosen, this is called the domain. The more homogeneous the correlations in the domain, the more precise estimates of correlations will be with true scores. The domain-sampling model was designed to estimate the measurement that

would be obtained if all the items in a domain were tested by using samples of items from a large representative group of items. The reliability coefficient for the entire group of items is calculated, resulting in an estimate of reliability for an entire "domain" of items. This theoretical framework for sampling test items has been employed by psychometricians throughout the years. In 1951, Lee Cronbach provided the following description of a domain of items:

> There is no practical testing problem where the items in the test and only these items constitute the trait under examination. We may be unable to compose more items because of our limited skill as testmakers but any group of items in a test of intelligence or knowledge or emotionality is regarded as a sample of items. If there weren't "plenty more where these came from," performance on the test would not represent performance on any more significant variable. (p. 308)

Lee, Brennan and Frisbie clarified: "In the strict statistical sense, this means that these items can be viewed as a simple random sample from an indefinitely large domain of items" (2000, p. 12). Thus, any particular set of items used to create a test represents samples of items from an infinite hypothetical domain; these item samples can be used to study the psychometric properties of a test. As Nunnally explained, "the domain-sampling model can be developed without consideration of the number of items sampled for particular measures. Each sample could contain many items, or at the lower extreme, only one item. Also, the model can be developed without concern for the type of item employed or the factorial composition of items" (Nunnally, 1978, p. 194).

The domain sampling model relates to CTT by considering that reliability is concerned with stability and repeatability of measurement; the theoretical basis underlying the domain-sampling approach is to estimate the test score that an examinee would obtain if all the items in the domain were incorporated into a test. The score that a test-taker would obtain over the entire domain is their true score ($X = T + \varepsilon + c_1 + c_2 + \ldots c_n$, where X = true score, ε = random error and c_x = systematic errors, such as test-wiseness). The consistency of measurement between

random samples of items with other such samples from the same domain constitutes a measure of test reliability (Nunnally, 1978; Cronbach, 2004). Nunnally (1978) described the logic behind the model by conceptualizing an infinitely large correlation matrix showing all correlations among items in a particular domain:

> The average correlation in the matrix r_{kk} would indicate the extent to which some common core existed in the items. The dispersion of correlations about the average would indicate the extent to which items varied in sharing the common core. If the assumption is made that all items have an equal amount of the common core, the average correlation in each column of the hypothetical matrix would be the same, which would be the same as the average correlation in the whole matrix. (p. 195)

Using the above assumption, Nunnally showed how it is possible to directly compute – not simply estimate – the correlation of any particular item, or group of items (i.e., in this study a testlet), with the sum of all items in the domain. Thus, the domain-sampling model considers any particular measure as being composed of a random sample of items from a hypothetical domain and – to the extent that correlations among items vary in the domain – there is random error associated with the average correlation in any particular sampling of items. Precision of the estimates is measured in terms of the standard error of item correlations which, in turn, is a function of the variance of the item correlations.

Domain-sampling model assumptions.

The following assumptions underlie the use of the domain-sampling model as described by Nunnally for investigating the reliability of tests:

- Observed scores differ from true scores on a random basis;
- The wider the spread of observed scores about true scores, the more error there is in employing the instrument;

- One standard deviation of errors typifies the amount of error to be expected in calculated reliability coefficients;
- Typical standard deviation of errors equates to the standard error of measurement (σ_{meas});
- The size of σ_{meas} is a direct indication of the amount of error involved in using a particular type of instrument;
- The variance of the sum of *k* sets of standard scores equals the sum of all the elements in the correlation matrix for those scores;
- All items in the domain are unidimensional, that is, they measure the same underlying construct or attribute;
- Correlations among items in the domain are normally distributed and statistically independent of one another; and
- The average correlation of each item with the all others is the same and is also the same as the average correlation in the matrix (Nunnally, 1978, pp. 193-196).

Nunnally (1978) noted that the domain-sampling model assumes correlations are normally distributed about the average value and are statistically independent of one another, but he also stated, "both assumptions are known to be slightly incorrect" (p. 207). Nunnally explained that, if an average correlation is positive, then the random distribution of correlations around the average tends to be negatively skewed and they are not completely independent of each other. However, he went on to say that these assumptions are violated only slightly, consequently the domain-sampling model will hold well in practice. This issue will not impact this research because the data will be computer-generated.

Testlets.

The concept of a testlet was introduced in 1987 by Wainer and Kelly as: "a group of items related to a single content area that is developed as a unit" (Wainer & Lewis, 1990, p. 1). Testlets have also been described as small tests; or as tests small enough to manipulate, but large enough to carry their own context as analysis units (Wainer & Kiely, 1987; Wainer & Lewis, 1990). As defined by Lee, Brennan and Frisbie, a testlet is "a subset of the items in a test form that is treated as a measurement unit in test construction, administration, and/or scoring" (Lee, 2002, p. 149).

Test reliability: minimum levels.

As previously noted, the correlation is the statistic used to assess reliability. Whereas the interpretation of this statistic is clear, the required magnitude for reliability coefficients is a source of debate. Over the years, a number of researchers have proposed minimum levels of reliability depending on both the context of the measurement instrument and its purpose (e.g., clinical, diagnostic, screening). As Cortina (1993) pointed out:

> The level of reliability that is adequate depends on the decision that is made with the scale. The finer the distinction that needs to be made, the better the reliability must be. For example, the reliability of the Scholastic Aptitude Test is quite adequate for distinguishing between a 750 scorer and a 450 scorer. Its reliability is not adequate for distinctions between scores of 749 and 750. Thus, any judgment of adequacy, even in research, needs to consider context. (p. 101)

In 1927 Kelly asserted that a reliability coefficient of 0.94 was needed to evaluate levels of individual accomplishment. More recently, Gregory (1999), Guilford and Fruchter (1978), Hopkins, Stanley, and Hopkins (1990), and Salvia and Ysseldyke (1988) suggested a reliability of 0.90 for accurate measurement in practical use. In addition to following the recommendation of Nunnally and Bernstein, who in 1994 advised that – for most research purposes – a reliability of 0.80 is adequate (p. 265), this study will also examine reliabilities of 0.70 and 0.90.

Monte Carlo methods.

Data simulation involves using computer models to emulate real life or to make predictions. Monte Carlo simulation is a computerized mathematical technique used to evaluate complex models by employing sets of random numbers as inputs; the simulation results in a range of possible outcomes and the probabilities of their occurrence. The technique was first used by scientists working on the atom bomb during World War II; it was named for Monte Carlo, the Monaco resort town renowned for its casinos. Since its introduction, Monte Carlo simulation has been used to model a wide variety of physical and conceptual systems (Metropolis, 1987).

As Harwell (1992) described, Monte Carlo studies have many beneficial attributes compared to other types of research:

> MC studies do not appear to suffer from the range or magnitude of definitional difficulties that often plague meta-analyses of empirical studies in education and psychology-for example, heterogeneity of measured constructs, studies of widely varying methodological quality, study selection bias, and dependencies among EMs [effect magnitudes]. This is due to the small number of simulation factors usually employed in MC studies, their nature, and the control that is exercised over the data generation process. These factors bolster the credibility of a carefully conducted summary of MC results. (p. 302)

In addition, as Headrick and Sawilowsky (1999) noted, "Monte Carlo simulations requiring correlated data from normal and nonnormal populations are frequently used to investigate the small sample properties of competing statistics, or the comparison of estimation techniques" (p. 25). Monte Carlo simulation methods were selected for this research because they can provide insight into what would happen to reliability if the same test, measuring the same underlying factor, were to be administered many times to different random and non-random samples. As Sawilowsky and Fahoome (2003) noted "Monte Carlo refers to repeated sampling from a probability distribution to determine the long run average of some parameter or characteristic"

(p. 46), in this case, the long run average of reliability coefficients for various sized testlets sampled randomly and non-randomly. In addition, Sawilowsky (2003) listed the characteristics of a high quality Monte Carlo simulation:

- the pseudo-random number generator has certain characteristics (e. g., a long "period" before repeating values);
- the pseudo-random number generator produces values that pass tests for randomness;
- the number of repetitions of the experiment is sufficiently large to ensure accuracy of results;
- the proper sampling technique is used;
- the algorithm used is valid for what is being modeled; and
- the study simulates the phenomenon in question.

CHAPTER 3

METHODOLOGY

Vacha-Haase (1998) noticed differences among reliability coefficient values calculated for different administrations of the same test. The obvious, but overlooked, explanation for this phenomenon – beyond the already known fact that different types of reliability will yield slightly different values – could be summarized as follows: Sampling error occurs even under pristine randomized procedures and it increases in practical administrations when implementing a test, particularly if standard process protocol is violated.

Rather than examining sampling error, Vacha-Haase invoked a concept similar to validity generalization. As discussed earlier, validity generalization is possible because the same instrument can be used for a variety of different purposes, or conversely, several different instruments can be valid for a particular purpose (for example, *The Mental Measurement Yearbook* for any given year lists several different measures with high validity that can be used to diagnose anxiety, depression, etc.). In any case, disparity across administrations related to validity evidence is not due to sampling error; hence, validity evidence is amenable to the meta-analytic procedure of synthesizing different literature.

Ignoring classical definitions of reliability and validity is the source of the mistake with reliability generalization. Classically, reliability is a property of a test; test scores are known to fluctuate due to sampling error or deviations from proper test administration protocol. Validity relates to the usage of a test for a particular purpose: it is a proposition, inference or conclusion that is valid or invalid. The scores (or measurements) generated by a test do not themselves provide information regarding the type or strength of the validity associated with the use of the

instrument. Test scores (or measurements) do not *have* validity; it is only appropriate to state that a score or measurement *leads to* valid conclusions or *enables* valid inferences to be made.

Vacha-Haase's concerns with variation in reliability estimates, therefore, are addressed with theoretical data by computing the reliabilities of various sized testlets sampled randomly and non-randomly from a domain of data correlated to a known value and comparing the average reliabilities of the various sized testlets to the known reliability of the universe. This study was designed to illustrate how the same instrument, administered repeatedly, can result in different reliability coefficients and to show that variation in reliability coefficients is due to sampling error. The following questions encompass the study's rationale:

Statistical Question 1: Can the r_{xy}, where (x, y) = 5, 10, 15, 25, 50, 75, 87 and 100, equal r_{XY}, where (X, Y) = 1,000,000, within the sampling error of SEM as predicted by classical measurement theory, eliminating the notion of reliability generalization?

Statistical Question 2: Does the, r_{xy} where (x, y) = 5, 10, 15, 25, 50, 75, 87 and 100, equal r_{XY}, where (X, Y) = 1,000,000, when administration protocols (i.e., lack of randomization) are violated as predicted by classical measurement theory, eliminating the notion of reliability generalization?

Data Creation

It is assumed that a matrix of pseudo-random numbers generated from a known data distribution represents a random sample of items from a hypothetical infinite domain of items. It is also assumed that pseudo random number generators are able to populate such a domain. Based on statistical tests of randomness (e.g., Chi-squared, Kolmogorov-Smirnov) and sufficiently long periods (cycles), randomly selecting pseudo-random numbers will result in representative samples of the larger domain, or universe; these will be called *mini-universes*.

Because most social and behavioral science instruments are composed of a yet smaller number of items taken together, called a scale or test, multiple items from the mini-universes will be used to calculate reliability per Nunnally (1978) and Lee (2002). They described how an infinitely large matrix of correlated items can be divided into groups composed of *h* items and how the sum of scores for each group of items may be considered as constituting a test. These are called testlets (Lee, 2002), and are the measurement units for analysis. The purpose of using correlated data and testlets is to estimate a subject's True Score (T in X = T + E), the score that the individual would obtain if they were to retake the same test an infinite number of times.

If items within the mini-universes are randomly sampled to create the testlets, then the testlets created will be randomly parallel. Means, standard deviations and estimates of correlations with true scores among the randomly parallel testlets will differ only by chance. The expectation is that – for random samples of parallel testlets correlated to a known value, 0.70, 0.80 or 0.90 – any variation in the calculated reliability coefficient between a testlet and its entire mini-universe will be due to sampling error. Thus, the metric of interest is the size of the correlation, as Morrow and Jackson (1993) stated: "The magnitude of the reliability coefficient is the issue, not statistical significance" (p. 354). For this reason, significance testing of the calculated coefficients will not be necessary; instead the difference between each testlet's average reliability and the overall average reliability of the testlet samples will be used to determine the extent to which differs from that of its mini-universe. Also, the standard deviation and SEM will serve to gauge how much, on average, the testlet reliabilities differ from the known reliability of the mini-universe.

Study Protocol

A program was written in Essential Lahey Fortran 90 (ELF90) to study the reliability of tests.

Part 1: Simulated data with random selection.

Algorithms for creating correlated data based on the Fleishman (1978) procedure as described by Headrick and Sawilowsky (2000) and presented by Sawilowsky and Fahoome (2003) were used to populate 2 × 1,000,000 matrices of correlated data for five different distributions: normal, Chi-squared (df = 1), Exponential, Double Exponential and t (df = 3). This procedure creates *X* and *Y* variables set to a specified correlation from the different distributions and at the same time preserves the underlying distribution shapes (skew and kurtosis).

First, using a Texas Instruments TI-83 graphing calculator, the constants *a*, *b* and *d* from Table 2 for each distribution were employed to solve for *r* in the equation:

$$r_{xy} = r^2(b^2 + 6bd + 9d^2 + 2a^2r^2 + 6d^2r^4), \tag{1.1}$$

where r_{xy} = 0.70, 0.80 or 0.90. After *r* was calculated from (1.1), it was used to produce standard normal variates, which are intermediate values. The intermediate values (x_i and y_i) were obtained using FORTRAN subroutines Rangen 2.0 and normb1.f90 (Sawilowsky & Fahoome, 2003) to randomly select three standard normal *z*-scores (z_1, z_2 and z_3) and placing these values into the following equations:

$$x_i = rz_1 + (\sqrt{1-r^2})(z_2) \tag{1.2}$$

$$y_i = rz_1 + (\sqrt{1-r^2})(z_3) \tag{1.3}$$

The intermediate x_i and y_i values generated from (1.2) and (1.3) were then applied to produce the i^{th} scores that were substituted in the Fleishman equations

$$X_i = a + bx_i + (-a)\,x_i^2 + dx_i^2 \tag{1.4}$$

$$Y_i = a + by_i + (-a)\,y_i^2 + dy_i^2 \tag{1.5}$$

to create the correlated data pairs (X_i, Y_i). The intermediate r values calculated in equation (1.1) ensure that the correlated scores generated in equations (1.2) – (1.5) maintain the properties of the distribution from which they were sampled. Thus, in addition to knowing the correlation of each data pair, the Fleishman procedure retains the shape of the underlying distribution, therefore controlling both skew (γ_1) and kurtosis (γ_2). The algorithm also produces data distributed with $\mu = 0$ and $\sigma = 1$; as a result, computed correlated values are standardized to this mean and standard deviation. Using the Fleishman method to generate data with a known correlation, r_{XY}, sets the extent to which test items (X_i, Y_i) are unidimensional, that is, the extent to which they measure the same underlying concept or construct. In addition, this r_{XY} – which is the reliability of the mini-universe – is the reliability coefficient of a norm group to which testlet (samples) reliabilities were compared.

Table 2: Study Distribution Statistical Properties and Solutions to the Fleishman Equation: Intermediate r Values by Distribution and Correlation

Distribution	r_{xy}	γ_1* (Skew)	γ_2* (Kurtosis)	a*	b*	d*	Intermediate r Value
Normal	.70	0	0	0	1	0	.83670
	.80			0	1	0	.89443
	.90			0	1	0	.94868
Chi-square (df = 1)	.70	2.828	12	−.5207	.6146	.02007	.88909
	.80			−.5207	.6146	.02007	.92960
	.90			−.5207	.6146	.02007	.96633
Exponential	.70	2	6	−.3137	.8263	.02271	.85998
	.80			−.3137	.8263	.02271	.91319
	.90			−.3137	.8263	.02271	.95973
Double Exponential	.70	0	3	0	.7824	.0679	.84248
	.80			0	.7824	.0679	.89877
	.90			0	.7824	.0679	.95110
t (df = 3)	.70	0	17	0	.3938	.1713	.86665
	.80			0	.3938	.1713	.91814
	.90			0	.3938	.1713	.96118

*Source: Sawilowsky & Fahoome (2003) from Headrick & Sawilowsky (2000, p. 427)

As equations (1.2) – (1.5) were solved and data was generated, the correlated data was placed into a 2 × 1,000,000 matrix, which is called a *mini-universe* (*mini* represents the fact that many more values could be calculated, but for practical purposes defined data sets were used). Thus, all data pairs in the matrix were calculated such that each X_i correlates to a Y_i from a particular distribution at a specified level of correlation. This process was performed fifteen times so that each of the five distributions had three 2 × 1,000,000 matrices of data correlated at 0.70, 0.80 and 0.90. Thus, each of the fifteen mini-universes holds 1 million pairs of correlated data simulating 2 million test scores where the data pairs (X_i, Y_i) represent paired test items, with the X_i's representing results from a first test administration and Y_i's representing results from a second administration; this mirrors a test-retest administration procedure.

After the mini-universes were generated, pairs of correlated data (X_i, Y_i) were randomly sampled from the 2 × 1,000,000 mini-universes into 2 x n testlets, where n = 5, 10, 15, 25, 50, 75, 87 and 100. Because a random number generator was used to simulate the (X_i, Y_i) pairs, the data was randomized within the mini-universe as it was produced by the program. This allowed testlets to be sampled randomly by simply taking sequential sets for all specified testlet sizes (2 x n), until each entire 2 × 1,000,000 million matrix was exhausted of (X_i, Y_i) pairs. The (X_i, Y_i) pairs represent parallel testlets of varying sample sizes (e.g., 2 x 5, 2 x 10, etc.). For example, using testlets sized n = 5 and the mini-universe for the normal distribution correlated at r_{xy} = 0.70, 200,000 testlets were randomly sampled. (See Table 3.) The mean, standard deviation, Pearson's Coefficient of Correlation r_{xy} (i.e., the reliability coefficient) and the standard error of measurement were calculated for all testlets. The testlet correlations represent reliability coefficients as calculated via test-retest method for a test administered multiple times to different sized groups of test subjects.

After the correlation coefficients were calculated and stored in the database, the upper bound, lower bound, mean, standard deviation and standard error of measurement was recorded for the correlations according to the formulas:

Upper Bound Correlation	MAX = highest correlation value	(2)
Lower Bound Correlation	MIN = lowest correlation value	(3)
Mean Reliability Coefficient	$\bar{x}_{r_{xy}} = \frac{\sum r_{xy}}{n}$	(4)
Standard Deviation of the Correlations	$s = \sqrt{\frac{\sum r_{xy}{}^2 - \left(\sum r_{xy}\right)^2}{n-1}}$	(5)
Standard Error of Measurement of the Reliability Coefficient	$SEM = s\sqrt{\left(1-\bar{x}_{r_{xy}}\right)}$	(7)

where X = correlation coefficients and r_{xy} is the average correlation coefficient for all 1 million testlets.

The precision with which reliability is estimated for a testlet is a direct function of the accuracy with which the correlation of items in a test estimates the correlation of all items in the domain (Nunnally, 1978, p. 210). It is expected that correlations calculated for each testlet will have very small standard deviations and SEMs and will nearly approximate the known correlation of the entire mini-universe.

Part 2: Simulated data with non-random selection.

The procedures in Part 1, as described above, were replicated with one change: randomization was violated to simulate practical conditions of administering a test to smaller non-randomly sampled groups of examinees. To create the non-random samples, the correlations in the mini-universes were ordered from low to high by adding the absolute values of the (X_i, Y_i) pairs and sorting the sums in ascending order from X_1Y_1 to $X_{1,000,000}Y_{1,000,000}$. After the data was ordered, multiple 2 x n testlets were again sampled, however, they were not sampled randomly – they were first sampled only from within the lowest 25% of the values (bottom 250,000) and then only from within the highest 25% of the values (top 250,000). (See Tables 4 and 5.) The expectation being that reliability coefficients observed for testlets would differ to a much greater extent from the known reliability of the mini-universe due to sampling error introduced by the use of non-random samples.

Table 3: Monte Carlo Simulation Variations, Random

Distribution	Correlation	Testlet Size (n)	# Parallel Testlets
Randomized Simulation, 0.70: 25 total			
Normal Chi-squared (df=1) Exponential Double Exponential t (df=3)	r_{xy}=0.70	5	200,000
		10	100,000
		15	66,667
		25	40,000
		50	20,000
		75	13,333
		87	11,494
		100	10,000
Randomized Simulation 0.80: 25 total			
Normal Chi-squared (df=1) Exponential Double Exponential t (df=3)	r_{xy}=0.80	5	200,000
		10	100,000
		15	66,667
		25	40,000
		50	20,000
		75	13,333
		87	11,494
		100	10,000
Randomized Simulation 0.90: 25 total			
Normal Chi-squared (df=1) Exponential Double Exponential t (df=3)	r_{xy}=0.90	5	200,000
		10	100,000
		15	66,667
		25	40,000
		50	20,000
		75	13,333
		87	11,494
		100	10,000

Table 4: Monte Carlo Simulation Variations, Non-Random Low*

Distribution	Correlation	Testlet Size (n)	# Parallel Testlets
Non-Randomized Simulations – Low: 25 total			
Normal Chi-squared (df=1) Exponential Double Exponential t (df=3)	r_{xy}=0.70	5	50,000
		10	25,000
		15	16,667
		25	10,000
		50	5,000
		75	3,333
		87	2,873
		100	2,500
Non-Randomized Simulations – Low: 25 total			
Normal Chi-squared (df=1) Exponential Double Exponential t (df=3)	r_{xy}=0.80	5	50,000
		10	25,000
		15	16,667
		25	10,000
		50	5,000
		75	3,333
		87	2,873
		100	2,500
Non-Randomized Simulations – Low: 25 total			
Normal Chi-squared (df=1) Exponential Double Exponential t (df=3)	r_{xy}=0.90	5	50,000
		10	25,000
		15	16,667
		25	10,000
		50	5,000
		75	3,333
		87	2,873
		100	2,500

*Correlates sorted in ascending order based on sum of absolute value of (X_i, Y_i) pairs; only the lowest 250,000 data pairs selected for non-random simulation

Table 5: Monte Carlo Simulation Variations, Non-Random High*

Distribution	Correlation	Testlet Size (n)	# Parallel Testlets
Non-Randomized Simulations – High: 25 total			
Normal Chi-squared (df=1) Exponential Double Exponential t (df=3)	r_{xy}=0.70	5	50,000
		10	25,000
		15	16,667
		25	10,000
		50	5,000
		75	3,333
		87	2,873
		100	2,500
Non-Randomized Simulations – High: 25 total			
Normal Chi-squared (df=1) Exponential Double Exponential t (df=3)	r_{xy}=0.80	5	50,000
		10	25,000
		15	16,667
		25	10,000
		50	5,000
		75	3,333
		87	2,873
		100	2,500
Non-Randomized Simulations – High: 25 total			
Normal Chi-squared (df=1) Exponential Double Exponential t (df=3)	r_{xy}=0.90	5	50,000
		10	25,000
		15	16,667
		25	10,000
		50	5,000
		75	3,333
		87	2,873
		100	2,500

*Correlates sorted in ascending order based on sum of absolute value of (X_i, Y_i) pairs; only the highest 250,000 data pairs selected for non-random simulation

CHAPTER 4

RESULTS

Monte Carlo simulations were used to create fifteen 2 × 1,000,000 mini-universes of data pairs correlated to a set level (0.70, 0.80 or 0.90) for five different distributions: normal, Chi-squared (df=1), exponential, double exponential and t (df=3). Table 6 shows the actual reliabilities for the mini-universes (r_{XY}); these are the values to which sample testlet reliabilities were compared. Differences between the set and actual correlations are due to random sampling error from selecting the three z-scores used in the Fleishman equations that create the correlates; as Thye (2000) noted "random measurement errors affect each observation randomly, causing a degree of unreliability" (p. 1279). The correlations of the mini-universes vary little from the set value in all cases.

Table 6: Mini-Universe Reliabilities (r_{XY})*

Distribution	Mini-Universe Reliability (r_{XY})		
	0.70	0.80	0.90
Normal	0.700330	0.800238	0.900144
Chi-Squared (df=1)	0.688895	0.792115	0.895831
Exponential	0.690861	0.798917	0.902728
Double Exponential	0.706736	0.805540	0.903367
t (df=3)	0.698982	0.806113	0.904287

*Actual correlations of 2 × 1,000,000 pairs of correlated data as generated by Fleishman equations

Tables 7-21 show the summary statistics resulting from randomly and non-randomly sampling various sized testlets from the mini-universes. For each size testlet studied (5, 10, 15, 25, 50, 75, 87, 100), the tables present the average correlation – i.e., average reliability

coefficient ($\bar{x}_{r_{xy}}$) – and the average standard deviations and standard errors of measurement. In addition, the upper and lower bound of the correlations from the entire group of testlets are also shown. Note that when working with correlations, a value of zero indicates a complete lack of relationship between variables studied whereas correlations near −1.00 or +1.00 indicate very strong relationships; for this reason the terms upper and lower bound are used to describe the range of values among observed correlations. Tables 22-26 show the differences between average reliabilities for each size testlet ($\bar{x}_{r_{xy}}$) and the reliability of the mini-universes (r_{XY}). Finally, Tables 27-29 summarize the results for testlets size n = 87. This data is tabled separately in order to highlight the statistics related to the same size group of tests that Vacha-Haase (1998) used in the reliability generalization study with the Bem Sex Role Inventory.

Table 7: Correlation Simulation Results for Normal Distribution, r_{XY} = 0.70

Testlet Size (n)	$\overline{x}_{r_{xy}}$ (Reliability)	Standard Deviation	SEM	Upper Bound	Lower Bound
RANDOM: All 1,000,000 (X, Y) Score Pairs					
5	0.648266	0.336244	0.199417	0.999898	-0.998839
10	0.678807	0.195213	0.110635	0.990772	-0.784411
15	0.686831	0.149353	0.083580	0.972348	-0.444311
25	0.692453	0.110018	0.061013	0.942147	-0.012465
50	0.696548	0.074702	0.041151	0.891257	0.321947
75	0.697971	0.059938	0.032940	0.881489	0.399607
87	0.698261	0.055614	0.030549	0.869163	0.438630
100	0.698476	0.051954	0.028529	0.859660	0.461481
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs					
5	0.046198	0.434856	0.424692	1.000000	-1.000000
10	0.047852	0.232120	0.226498	0.990044	-0.968003
15	0.047259	0.173525	0.169375	0.719866	-0.676390
25	0.047078	0.126793	0.123772	0.490121	-0.464661
50	0.046547	0.088548	0.086463	0.394986	-0.238392
75	0.046779	0.073557	0.071816	0.298490	-0.189026
87	0.046531	0.069188	0.067559	0.274246	-0.198798
100	0.046549	0.064468	0.062949	0.269850	-0.169655
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs					
5	0.751809	0.473096	0.235690	0.999987	-1.000000
10	0.839778	0.152595	0.061080	0.997202	-1.000000
15	0.840701	0.110665	0.044169	0.993445	-1.000000
25	0.838213	0.097925	0.039388	0.988383	0.377802
50	0.835971	0.089252	0.036148	0.983180	0.485573
75	0.835125	0.086444	0.035100	0.983279	0.568150
87	0.834948	0.085649	0.034796	0.981230	0.559541
100	0.834642	0.085133	0.034619	0.979346	0.553004

Table 8: Correlation Simulation Results for Normal Distribution, r_{XY} = 0.80

Testlet Size (n)	$\overline{x}_{r_{xy}}$ (Reliability)	Standard Deviation	SEM	Upper Bound	Lower Bound
RANDOM: All 1,000,000 (X, Y) Score Pairs					
5	0.754190	0.269254	0.133494	0.999957	-0.995279
10	0.782099	0.145694	0.068010	0.994145	-0.638367
15	0.789046	0.108939	0.050036	0.983339	-0.259832
25	0.793804	0.079072	0.035906	0.964311	0.183494
50	0.797165	0.053190	0.023955	0.933423	0.504464
75	0.798331	0.042537	0.019102	0.924510	0.574481
87	0.798549	0.039494	0.017726	0.916329	0.606996
100	0.798740	0.036850	0.016532	0.909426	0.620531
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs					
5	0.065593	0.436693	0.422129	1.000000	-1.000000
10	0.067836	0.233976	0.225901	0.921188	-0.999881
15	0.067597	0.175220	0.169194	0.792544	-0.772006
25	0.066721	0.130811	0.126372	0.612490	-0.454470
50	0.066165	0.094919	0.091725	0.479109	-0.260983
75	0.066045	0.081040	0.078318	0.367343	-0.187623
87	0.065567	0.077421	0.074839	0.371189	-0.202954
100	0.065906	0.073736	0.071265	0.323894	-0.216848
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs					
5	0.807223	0.475725	0.208874	0.999970	-1.000000
10	0.909173	0.111247	0.033527	0.997343	-1.000000
15	0.911257	0.056840	0.016933	0.995821	-1.000000
25	0.909363	0.048480	0.014596	0.994195	0.631623
50	0.907833	0.043983	0.013353	0.989849	0.743409
75	0.907264	0.042372	0.012903	0.989214	0.713123
87	0.907041	0.041914	0.012779	0.987120	0.746294
100	0.906970	0.041596	0.012687	0.987637	0.759218

Table 9: Correlation Simulation Results for Normal Distribution, r_{XY} = 0.90

Testlet Size (n)	$\overline{x}_{r_{xy}}$ (Reliability)	Standard Deviation	SEM	Upper Bound	Lower Bound
RANDOM: All 1,000,000 (X, Y) Score Pairs					
5	0.868924	0.173854	0.062943	0.999968	-0.985296
10	0.888731	0.083147	0.027735	0.997155	-0.363735
15	0.893277	0.059949	0.019584	0.992391	0.096182
25	0.896271	0.042632	0.013731	0.983286	0.500806
50	0.898309	0.028339	0.009037	0.969543	0.729307
75	0.899011	0.022577	0.007175	0.963705	0.782618
87	0.899131	0.020983	0.006664	0.959613	0.791521
100	0.899254	0.019550	0.006205	0.955850	0.798688
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs					
5	0.108589	0.443863	0.419071	1.000000	-1.000000
10	0.114835	0.242690	0.228331	0.962486	-0.961953
15	0.114506	0.188188	0.177086	0.834822	-0.803851
25	0.113364	0.145043	0.136574	0.687799	-0.427953
50	0.112307	0.112915	0.106386	0.512017	-0.246519
75	0.111620	0.101592	0.095755	0.424637	-0.179846
87	0.111789	0.098187	0.092536	0.413281	-0.145094
100	0.111785	0.095554	0.090055	0.398565	-0.166750
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs					
5	0.847437	0.471799	0.184282	0.999991	-1.000000
10	0.955611	0.090441	0.019055	0.999473	-1.000000
15	0.958201	0.024848	0.005080	0.998831	0.646015
25	0.957248	0.021838	0.004515	0.997064	0.841458
50	0.956484	0.019704	0.004110	0.994973	0.877120
75	0.956219	0.018958	0.003967	0.994119	0.874562
87	0.956154	0.018753	0.003927	0.993684	0.883418
100	0.956104	0.018603	0.003898	0.994167	0.882716

Table 10: Correlation Simulation Results for Chi-Squared (df=1) Distribution, r_{XY} = 0.69

Testlet Size (n)	$\overline{x}_{r_{xy}}$ (Reliability)	Standard Deviation	SEM	Upper Bound	Lower Bound
RANDOM: All 1,000,000 (X, Y) Score Pairs					
5	0.578244	0.424615	0.275756	0.999992	-0.996677
10	0.631987	0.279223	0.169388	0.999421	-0.736336
15	0.650152	0.223653	0.132286	0.997298	-0.465799
25	0.664332	0.171386	0.099296	0.984629	-0.277864
50	0.675827	0.121412	0.069127	0.958771	0.107786
75	0.680089	0.099654	0.056365	0.940450	0.165121
87	0.681311	0.092823	0.052401	0.928377	0.135081
100	0.682199	0.086662	0.048855	0.923595	0.216394
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs					
5	-0.134091	0.521090	0.554928	1.000000	-1.000000
10	-0.095193	0.311340	0.325822	0.991465	-1.000000
15	-0.081184	0.236128	0.245526	0.663107	-1.000000
25	-0.072298	0.171661	0.177758	0.461480	-0.915003
50	-0.066434	0.119993	0.123914	0.304612	-0.666995
75	-0.064527	0.099646	0.102811	0.228908	-0.553689
87	-0.064112	0.093552	0.096505	0.183379	-0.615839
100	-0.063754	0.089738	0.092555	0.161117	-0.548462
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs					
5	-0.732711	0.591735	0.778916	1.000000	-1.000000
10	-0.611498	0.679322	0.862363	1.000000	-1.000000
15	-0.541692	0.713028	0.885330	1.000000	-1.000000
25	-0.485729	0.721425	0.879349	1.000000	-1.000000
50	-0.473536	0.694947	0.843591	0.999983	-1.000000
75	-0.477071	0.682817	0.829860	0.984213	-0.999999
87	-0.479379	0.677985	0.824632	0.976572	-0.999999
100	-0.479383	0.676331	0.822621	0.899244	-0.999999

Table 11: Correlation Simulation Results for Chi-Squared (df=1) Distribution, r_{XY} = 0.81

Testlet Size (n)	$\overline{x}_{r_{xy}}$ (Reliability)	Standard Deviation	SEM	Upper Bound	Lower Bound
RANDOM: All 1,000,000 (X, Y) Score Pairs					
5	0.692427	0.359106	0.199158	0.999994	-0.998025
10	0.743291	0.221001	0.111973	0.999512	-0.678000
15	0.759299	0.172488	0.084625	0.998235	-0.337154
25	0.771740	0.128814	0.061543	0.989584	-0.125798
50	0.781376	0.089518	0.041856	0.973650	0.281638
75	0.784936	0.072826	0.033773	0.961050	0.361304
87	0.785849	0.067909	0.031426	0.954118	0.325845
100	0.786615	0.063183	0.029187	0.951861	0.403211
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs					
5	-0.062570	0.510123	0.525841	1.000000	-1.000000
10	-0.027156	0.293765	0.297727	1.000000	-1.000000
15	-0.015856	0.215929	0.217634	0.699971	-1.000000
25	-0.009824	0.154467	0.155223	0.538468	-0.865845
50	-0.005572	0.104092	0.104381	0.353917	-0.643846
75	-0.004180	0.084050	0.084226	0.341446	-0.318820
87	-0.003948	0.078240	0.078395	0.315674	-0.342294
100	-0.003551	0.072652	0.072781	0.214754	-0.373369
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs					
5	-0.777367	0.585524	0.780609	1.000000	-1.000000
10	-0.641824	0.709474	0.909075	1.000000	-1.000000
15	-0.557654	0.761325	0.950179	1.000000	-1.000000
25	-0.464938	0.799596	0.967788	1.000000	-1.000000
50	-0.406994	0.800810	0.949896	1.000000	-1.000000
75	-0.402788	0.788880	0.934344	1.000000	-1.000000
87	-0.404829	0.785364	0.930856	0.999616	-1.000000
100	-0.405026	0.782814	0.927899	0.999918	-0.999999

Table 12: Correlation Simulation Results for Chi-Squared (df=1) Distribution, r_{XY} = 0.90

Testlet Size (n)	$\overline{x}_{r_{xy}}$ (Reliability)	Standard Deviation	SEM	Upper Bound	Lower Bound
RANDOM: All 1,000,000 (X, Y) Score Pairs					
5	0.826034	0.251336	0.104830	0.999996	-0.980836
10	0.864316	0.136272	0.050196	0.999704	-0.466942
15	0.875128	0.101735	0.035950	0.998944	-0.202013
25	0.883306	0.073115	0.024976	0.995442	0.203398
50	0.889351	0.049431	0.016443	0.987085	0.565133
75	0.891534	0.039756	0.013093	0.981154	0.638914
87	0.892032	0.037115	0.012196	0.977841	0.611253
100	0.892523	0.034370	0.011268	0.976557	0.665759
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs					
5	0.045126	0.502563	0.491093	1.000000	-1.000000
10	0.080316	0.280146	0.268661	0.913167	-1.000000
15	0.086353	0.210598	0.201300	0.822164	-1.000000
25	0.090216	0.157517	0.150244	0.705517	-0.731018
50	0.091861	0.116973	0.111471	0.501320	-0.315190
75	0.092267	0.101971	0.097153	0.416850	-0.194245
87	0.092241	0.099649	0.094942	0.455258	-0.189479
100	0.092667	0.095883	0.091333	0.400145	-0.202196
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs					
5	-0.862054	0.501079	0.683758	1.000000	-1.000000
10	-0.753072	0.648658	0.858847	1.000000	-1.000000
15	-0.667958	0.731785	0.945097	1.000000	-1.000000
25	-0.551014	0.816234	1.016540	1.000000	-1.000000
50	-0.411054	0.881503	1.047120	1.000000	-1.000000
75	-0.366618	0.891810	1.042550	1.000000	-1.000000
87	-0.360526	0.891725	1.040120	0.999998	-1.000000
100	-0.353697	0.891156	1.036850	0.999995	-1.000000

Table 13: Correlation Simulation Results for Exponential Distribution, r_{XY} = 0.69

Testlet Size (n)	$\overline{x}_{r_{xy}}$ (Reliability)	Standard Deviation	SEM	Upper Bound	Lower Bound
RANDOM: All 1,000,000 (X, Y) Score Pairs					
5	0.634620	0.367176	0.221946	0.999991	-0.995957
10	0.665239	0.234885	0.135901	0.998409	-0.716823
15	0.674201	0.186898	0.106679	0.991818	-0.368423
25	0.680697	0.143386	0.081023	0.983465	-0.148920
50	0.685639	0.101079	0.056673	0.939688	0.237231
75	0.687359	0.083355	0.046607	0.908906	0.317897
87	0.687899	0.077317	0.043194	0.896931	0.284520
100	0.688268	0.072233	0.040330	0.895871	0.359762
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs					
5	-0.024895	0.481736	0.487696	1.000000	-1.000000
10	-0.002733	0.264632	0.264994	0.948046	-1.000000
15	0.002983	0.194359	0.194069	0.719872	-0.994021
25	0.006769	0.139992	0.139517	0.463452	-0.627443
50	0.008481	0.093655	0.093257	0.370334	-0.366490
75	0.009055	0.074944	0.074604	0.263032	-0.311203
87	0.009225	0.068417	0.068101	0.220837	-0.283320
100	0.009293	0.063923	0.063625	0.234001	-0.254494
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs					
5	-0.660248	0.696148	0.896991	1.000000	-1.000000
10	-0.503512	0.783875	0.961170	1.000000	-1.000000
15	-0.437448	0.799476	0.958521	1.000000	-1.000000
25	-0.385978	0.796799	0.938052	1.000000	-1.000000
50	-0.380432	0.775087	0.910664	0.999900	-1.000000
75	-0.382407	0.768214	0.903233	0.987718	-1.000000
87	-0.383959	0.765221	0.900220	0.951908	-1.000000
100	-0.384412	0.763333	0.898146	0.976636	-0.999999

Table 14: Correlation Simulation Results for Exponential Distribution, r_{XY} = 0.80

Testlet Size (n)	$\overline{x}_{r_{xy}}$ (Reliability)	Standard Deviation	SEM	Upper Bound	Lower Bound
RANDOM: All 1,000,000 (X, Y) Score Pairs					
5	0.746638	0.294890	0.148433	0.999994	-0.994496
10	0.775754	0.177353	0.083985	0.998934	-0.721202
15	0.784004	0.137929	0.064103	0.995575	-0.132289
25	0.789959	0.103911	0.047623	0.990293	0.066454
50	0.794342	0.072327	0.032800	0.963562	0.435149
75	0.795853	0.059454	0.026863	0.941621	0.500931
87	0.796282	0.055152	0.024893	0.937008	0.477813
100	0.796611	0.051414	0.023187	0.937644	0.531078
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs					
5	0.017849	0.475969	0.471702	1.000000	-1.000000
10	0.040324	0.260955	0.255639	0.943797	-1.000000
15	0.044076	0.194058	0.189733	0.763568	-1.000000
25	0.046092	0.142154	0.138840	0.680202	-0.614279
50	0.047537	0.099326	0.096937	0.368721	-0.320435
75	0.047985	0.081803	0.079817	0.354589	-0.242866
87	0.048004	0.077678	0.075791	0.310528	-0.264214
100	0.048035	0.073330	0.071547	0.289611	-0.226492
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs					
5	-0.731452	0.663787	0.873441	1.000000	-1.000000
10	-0.568270	0.793088	0.993188	1.000000	-1.000000
15	-0.473458	0.840918	1.020760	1.000000	-1.000000
25	-0.380641	0.869136	1.021240	1.000000	-1.000000
50	-0.338473	0.864122	0.999724	0.999983	-1.000000
75	-0.338399	0.856936	0.991382	0.999963	-1.000000
87	-0.339000	0.854823	0.989160	0.999387	-1.000000
100	-0.339119	0.852774	0.986833	0.999677	-1.000000

Table 15: Correlation Simulation Results for Exponential Distribution, r_{XY} = 0.90

Testlet Size (n)	$\overline{x}_{r_{xy}}$ (Reliability)	Standard Deviation	SEM	Upper Bound	Lower Bound
RANDOM: All 1,000,000 (X, Y) Score Pairs					
5	0.866287	0.190097	0.069512	0.999999	-0.989303
10	0.887542	0.101821	0.034146	0.999436	-0.186904
15	0.893135	0.076306	0.024945	0.998323	0.101354
25	0.897082	0.055967	0.017955	0.995559	0.443504
50	0.899889	0.038269	0.012109	0.983227	0.685447
75	0.900831	0.031292	0.009854	0.972829	0.723638
87	0.901075	0.029046	0.009136	0.971867	0.724805
100	0.901284	0.026994	0.008481	0.971417	0.748726
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs					
5	0.097450	0.479610	0.455642	1.000000	-1.000000
10	0.122302	0.266891	0.250039	1.000000	-1.000000
15	0.124563	0.206078	0.192816	0.877943	-0.965629
25	0.124953	0.159655	0.149348	0.722197	-0.478932
50	0.124059	0.125570	0.117523	0.514049	-0.229497
75	0.123963	0.114373	0.107049	0.504205	-0.180203
87	0.123992	0.111270	0.104143	0.496912	-0.185403
100	0.123774	0.108545	0.101605	0.470081	-0.188443
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs					
5	-0.814573	0.578475	0.779242	1.000000	-1.000000
10	-0.679793	0.729886	0.945982	1.000000	-1.000000
15	-0.583194	0.806849	1.015220	1.000000	-1.000000
25	-0.458648	0.879215	1.061870	1.000000	-1.000000
50	-0.344228	0.922000	1.068970	1.000000	-1.000000
75	-0.317839	0.926758	1.063890	1.000000	-1.000000
87	-0.320015	0.924339	1.061990	0.999992	-1.000000
100	-0.313881	0.925641	1.061010	1.000000	-1.000000

Table 16: Correlation Simulation Results for Double Exponential Distribution, r_{XY} = 0.71

Testlet Size (n)	$\overline{x}_{r_{xy}}$ (Reliability)	Standard Deviation	SEM	Upper Bound	Lower Bound
RANDOM: All 1,000,000 (X, Y) Score Pairs					
5	0.656802	0.332067	0.194535	0.999952	-0.994667
10	0.686255	0.193686	0.108489	0.993468	-0.784625
15	0.693887	0.148830	0.082344	0.976458	-0.435905
25	0.699208	0.110267	0.060475	0.944224	0.003641
50	0.703125	0.075302	0.041029	0.901741	0.315973
75	0.704457	0.060712	0.033005	0.888952	0.409576
87	0.704757	0.056345	0.030616	0.875635	0.444187
100	0.704972	0.052678	0.028613	0.857856	0.458344
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs					
5	0.045252	0.435398	0.425433	1.000000	-1.000000
10	0.047900	0.232701	0.227059	0.964293	-0.949413
15	0.047574	0.173105	0.168937	0.772539	-0.781319
25	0.047225	0.126912	0.123879	0.552132	-0.481813
50	0.047112	0.087315	0.085234	0.343171	-0.292022
75	0.046647	0.074470	0.072712	0.318446	-0.192458
87	0.046762	0.069796	0.068145	0.299137	-0.180351
100	0.046718	0.064873	0.063339	0.264353	-0.194300
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs					
5	0.544434	0.706743	0.477020	0.999910	-1.000000
10	0.674942	0.508997	0.290199	0.996180	-1.000000
15	0.698341	0.455984	0.250442	0.989690	-1.000000
25	0.710757	0.419601	0.225667	0.980792	-1.000000
50	0.723049	0.379274	0.199598	0.941392	-1.000000
75	0.725590	0.369024	0.193310	0.933462	-0.999998
87	0.727451	0.361926	0.188948	0.930583	-0.999996
100	0.729178	0.354702	0.184589	0.932354	-0.999992

Table 17: Correlation Simulation Results for Double Exponential Distribution, r_{XY} = 0.81

Testlet Size (n)	$\overline{x}_{r_{xy}}$ (Reliability)	Standard Deviation	SEM	Upper Bound	Lower Bound
RANDOM: All 1,000,000 (X, Y) Score Pairs					
5	0.761492	0.264444	0.129147	0.999940	-0.996050
10	0.788279	0.143827	0.066179	0.995885	-0.617093
15	0.794871	0.108077	0.048949	0.985160	-0.236937
25	0.799385	0.078928	0.035352	0.967164	0.188722
50	0.802595	0.053421	0.023735	0.941677	0.514369
75	0.803689	0.042965	0.019037	0.929458	0.593579
87	0.803914	0.039904	0.017670	0.920717	0.604201
100	0.804101	0.037260	0.016491	0.908962	0.625244
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs					
5	0.065084	0.439749	0.425198	1.000000	-1.000000
10	0.069949	0.236133	0.227725	0.929529	-1.000000
15	0.068080	0.177921	0.171758	0.780209	-0.690744
25	0.067326	0.132301	0.127770	0.563912	-0.478871
50	0.066597	0.095085	0.091865	0.414681	-0.218338
75	0.066444	0.080841	0.078109	0.377798	-0.182981
87	0.066340	0.077924	0.075295	0.386326	-0.230243
100	0.066348	0.073932	0.071438	0.332187	-0.162065
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs					
5	0.590215	0.730474	0.467609	0.999976	-1.000000
10	0.734688	0.539045	0.277654	0.997244	-1.000000
15	0.760967	0.488590	0.238876	0.992107	-1.000000
25	0.777099	0.452362	0.213571	0.984782	-1.000000
50	0.792519	0.411342	0.187367	0.970206	-0.999998
75	0.798496	0.393694	0.176726	0.961731	-0.999995
87	0.800544	0.386527	0.172625	0.955075	-0.999993
100	0.802072	0.380828	0.169427	0.960057	-0.999993

Table 18: Correlation Simulation Results for Double Exponential Distribution, r_{XY} = 0.90

Testlet Size (n)	$\overline{x}_{r_{xy}}$ (Reliability)	Standard Deviation	SEM	Upper Bound	Lower Bound
RANDOM: All 1,000,000 (X, Y) Score Pairs					
5	0.873671	0.169362	0.060196	0.999994	-0.987802
10	0.892528	0.081529	0.026728	0.998001	-0.323797
15	0.896825	0.059118	0.018989	0.992788	0.140884
25	0.899667	0.042324	0.013406	0.985344	0.491519
50	0.901609	0.028319	0.008883	0.973995	0.734891
75	0.902266	0.022709	0.007099	0.966124	0.791652
87	0.902388	0.021116	0.006597	0.961807	0.789092
100	0.902507	0.019685	0.006146	0.956051	0.807806
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs					
5	0.114360	0.445719	0.419459	1.000000	-1.000000
10	0.119697	0.245263	0.230116	0.938774	-0.978319
15	0.118283	0.189663	0.178094	0.780296	-0.634929
25	0.117200	0.147994	0.139052	0.745307	-0.491547
50	0.115518	0.115792	0.108899	0.549299	-0.273044
75	0.115415	0.104302	0.098098	0.474291	-0.188625
87	0.115041	0.101157	0.095161	0.474798	-0.187096
100	0.115156	0.098462	0.092619	0.424134	-0.183523
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs					
5	0.623496	0.743849	0.456425	0.999991	-1.000000
10	0.777263	0.559111	0.263873	0.999617	-1.000000
15	0.806952	0.509442	0.223835	0.998451	-1.000000
25	0.827120	0.470482	0.195621	0.991661	-1.000000
50	0.843738	0.433124	0.171214	0.985225	-0.999998
75	0.849864	0.417342	0.161709	0.983464	-0.999991
87	0.850807	0.414677	0.160171	0.982620	-0.999984
100	0.852060	0.411006	0.158085	0.981320	-0.999981

Table 19: Correlation Simulation Results for t (df=3) Distribution, r_{XY} = 0.70

Testlet Size (n)	$\overline{x}_{r_{xy}}$ (Reliability)	Standard Deviation	SEM	Upper Bound	Lower Bound
RANDOM: All 1,000,000 (X, Y) Score Pairs					
5	0.638410	0.368369	0.221509	0.999987	-0.996657
10	0.671196	0.235403	0.134984	0.998672	-0.775982
15	0.680893	0.187075	0.105678	0.992736	-0.344606
25	0.687960	0.143345	0.080073	0.984247	-0.174018
50	0.693301	0.101028	0.055950	0.945010	0.239317
75	0.695167	0.083320	0.046002	0.915033	0.325373
87	0.695747	0.077290	0.042632	0.904212	0.283526
100	0.696151	0.072188	0.039792	0.903241	0.357231
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs					
5	-0.067917	0.495553	0.512105	1.000000	-1.000000
10	-0.038055	0.282323	0.287645	0.948219	-1.000000
15	-0.029626	0.207835	0.210891	0.795783	-0.979547
25	-0.024563	0.150812	0.152653	0.623350	-0.811768
50	-0.021235	0.102219	0.103299	0.300116	-0.524966
75	-0.020559	0.082225	0.083066	0.271324	-0.373250
87	-0.020282	0.076944	0.077720	0.214242	-0.342450
100	-0.019820	0.071771	0.072479	0.212188	-0.311002
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs					
5	-0.986952	0.051418	0.072478	0.968705	-1.000000
10	-0.981680	0.040323	0.056763	0.340329	-1.000000
15	-0.978852	0.039873	0.056090	-0.186233	-1.000000
25	-0.976412	0.038755	0.054483	-0.231862	-1.000000
50	-0.974366	0.037899	0.053252	-0.306706	-1.000000
75	-0.973605	0.037769	0.053060	-0.324579	-1.000000
87	-0.973407	0.038120	0.053551	-0.295313	-1.000000
100	-0.973192	0.038225	0.053695	-0.298611	-1.000000

Table 20: Correlation Simulation Results for t (df=3) Distribution, r_{XY} = 0.81

Testlet Size (n)	$\overline{x}_{r_{xy}}$ (Reliability)	Standard Deviation	SEM	Upper Bound	Lower Bound
RANDOM: All 1,000,000 (X, Y) Score Pairs					
5	0.750764	0.294909	0.147229	0.999998	-0.996548
10	0.781469	0.176705	0.082605	0.998894	-0.792781
15	0.790244	0.137030	0.062759	0.996371	-0.151190
25	0.796594	0.102970	0.046440	0.990721	0.049281
50	0.801243	0.071587	0.031915	0.967008	0.436529
75	0.802849	0.058816	0.026115	0.946234	0.502929
87	0.803302	0.054567	0.024201	0.941774	0.481709
100	0.803654	0.050847	0.022531	0.942410	0.532727
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs					
5	-0.007394	0.482665	0.484446	1.000000	-1.000000
10	0.016482	0.266464	0.264259	0.945731	-1.000000
15	0.021847	0.195461	0.193314	0.736110	-0.994147
25	0.024736	0.142877	0.141099	0.604358	-0.655086
50	0.027130	0.097709	0.096374	0.353821	-0.365411
75	0.027640	0.080358	0.079240	0.293948	-0.226840
87	0.027748	0.073804	0.072773	0.278026	-0.241765
100	0.028056	0.070089	0.069099	0.274772	-0.251767
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs					
5	-0.996616	0.024472	0.034580	0.930882	-1.000000
10	-0.994341	0.024317	0.034341	0.512765	-1.000000
15	-0.993015	0.025460	0.035943	0.418291	-1.000000
25	-0.991607	0.024308	0.034305	-0.080461	-1.000000
50	-0.990340	0.024820	0.035016	-0.068406	-1.000000
75	-0.989757	0.026096	0.036811	-0.038915	-1.000000
87	-0.989564	0.026934	0.037990	-0.015805	-1.000000
100	-0.989444	0.027457	0.038727	-0.034741	-1.000000

Table 21: Correlation Simulation Results for t (df=3) Distribution, r_{XY} = 0.90

Testlet Size (n)	$\overline{x}_{r_{xy}}$ (Reliability)	Standard Deviation	SEM	Upper Bound	Lower Bound
RANDOM: All 1,000,000 (X, Y) Score Pairs					
5	0.865651	0.193293	0.070849	0.999998	-0.991327
10	0.888125	0.103191	0.034515	0.999351	-0.244896
15	0.894074	0.077108	0.025096	0.998511	0.061238
25	0.898279	0.056395	0.017987	0.995614	0.422539
50	0.901261	0.038506	0.012100	0.984420	0.677815
75	0.902265	0.031457	0.009834	0.974393	0.720595
87	0.902525	0.029203	0.009117	0.973110	0.722330
100	0.902749	0.027128	0.008460	0.972884	0.745517
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs					
5	0.082972	0.479012	0.458710	1.000000	-1.000000
10	0.104419	0.265566	0.251319	0.926029	-1.000000
15	0.106623	0.202959	0.191834	0.797181	-0.927650
25	0.107236	0.157279	0.148607	0.659667	-0.619620
50	0.107199	0.122312	0.115570	0.582237	-0.282734
75	0.107486	0.109689	0.103627	0.480817	-0.230149
87	0.107164	0.106067	0.100223	0.486779	-0.183296
100	0.107548	0.103286	0.097574	0.428170	-0.171230
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs					
5	-0.999897	0.008323	0.011771	0.814213	-1.000000
10	-0.999772	0.011218	0.015863	0.723064	-1.000000
15	-0.999672	0.012899	0.018240	0.617836	-1.000000
25	-0.999532	0.013450	0.019019	0.289146	-1.000000
50	-0.999274	0.018551	0.026230	0.284233	-1.000000
75	-0.999080	0.023399	0.033083	0.336396	-1.000000
87	-0.998986	0.024319	0.034383	0.286400	-1.000000
100	-0.998890	0.025824	0.036510	0.273592	-1.000000

Table 22: Correlation Simulation Result Comparison between Mini-Universe Reliability (r_{XY}) and Average Reliability ($\bar{x}_{r_{xy}}$) for Normal Distribution

	Testlet Size (n =)							
	5	10	15	25	50	75	87	100
r_{XY} = 0.700330								
RANDOM: All 1,000,000 (X, Y) Score Pairs								
$\bar{x}_{r_{xy}}$	0.648266	0.678807	0.686831	0.692453	0.696548	0.697971	0.698261	0.698476
$r_{XY} - \bar{x}_{r_{xy}}$	0.052064	0.021523	0.013499	0.007877	0.003782	0.002359	0.002069	0.001854
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs								
$\bar{x}_{r_{xy}}$	0.046198	0.047852	0.047259	0.047078	0.046547	0.046779	0.046531	0.046549
$r_{XY} - \bar{x}_{r_{xy}}$	0.654133	0.652478	0.653071	0.653252	0.653783	0.653552	0.653799	0.653781
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs								
$\bar{x}_{r_{xy}}$	0.751809	0.839778	0.840701	0.838213	0.835971	0.835125	0.834948	0.834642
$r_{XY} - \bar{x}_{r_{xy}}$	-0.051479	-0.139448	-0.140371	-0.137883	-0.135641	-0.134795	-0.134618	-0.134312
r_{XY} = 0.800238								
RANDOM: All 1,000,000 (X, Y) Score Pairs								
$\bar{x}_{r_{xy}}$	0.754190	0.782099	0.789046	0.793804	0.797165	0.798331	0.798549	0.798740
$r_{XY} - \bar{x}_{r_{xy}}$	0.046048	0.018139	0.011192	0.006434	0.003073	0.001907	0.001689	0.001498
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs								
$\bar{x}_{r_{xy}}$	0.065593	0.067836	0.067597	0.066721	0.066165	0.066045	0.065567	0.065906
$r_{XY} - \bar{x}_{r_{xy}}$	0.734646	0.732402	0.732641	0.733517	0.734073	0.734193	0.734672	0.734332
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs								
$\bar{x}_{r_{xy}}$	0.807223	0.909173	0.911257	0.909363	0.907833	0.907264	0.907041	0.906970
$r_{XY} - \bar{x}_{r_{xy}}$	-0.006985	-0.108935	-0.111019	-0.109125	-0.107595	-0.107026	-0.106803	-0.106732
r_{XY} = 0.900144								
RANDOM: All 1,000,000 (X, Y) Score Pairs								
$\bar{x}_{r_{xy}}$	0.868924	0.888731	0.893277	0.896271	0.898309	0.899011	0.899131	0.899254
$r_{XY} - \bar{x}_{r_{xy}}$	0.031220	0.011413	0.006867	0.003873	0.001835	0.001133	0.001013	0.000890
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs								
$\bar{x}_{r_{xy}}$	0.108589	0.114835	0.114506	0.113364	0.112307	0.111620	0.111789	0.111785
$r_{XY} - \bar{x}_{r_{xy}}$	0.791555	0.785309	0.785638	0.786780	0.787837	0.788524	0.788355	0.788359
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs								
$\bar{x}_{r_{xy}}$	0.847437	0.955611	0.958201	0.957248	0.956484	0.956219	0.956154	0.956104
$r_{XY} - \bar{x}_{r_{xy}}$	0.052707	-0.055467	-0.058057	-0.057104	-0.056340	-0.056075	-0.056010	-0.055960

Table 23: Correlation Simulation Result Comparison between Mini-Universe Reliability (r_{XY}) and Average Reliability ($\overline{x}_{r_{xy}}$) for Chi-Squared (df=1) Distribution

	Testlet Size (n =)							
	5	10	15	25	50	75	87	100
r_{XY} = 0.688895								
RANDOM: All 1,000,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	0.578244	0.631987	0.650152	0.664332	0.675827	0.680089	0.681311	0.682199
$r_{XY} - \overline{x}_{r_{xy}}$	0.110651	0.056908	0.038743	0.024563	0.013068	0.008806	0.007584	0.006696
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	-0.134091	-0.095193	-0.081184	-0.072298	-0.066434	-0.064527	-0.064112	-0.063754
$r_{XY} - \overline{x}_{r_{xy}}$	0.822986	0.784088	0.770079	0.761193	0.755329	0.753422	0.753007	0.752649
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	-0.732711	-0.611498	-0.541692	-0.485729	-0.473536	-0.477071	-0.479379	-0.479383
$r_{XY} - \overline{x}_{r_{xy}}$	1.421606	1.300393	1.230587	1.174624	1.162431	1.165966	1.168274	1.168278
r_{XY} = 0.792115								
RANDOM: All 1,000,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	0.692427	0.743291	0.759299	0.771740	0.781376	0.784936	0.785849	0.786615
$r_{XY} - \overline{x}_{r_{xy}}$	0.099688	0.048824	0.032816	0.020375	0.010739	0.007179	0.006266	0.005500
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	-0.062570	-0.027156	-0.015856	-0.009824	-0.005572	-0.004180	-0.003948	-0.003551
$r_{XY} - \overline{x}_{r_{xy}}$	0.854685	0.819271	0.807971	0.801939	0.797687	0.796295	0.796063	0.795666
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	-0.777367	-0.641824	-0.557654	-0.464938	-0.406994	-0.402788	-0.404829	-0.405026
$r_{XY} - \overline{x}_{r_{xy}}$	1.569482	1.433939	1.349769	1.257053	1.199109	1.194903	1.196944	1.197141
r_{XY} = 0.895831								
RANDOM: All 1,000,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	0.826034	0.864316	0.875128	0.883306	0.889351	0.891534	0.892032	0.892523
$r_{XY} - \overline{x}_{r_{xy}}$	0.069797	0.031515	0.020703	0.012525	0.006480	0.004297	0.003799	0.003308
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	0.045126	0.080316	0.086353	0.090216	0.091861	0.092267	0.092241	0.092667
$r_{XY} - \overline{x}_{r_{xy}}$	-0.940957	-0.976147	-0.982184	-0.986047	-0.987692	-0.988098	-0.988072	-0.988498
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	-0.862054	-0.753072	-0.667958	-0.551014	-0.411054	-0.366618	-0.360526	-0.353697
$r_{XY} - \overline{x}_{r_{xy}}$	1.757885	1.648903	1.563789	1.446845	1.306885	1.262449	1.256357	1.249528

Table 24: Correlation Simulation Result Comparison between Mini-Universe Reliability (r_{XY}) and Average Reliability ($\overline{x}_{r_{xy}}$) for Exponential Distribution

	Testlet Size (n =)							
	5	10	15	25	50	75	87	100
r_{XY} = 0.690861								
RANDOM: All 1,000,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	0.634620	0.665239	0.674201	0.680697	0.685639	0.687359	0.687899	0.688268
$r_{XY} - \overline{x}_{r_{xy}}$	0.056241	0.025622	0.016660	0.010164	0.005222	0.003502	0.002962	0.002593
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	-0.024895	-0.002733	0.002983	0.006769	0.008481	0.009055	0.009225	0.009293
$r_{XY} - \overline{x}_{r_{xy}}$	0.715756	0.693594	0.687878	0.684092	0.682380	0.681806	0.681636	0.681568
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	-0.660248	-0.503512	-0.437448	-0.385978	-0.380432	-0.382407	-0.383959	-0.384412
$r_{XY} - \overline{x}_{r_{xy}}$	1.351109	1.194373	1.128309	1.076839	1.071293	1.073268	1.074820	1.075273
r_{XY} = 0.798917								
RANDOM: All 1,000,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	0.746638	0.775754	0.784004	0.789959	0.794342	0.795853	0.796282	0.796611
$r_{XY} - \overline{x}_{r_{xy}}$	0.052279	0.023163	0.014913	0.008958	0.004575	0.003064	0.002635	0.002306
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	0.017849	0.040324	0.044076	0.046092	0.047537	0.047985	0.048004	0.048035
$r_{XY} - \overline{x}_{r_{xy}}$	0.781069	0.758593	0.754841	0.752825	0.751380	0.750932	0.750913	0.750882
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	-0.731452	-0.568270	-0.473458	-0.380641	-0.338473	-0.338399	-0.339000	-0.339119
$r_{XY} - \overline{x}_{r_{xy}}$	1.530369	1.367187	1.272375	1.179558	1.137390	1.137316	1.137917	1.138036
r_{XY} = 0.902728								
RANDOM: All 1,000,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	0.866287	0.887542	0.893135	0.897082	0.899889	0.900831	0.901075	0.901284
$r_{XY} - \overline{x}_{r_{xy}}$	0.036441	0.015186	0.009593	0.005646	0.002839	0.001897	0.001653	0.001444
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	0.097450	0.122302	0.124563	0.124953	0.124059	0.123963	0.123992	0.123774
$r_{XY} - \overline{x}_{r_{xy}}$	0.805278	0.780426	0.778165	0.777775	0.778669	0.778765	0.778736	0.778954
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	-0.814573	-0.679793	-0.583194	-0.458648	-0.344228	-0.317839	-0.320015	-0.313881
$r_{XY} - \overline{x}_{r_{xy}}$	1.717301	1.582521	1.485922	1.361376	1.246956	1.220567	1.222743	1.216609

Table 25: Correlation Simulation Result Comparison between Mini-Universe Reliability (r_{XY}) and Average Reliability ($\overline{x}_{r_{xy}}$) for Double Exponential Distribution

	Testlet Size (n =)							
	5	10	15	25	50	75	87	100
r_{XY} = 0.706736								
RANDOM: All 1,000,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	0.656802	0.686255	0.693887	0.699208	0.703125	0.704457	0.704757	0.704972
$r_{XY} - \overline{x}_{r_{xy}}$	0.049934	0.020481	0.012849	0.007528	0.003611	0.002279	0.001979	0.001764
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	0.045252	0.047900	0.047574	0.047225	0.047112	0.046647	0.046762	0.046718
$r_{XY} - \overline{x}_{r_{xy}}$	0.661484	0.658836	0.659163	0.659511	0.659624	0.660089	0.659974	0.660018
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	0.544434	0.674942	0.698341	0.710757	0.723049	0.725590	0.727451	0.729178
$r_{XY} - \overline{x}_{r_{xy}}$	0.162302	0.031794	0.008395	-0.004021	-0.016313	-0.018854	-0.020715	-0.022442
r_{XY} = 0.805540								
RANDOM: All 1,000,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	0.761492	0.788279	0.794871	0.799385	0.802595	0.803689	0.803914	0.804101
$r_{XY} - \overline{x}_{r_{xy}}$	0.044048	0.017261	0.010669	0.006155	0.002945	0.001851	0.001626	0.001439
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	0.065084	0.069949	0.068080	0.067326	0.066597	0.066444	0.066340	0.066348
$r_{XY} - \overline{x}_{r_{xy}}$	0.740456	0.735591	0.737460	0.738214	0.738943	0.739096	0.739200	0.739193
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	0.590215	0.734688	0.760967	0.777099	0.792519	0.798496	0.800544	0.802072
$r_{XY} - \overline{x}_{r_{xy}}$	0.215325	0.070852	0.044573	0.028441	0.013021	0.007044	0.004996	0.003468
r_{XY} = 0.903367								
RANDOM: All 1,000,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	0.866287	0.887542	0.893135	0.897082	0.899889	0.900831	0.901075	0.901284
$r_{XY} - \overline{x}_{r_{xy}}$	0.036441	0.015186	0.009593	0.005646	0.002839	0.001897	0.001653	0.001444
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	0.097450	0.122302	0.124563	0.124953	0.124059	0.123963	0.123992	0.123774
$r_{XY} - \overline{x}_{r_{xy}}$	0.805278	0.780426	0.778165	0.777775	0.778669	0.778765	0.778736	0.778954
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs								
$\overline{x}_{r_{xy}}$	-0.814573	-0.679793	-0.583194	-0.458648	-0.344228	-0.317839	-0.320015	-0.313881
$r_{XY} - \overline{x}_{r_{xy}}$	1.717301	1.582521	1.485922	1.361376	1.246956	1.220567	1.222743	1.216609

Table 26: Correlation Simulation Result Comparison between Mini-Universe Reliability (r_{XY}) and Average Reliability ($\bar{x}_{r_{xy}}$) for t (df=3) Distribution

	Testlet Size (n =)							
	5	10	15	25	50	75	87	100
r_{XY} = 0.698982								
RANDOM: All 1,000,000 (X, Y) Score Pairs								
$\bar{x}_{r_{xy}}$	0.638410	0.671196	0.680893	0.687960	0.693301	0.695167	0.695747	0.696151
$r_{XY} - \bar{x}_{r_{xy}}$	0.060572	0.027786	0.018089	0.011022	0.005681	0.003815	0.003235	0.002831
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs								
$\bar{x}_{r_{xy}}$	-0.067917	-0.038055	-0.029626	-0.024563	-0.021235	-0.020559	-0.020282	-0.019820
$r_{XY} - \bar{x}_{r_{xy}}$	0.766899	0.737037	0.728608	0.723545	0.720217	0.719541	0.719264	0.718802
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs								
$\bar{x}_{r_{xy}}$	-0.986952	-0.981680	-0.978852	-0.976412	-0.974366	-0.973605	-0.973407	-0.973192
$r_{XY} - \bar{x}_{r_{xy}}$	1.685934	1.680662	1.677834	1.675394	1.673348	1.672587	1.672389	1.672174
r_{XY} = 0.806113								
RANDOM: All 1,000,000 (X, Y) Score Pairs								
$\bar{x}_{r_{xy}}$	0.750764	0.781469	0.790244	0.796594	0.801243	0.802849	0.803302	0.803654
$r_{XY} - \bar{x}_{r_{xy}}$	0.055349	0.024644	0.015869	0.009519	0.004870	0.003264	0.002811	0.002459
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs								
$\bar{x}_{r_{xy}}$	-0.007394	0.016482	0.021847	0.024736	0.027130	0.027640	0.027748	0.028056
$r_{XY} - \bar{x}_{r_{xy}}$	0.813507	0.789631	0.784266	0.781377	0.778983	0.778473	0.778365	0.778057
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs								
$\bar{x}_{r_{xy}}$	-0.996616	-0.994341	-0.993015	-0.991607	-0.990340	-0.989757	-0.989564	-0.989444
$r_{XY} - \bar{x}_{r_{xy}}$	1.802729	1.800454	1.799128	1.797720	1.796453	1.795870	1.795677	1.795557
r_{XY} = 0.904287								
RANDOM: All 1,000,000 (X, Y) Score Pairs								
$\bar{x}_{r_{xy}}$	0.865651	0.888125	0.894074	0.898279	0.901261	0.902265	0.902525	0.902749
$r_{XY} - \bar{x}_{r_{xy}}$	0.038636	0.016162	0.010213	0.006008	0.003026	0.002022	0.001762	0.001538
NON-RANDOM ASCENDING: Lowest 250,000 (X, Y) Score Pairs								
$\bar{x}_{r_{xy}}$	0.082972	0.104419	0.106623	0.107236	0.107199	0.107486	0.107164	0.107548
$r_{XY} - \bar{x}_{r_{xy}}$	0.821315	0.799868	0.797664	0.797051	0.797088	0.796801	0.797123	0.796739
NON-RANDOM DESCENDING: Highest 250,000 (X, Y) Score Pairs								
$\bar{x}_{r_{xy}}$	-0.999897	-0.999772	-0.999672	-0.999532	-0.999274	-0.999080	-0.998986	-0.998890
$r_{XY} - \bar{x}_{r_{xy}}$	1.904184	1.904059	1.903959	1.903819	1.903561	1.903367	1.903273	1.903177

Table 27: Correlation Simulation Results for All Distributions *n* = 87, Set *r* = 0.70

Distribution	Sampling Method	$\overline{x}_{r_{xy}}$ *	Standard Deviation	SEM	Upper Bound	Lower Bound
Normal	Random	0.698261	0.055614	0.030549	0.869163	0.438630
	Non-Random Ascending	0.046531	0.069188	0.067559	0.274246	-0.198798
	Non-Random Descending	0.834948	0.085649	0.034796	0.981230	0.559541
Chi-Squared (df=1)	Random	0.681311	0.092823	0.052401	0.928377	0.135081
	Non-Random Ascending	-0.064112	0.093552	0.096505	0.183379	-0.615839
	Non-Random Descending	-0.479379	0.677985	0.824632	0.976572	-0.999999
Exponential	Random	0.687899	0.077317	0.043194	0.896931	0.284520
	Non-Random Ascending	0.009225	0.068417	0.068101	0.220837	-0.283320
	Non-Random Descending	-0.383959	0.765221	0.900220	0.951908	-1.000000
Double Exponential	Random	0.704757	0.056345	0.030616	0.875635	0.444187
	Non-Random Ascending	0.046762	0.069796	0.068145	0.299137	-0.180351
	Non-Random Descending	0.727451	0.361926	0.188948	0.930583	-0.999996
t (df=3)	Random	0.695747	0.077290	0.042632	0.904212	0.283526
	Non-Random Ascending	-0.020282	0.076944	0.077720	0.214242	-0.342450
	Non-Random Descending	-0.973407	0.038120	0.053551	-0.295313	-1.000000

*Average reliability for *n* = 87

Table 28: Correlation Simulation Results for All Distributions *n* = 87, Set *r* = 0.80

Distribution	Sampling Method	$\overline{x}_{r_{xy}}$ *	Standard Deviation	SEM	Upper Bound	Lower Bound
Normal	Random	0.798549	0.039494	0.017726	0.916329	0.606996
	Non-Random Ascending	0.065567	0.077421	0.074839	0.371189	-0.202954
	Non-Random Descending	0.907041	0.041914	0.012779	0.987120	0.746294
Chi-Squared (df=1)	Random	0.785849	0.067909	0.031426	0.954118	0.325845
	Non-Random Ascending	-0.003948	0.078240	0.078395	0.315674	-0.342294
	Non-Random Descending	-0.404829	0.785364	0.930856	0.999616	-1.000000
Exponential	Random	0.796282	0.055152	0.024893	0.937008	0.477813
	Non-Random Ascending	0.048004	0.077678	0.075791	0.310528	-0.264214
	Non-Random Descending	-0.339000	0.854823	0.989160	0.999387	-1.000000
Double Exponential	Random	0.803914	0.039904	0.017670	0.920717	0.604201
	Non-Random Ascending	0.066340	0.077924	0.075295	0.386326	-0.230243
	Non-Random Descending	0.800544	0.386527	0.172625	0.955075	-0.999993
t (df=3)	Random	0.803302	0.054567	0.024201	0.941774	0.481709
	Non-Random Ascending	0.027748	0.073804	0.072773	0.278026	-0.241765
	Non-Random Descending	-0.989564	0.026934	0.037990	-0.015805	-1.000000

*Average reliability for *n* = 87

Table 29: Correlation Simulation Results for All Distributions *n* = 87, Set *r* = 0.90

Distribution	Sampling Method	$\overline{x}_{r_{xy}}$ *	Standard Deviation	SEM	Upper Bound	Lower Bound
Normal	Random	0.899131	0.020983	0.006664	0.959613	0.791521
	Non-Random Ascending	0.111789	0.098187	0.092536	0.413281	-0.145094
	Non-Random Descending	0.956154	0.018753	0.003927	0.993684	0.883418
Chi-Squared (df=1)	Random	0.892032	0.037115	0.012196	0.977841	0.611253
	Non-Random Ascending	0.092241	0.099649	0.094942	0.455258	-0.189479
	Non-Random Descending	-0.360526	0.891725	1.040120	0.999998	-1.000000
Exponential	Random	0.901075	0.029046	0.009136	0.971867	0.724805
	Non-Random Ascending	0.123992	0.111270	0.104143	0.496912	-0.185403
	Non-Random Descending	-0.320015	0.924339	1.061990	0.999992	-1.000000
Double Exponential	Random	0.902388	0.021116	0.006597	0.961807	0.789092
	Non-Random Ascending	0.115041	0.101157	0.095161	0.474798	-0.187096
	Non-Random Descending	0.850807	0.414677	0.160171	0.982620	-0.999984
t (df=3)	Random	0.902525	0.029203	0.009117	0.973110	0.722330
	Non-Random Ascending	0.107164	0.106067	0.100223	0.486779	-0.183296
	Non-Random Descending	-0.998986	0.024319	0.034383	0.286400	-1.000000

*Average reliability for *n* = 87

CHAPTER 5

DISCUSSION

The Pearson product-moment correlation coefficient (Pearson's *r*) is used to specify the degree of linear relationship between two variables expressed in the form of standard scores (Nunnally, 1978). The correlation coefficient represents the extent to which individuals exhibit the same score from one set of measures to the next; therefore, casting reliability in terms of the correlation between parallel tests is another way of describing precision of measurement. When the coefficient is close to zero, it indicates that an individual's scores over some number of parallel tests show a great deal of variation, meaning the instrument (i.e., test) provides unstable or inconsistent measurements; by contrast, when the coefficient is close to +1.00 it means the individual's scores are very nearly the same indicating that the instrument provides stable, consistent measurements.

Creating the mini-universes and setting the degree of relationship (correlation) between the (*X*, *Y*) correlates (0.70, 0.80, 0.90) sets the degree to which systematic error is present from test to test. It is then possible to attribute differences between the known correlation of the mini-universe and the calculated correlations (i.e., the reliability coefficients) of testlets sampled from the mini-universe to random (non-systematic) error, or sampling error, that would be expected across different test administrations. Creating the mini-universes mirrors the norming process for a test and simulations using random sampling result in a range of reliabilities for various size testlets. These coefficients should be close to the known reliability of the mini-universe and have small standard deviations and standard errors of measurement because sampling error is very low when randomization is not violated. When randomness is removed by ordering the correlations in the mini-universe from low to high (or vice versa), sample range is restricted and sampling

error is introduced – this mirrors typical test administrations to small, non-random groups of examinees – the reliability of the testlets sampled from these non-random pools should differ from that of the known mini-universe reliability and greater standard deviations and standard errors of measurement should be observed.

The purpose of this study was to determine if the fluctuation in estimates of reliability under proper experimental conditions can be fully explained via classical measurement theory without resorting to reliability generalization. To answer this research question, two statistical questions were posed:

(1) Can the r_{xy}, where (x, y) = 5, 10, 15, 25, 50, 75, 87 and 100, equal r_{XY}, where (X, Y) = 1,000,000, within the sampling error of SEM as predicted by classical measurement theory, eliminating the notion of reliability generalization?

(2) Does the r_{xy}, where (x, y) = 5, 10, 15, 25, 50, 75, 87 and 100, equal r_{XY}, where (X, Y) = 1,000,000, when administration protocols (i.e., lack of randomization) are violated as predicted by classical measurement theory, eliminating the notion of reliability generalization?

The answer to the research question based on results for the statistical questions is yes; the fluctuation in reliability estimates can be fully explained via classical measurement theory under proper experimental conditions without resorting to reliability generalization.

Random Samples

With respect to statistical question (1), results for randomly sampled testlets show that, in general, as the testlet size increases, the reliability of the testlets converges to that of the known reliability of the mini-universe. This is a well-known phenomenon in psychometrics – that as test length increases reliability also tends to increase. As stated by Ponterotto and Ruckdeschel

(2007) “to the extent that increasing sample size promotes larger variance in construct measurement, then larger samples will lead to greater reliability in measurement” (p. 1001). Nunnally (1978) also stated “the reliability of scores obtained on a sample of items from a domain increases with the number of items sampled” (p. 209). This convergence trend is observed in random simulations for all five distributions when r_{XY} approximates 0.70, 0.80 and 0.90. In addition, for all distributions, both the standard deviation and the standard error of measurement for all randomly sampled testlets decreased as testlet size increased.

The average reliabilities for n = 87 randomly sampled testlets are compiled in Table 30. Average reliability for all testlet sizes varies little from that of the known reliability of the mini-universes and in all cases standard deviations and standard error of measurement are very small. Recall that classical test theory is based on the notion that observed scores (X) are composites of true scores (T) and measurement error (E) (Nunnally, 1978; Lord & Novick, 1968) in the relationship expressed as X = T + E. In this expression, T is theoretically pure, meaning that this value would not change if the same person was given perfectly parallel versions of a test an infinite number of times; thus, T is not influenced by measurement error. In practice, it is only possible to know X, the observed score, because error is always present to some degree, and “these random deviations from true score are caused by measurement error” (Thye, 2000, p. 1280).

That average results would be very close to the known reliability of the mini-universes was expected; that is, administering the same test over and over to random groups should result in an average reliability that is very close to that of the mini-universe or known reliability. However, sampling error – which influences even randomly selected groups to some degree – can cause reliabilities for individual test administrations to differ from that of the norm group

and from that of the same test given to a different group of test takers. In this study, testlets size *n* = 87 sampled from the normal distribution produced reliability coefficients as low as 0.44 (lower bound) and as high as 0.87 (upper bound) from a mini-universe with a known reliability (r_{XY}) of 0.70 and low and high values (lower and upper bounds) of 0.61 and 0.92 for r_{XY}=0.80, and 0.79 and 0.96 for r_{XY}=0.90 (see Table 27). Similar results are observed for all five distributions at all three levels of correlation. This shows that even under pristine conditions when systematic error is held constant random error influences the reliability of a test. Thus, it is not surprising that a test administrator would find some difference between the reliability calculated for a small group of examinees compared to the reliability of the norm group, however, if random sampling is employed to select the group of examines, that difference should be small.

Table 30: Random Correlation Simulation Results for All Distributions *n* = 87

Distribution	$\bar{x}_{r_{xy}}$ *	Standard Deviation	SEM	r_{XY} **	$r_{XY} - \bar{x}_{r_{xy}}$
Normal	0.698261	0.055614	0.030549	0.700330	0.002069
Chi-Squared (df=1)	0.681311	0.092823	0.052401	0.688895	0.007584
Exponential	0.687899	0.077317	0.043194	0.690861	0.002962
Double Exponential	0.704757	0.056345	0.030616	0.706736	0.001979
t (df=3)	0.695747	0.077290	0.042632	0.698982	0.003235
Normal	0.798549	0.039494	0.017726	0.800238	0.001689
Chi-Squared (df=1)	0.785849	0.067909	0.031426	0.792115	0.006266
Exponential	0.796282	0.055152	0.024893	0.798917	0.002635
Double Exponential	0.803914	0.039904	0.017670	0.805540	0.001626
t (df=3)	0.803302	0.054567	0.024201	0.806113	0.002811
Normal	0.899131	0.020983	0.006664	0.900144	0.001013
Chi-Squared (df=1)	0.892032	0.037115	0.012196	0.895831	0.003799
Exponential	0.901075	0.029046	0.009136	0.902728	0.001653
Double Exponential	0.902388	0.021116	0.006597	0.903367	0.000979
t (df=3)	0.902525	0.029203	0.009117	0.904287	0.001762

*Average reliability for *n* = 87; **Mini-universe actual reliability

Non-Random Samples

To remove randomization from the testlets to answer statistical question (2), correlates in the mini-universes were ordered from low to high by adding the absolute value of the (*X*, *Y*) pairs and sorting based on the sum. After ordering the data, testlets were sampled from within the lowest 250,000 correlates and then from within the highest 250,000 correlates, resulting in severe sampling range restriction. Other methods exist for removing randomization from the data, but this particular method provides a way to isolate the smallest and largest scores which, when testlets are sampled from these groups, simulate the worst case scenarios for test administration. Sampling from only the lowest correlates simulates administering a test to a very low performing group of examinees and sampling from the highest correlates simulates administering a test to a very high performing group of examinees.

When sampling error is introduced, the average correlations for most of the testlets, regardless of size, change drastically and fail to properly estimate the known correlation of the mini-universe. Tables 31-32 show the average reliabilities for n = 87 non-randomly sampled testlets. For example, looking at the normal distribution with a known mini-universe reliability of 0.70 and non-randomly selecting testlets from only the lowest 250,000 correlates, the average reliability for testlets size n = 87 drops to 0.05; likewise, when selecting from only the 250,000 highest correlates the average reliability increases to 0.84. Compared to a reliability of 0.70 when the n = 87 testlets are randomly sampled, neither of these values is an accurate reflection of the known reliability that they should be estimating. Similar results are observed across all distributions for all testlet sizes at all three levels (0.70, 0.80 and 0.90). This shows that reliability of a test will differ depending on the characteristics of the group of examinees taking the test; this was described by Harvill (1991) as follows:

The reliability of a test is not a fixed value. It will vary among different methods for determining reliability using a single group of examinees and among different groups of examinees using a single method for estimating reliability. The manual for a standardized test may report many reliability coefficients obtained for that test using different methods and different groups of examinees. (p. 182)

Table 31: Non-Random Ascending Correlation Simulation Results, All Distributions *n* = 87

Distribution	$\bar{x}_{r_{xy}}$ *	Standard Deviation	SEM	r_{XY} **	$r_{XY} - \bar{x}_{r_{xy}}$
Normal	0.046531	0.069188	0.067559	0.700330	0.653799
Chi-Squared (df=1)	-0.064112	0.093552	0.096505	0.688895	0.753007
Exponential	0.009225	0.068417	0.068101	0.690861	0.681636
Double Exponential	0.046762	0.069796	0.068145	0.706736	0.659974
t (df=3)	-0.020282	0.076944	0.077720	0.698982	0.719264
Normal	0.065567	0.077421	0.074839	0.800238	0.734671
Chi-Squared (df=1)	-0.003948	0.078240	0.078395	0.792115	0.796063
Exponential	0.048004	0.077678	0.075791	0.798917	0.750913
Double Exponential	0.066340	0.077924	0.075295	0.805540	0.739200
t (df=3)	0.027748	0.073804	0.072773	0.806113	0.778365
Normal	0.111789	0.098187	0.092536	0.900144	0.788355
Chi-Squared (df=1)	0.092241	0.099649	0.094942	0.895831	-0.988072
Exponential	0.123992	0.111270	0.104143	0.902728	0.778736
Double Exponential	0.115041	0.101157	0.095161	0.903367	0.788326
t (df=3)	0.107164	0.106067	0.100223	0.904287	0.797123

*Average reliability for *n* = 87; **Mini-universe actual reliability

Table 32: Non-Random Descending Correlation Simulation Results, All Distributions *n* = 87

Distribution	$\bar{x}_{r_{xy}}$ *	Standard Deviation	SEM	r_{XY} **	$r_{XY} - \bar{x}_{r_{xy}}$
Normal	0.834948	0.085649	0.034796	0.700330	-0.134618
Chi-Squared (df=1)	-0.479379	0.677985	0.824632	0.688895	1.168274
Exponential	-0.383959	0.765221	0.900220	0.690861	1.074820
Double Exponential	0.727451	0.361926	0.188948	0.706736	-0.020715
t (df=3)	-0.973407	0.038120	0.053551	0.698982	1.672389
Normal	0.907041	0.041914	0.012779	0.800238	-0.106803
Chi-Squared (df=1)	-0.404829	0.785364	0.930856	0.792115	1.196944
Exponential	-0.339000	0.854823	0.989160	0.798917	1.137917
Double Exponential	0.800544	0.386527	0.172625	0.805540	0.004996
t (df=3)	-0.989564	0.026934	0.037990	0.806113	1.795677
Normal	0.956154	0.018753	0.003927	0.900144	-0.05601
Chi-Squared (df=1)	-0.360526	0.891725	1.040120	0.895831	1.256357
Exponential	-0.320015	0.924339	1.061990	0.902728	1.222743
Double Exponential	0.850807	0.414677	0.160171	0.903367	0.052560
t (df=3)	-0.998986	0.024319	0.034383	0.904287	1.903273

*Average reliability for *n* = 87; **Mini-universe actual reliability

Summary

The proposal of the reliability generalization (RG) method in 1998 spawned a cottage industry and created a split in the psychometric community; this is disquieting given the fact that RG essentially ignores decades of work in psychometric theory and application. As discussed earlier, several issues pertinent to the method have been put forth since its introduction. Problems identified with using RG to examine reliability coefficients across studies (or across test administrations) are not limited to, but include:

- Comparing reliability coefficients obtained via different calculation methods (e.g., test-retest, Cronbach's α, KR-20, etc.) which are known to differ, even when calculated for the same group and test administration;

- Failing to consider different errors associated with different types of reliability;
- Not considering different test forms or formats and how reliability can differ depending on the form used;
- Neglecting to account for errors associated with test administration conditions, including not confirming that test administration protocols were followed properly;
- Failure to take into consideration the composition of test groups (sample) and sampling error.

As discussed in Chapter 3, extensive arguments describing how and why these issues are problematic in RG studies have been put forth by several researchers. In addition, researchers such as Symonds (1928), Sawilowsky (2000) and Schumaker and Smith (2007) documented a variety of other factors that influence test reliability, these include: the test itself, conditions under which a test is administered, the individuals taking the test and interactions among these factors. (See Table 1.) Each issue on its own gives cause to question the RG method, but taken together they form a strong basis to argue that RG is not a suitable method for the study of reliability. Additional support for this argument may also be found somewhat unexpectedly among RG studies that have been conducted since the method's introduction.

Reviewing research carried out over the years, a number of issues with the RG method are found within RG studies themselves. For example, some of the studies purport to examine reliability, but upon closer inspection actually appear to focus on measurement error (e.g., Vacha-Haase, 1998; Baugh, 2002) and most fail to describe or even acknowledge known issues regarding combining reliability coefficients (e.g., Charter, 2003). Others report finding very narrow confidence bands in reliability coefficients across studies for different groups using the same test (e.g., Henson, 2004) and some simply restate already known aspects of reliability, such

as the fact that sampling group variability influences reliability (e.g., Fan & Yin, 2003). Thus, those who have employed RG have done little to illustrate benefits to using the method to study reliability or to instill confidence in RG results. In addition to the above, one other issue with reliability that is acknowledged among both proponents and opponents of RG is that of failures in study result reporting.

Not surprisingly, Vacha-Hasse (1998) observed differences among reliability coefficient values calculated from different administrations of the same test. In order to address these differences, the RG method was proposed, seemingly without consideration of the various factors which are known to influence reliability. For example, it was noted previously that it is inappropriate to definitively assert that the reliability of a test is a specific value (Sawilowsky, 2000): This is due to the fact that reliability coefficients can be calculated using several different approaches and each will provide a somewhat different estimate of reliability - even for the same administration to a group of examinees - because each accounts for different types of errors. This study employed test-retest reliability to specifically examine the impact of one factor on reliability coefficients: random versus non-random sampling. Interestingly, Vacha-Haase (1998) addressed this by quoting Dawis (1987) who said "because reliability is a function of sample as well as of instrument, it should be evaluated on a sample from the intended target population, an obvious but sometimes overlooked point" (p. 839). It has been shown in this research that this is indeed the case: Reliability coefficients calculated for the same test but with different samples of examinees will differ from the reliability calculated for the norm group. The size of the difference depends upon how much sampling error is present, which is in part due to the sampling procedure used (random vs. non-random) to select the group of examinees to whom the test is administered. In short: reliability is sample dependent. Using Monte Carlo simulations to

illustrate how reliability coefficients vary considering just one aspect of measurement error, it can be said that reliability generalization is an unnecessary and potentially misleading practice. As Nunnally (1978) aptly stated:

> …it is meaningful to think of a test as having a number of different coefficients of reliability, depending on the major sources of measurement error that are considered. In practice, however, it is useful to speak of *a* reliability coefficient for a test which summarizes the amount of measurement error expected from using the instrument. This striving for simplicity is understandable, but at least two types of reliability coefficients should be computed and reported for any test that will be employed widely.
>
> To the extent that different approaches to obtaining the reliability coefficient produce somewhat different results, the coefficient that should be used in gauging the stability of traits and in making statistical corrections depends on the way in which the measurement method will be employed. (p. 237)

Implications for Future Research

This study employed Monte Carlo methods to simulate data from five different distributions. The five distributions selected represent the various characteristics commonly present in data collected in social and behavioral research. Yuan and Bentler (2002) commented on this, when they stated that "the asymptotic distribution for each of the [reliability] coefficient estimates, obtained based on a normal sampling distribution, is still valid within a large class of non-normal distributions. Therefore…can still be used even with skewed and kurtotic data such as are typical in the social and behavioral sciences" (p. 251). Simulated data was employed in this research in order to isolate random sampling error to show its effects on reliability coefficients. Future research could utilize real data from either a norm group or other research studies to populate a correlation matrix and calculate reliability coefficients using a similar sampling technique. Conducting a similar study using real data – as opposed to simulated test results – will provide additional evidence for statements regarding the reliability of tests.

Conclusion

This research demonstrates that calculated reliability can and will differ between test administrations due to sampling, and that these differences in calculated coefficients are predictable and can be explained mathematically. Along with this, results from this work illustrate how critical it is that investigators describe the sample and sampling method used in a study. It is thus recommended that workers would be wise to forgo reliability generalization and instead focus on developing practices to encourage researchers to report any reliability coefficients noted in test manuals (e.g., that for the norm group) along with the reliability coefficient(s) calculated for tests administered in their own studies, the type(s) of reliability calculated, sampling plan, sample characteristics and test administration conditions. The results of this Monte Carlo study demonstrate that the concept, nomenclature and burgeoning cottage industry of reliability generalization are merely *lapsus linguae*.

REFERENCES

Algina, J., Oshima, T. C. & Tang, K. L. (1991). Robustness of Yao's, James's, and Johansen's tests under variance-covariance heteroscedasticity and nonnormality. *Journal of Educational Statistics, 16*(2), 125-139.

Allen, M. & Yen, W. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Alsawalmeh, Y. M. & Feldt, L. S. (1992). Test of the hypothesis that the intraclass reliability coefficient is the same for two measurement procedures. *Applied Psychological Measurement, 16*(2), 195-205.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). The standards for educational and psychological testing. Washington, DC: American Psychological Association.

Anastasi, A. (1976). *Psychological testing*. New York, NY: MacMillan Publishing Company, Inc.

Armstrong, R. D., Jones, D. H. & Wang, Z. (1994). Automated parallel test construction using classical test theory. *Journal of Educational and Behavioral Statistics*, *19*(1), 73-90.

Armstrong, R. D., Jones, D. H. & Wang, Z. (1998). Optimization of classical reliability in test construction. *Journal of Educational and Behavioral Statistics*, *23*(1), 1-17.

Barrett, P. (2005). What if there were no psychometrics?: Constructs, complexity, and measurement. *Journal of Personality Assessment, 85*(2), 134-140.

Bartholomew, D. J. (1998). Scaling unobservable constructs in social science. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, *47*(1), 1-13.

Baugh, F. (2002). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement*, *62*(2), 254-263.

Bechger, T. M., et al. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement, 27*(5), 319-334.

Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods*, *5*(3), 370-379.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliablity. *Psychometrika*, *74*(1), 137-143.

Beretvas, S. N., et al. (2008). A reliability generalization study of scores on Rotter's and Nowicki-Strickland's locus of control scales. *Educational and Psychological Measurement*, *68*(1), 97-119.

Birenbaum, M., Tatsuoka, K. K. & Nasser, F. (1997). On the agreement of diagnostic classifications from parallel subtests: score reliability at the micro level. *Educational and Psychological Measurement*, *57*, 541-558.

Bonnett, D. G. (2002). Sample size requirements for testing and estimating α coefficients. *Journal of Educational and Behavioral Statistics*, *27*, 335-340.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425-440.

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating easy readers. *Journal of Educational Statistics*, *13*(1), 1-18.

Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational Measurement: Issues and Practice*, *16*(4), pp. 14-20.

Brennan, R. L. (2000). (Mis)conceptions about generalizability theory. *Educational Measurement: Issues and Practice*, *19*(1), pp. 5-10.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *38*(4), 295-317.

Brennan, R. L., Gao, X. & Colton, D. A. (1995). Generalizability analyses of work keys listening and writing tests. *Educational and Psychological Measurement*, *55*, 157-176.

Bridge, P. D. & Sawilowsky, S. S. (1999). Increasing physicians' awareness of the impact of statistics on research outcomes: Comparative power the t-test and Wilcoxon rank-sum test in small samples applied research. *Journal of Clinical Epidemiology*, *52*(3), 229-235.

Bryant, N. C. & Barnes, L. B. (1997). Development and validation of the attitude toward educational measurement inventory. *Educational and Psychological Measurement*, *57*, 870-875.

Capraro, M. M., Capraro, R. M., & Henson, R. K. (2001). Measurement error of scores on the Mathematics Anxiety Rating Scale across studies. *Educational and Psychological Measurement*, *61*, 373–386.

Caruso, J. C. (2000). Reliability generalization of the NEO personality scales. *Educational and Psychological Measurement*, *60*(2), 236-254.

Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *The Journal of General Psychology*, *130*(3), 290-304.

Charter, R. A. (2003). Combining reliability coefficients: Possible application to meta-analysis and reliability generalization. *Psychological Reports*, *93*, 643-647.

Charter, R. A. (2008). Statistical approaches to achieving sufficiently high test score reliabilities for research purposes. *The Journal of General Psychology*, *135*(3), 241-251.

Chronbach, L. J. (1990). *Essentials of psychological testing*. New York, NY: Harper & Row Publishers, Inc.

Cooper-Hakim, A., & Viswesvaran, C. (2002). A meta-analytic review of the MacAndrew alcoholism scale. *Educational and Psychological Measurement*, *62*(5), 818-829.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98-104.

Crocker, L. & Algina. J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace Jovanovich College Publishers.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Cronbach, L. J. (1988). Internal consistency of tests: Analyses old and new. *Psychometrika*, *53*, 63-70.

Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*(3), 391-418.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology*, *34,* 481-489.

Egger, M., Smith, G. D. & Phillips, A. N. (1997.) Meta-analysis: Principles and procedures. *British Medical Journal*, *315*(7121), 1533-1537.

Eide, P, et al. (2002). Test-retest reliability of the emotional stroop task: Examining the paradox of measurement change. *The Journal of Psychology*, *136*(5), 514-520.

Epstein, M. H., Ryser, G. & Pearson, N. (2002). Standardization of the behavioral and emotional rating scale: Factor structure, reliability, and criterion validity. *The Journal of Behavioral Health Services & Research*, *29*(2), 208-216.

Fan, X., & Yin, P. (2003). Examinee characteristics and score reliability: An empirical investigation. *Educational and Psychological Measurement*, *63*(3), 357-368.

Feldt, L. S. & Ankenmann, R. D. (1998). Appropriate sample size for comparing alpha reliabilities. *Applied Psychological Measurement*, *22*(2), 170-178.

Feldt, L. S. & Brennan, R . L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (*3rd Ed.*). Washington, D C: The American Council on Education and the National Council on Measurement in Education.

Feldt, L. S. & Charter, R. A. (2003). Estimation of internal consistency reliability when test parts vary in effective length. *Measurement and Evaluation in Counseling and Development*, *36*(1), 23-27.

Feldt, L. S., Woodruff, D. J. & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*(1), 93-103.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521-532.

Frisbie, D. & Druva, C. A. (1986). Estimating the reliability of multiple true-false tests. Journal of *Educational Measurement*, *23*(2), 99-105.

Gallagher, D., Nies, G., & Thompson, L. (1982). Reliability of the Beck depression inventory with older adults. *Journal of Counseling and Clinical Psychology*, *50*(1), 152-153.

Gaston, J. E. & Vogl, L. (2005). Psychometric properties of the general well-being index. *Quality of Life Research*, *14*, 71-75.

Ghiselli, E. E., Campbell, J. P. & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco, CA: W. H. Freeman & Company.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3-8.

Gliner, J. A., Morgan, G. A. & Harmon, R. J. (2001). Measurement reliability. *Journal of the American Academy of Child and Adolescent Psychiatry*, *40*(4), 486-488.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement*, *66*(6), 930-944.

Green, S. B. & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*(1), 121-135.

Hahn, G. J., Doganaksoy, N. & Meeker, W. Q. (1999). Reliability improvement. *Quality Progress*, *32*(5), 133-136.

Hambleton, R. K. & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*, 38-47.

Hambleton, R. K. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, *4*, 1-48.

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practice*, *10*(2), 181-189.

Harwell, M. R. (1992). Summarizing Monte Carlo results in methodological research. *Journal of Educational Statistics*, *17*(4), 297-313.

Headrick, T. C. & Sawilowsky, S. S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the Fleishman power method. *Psychometrika*, *64*(1), 25-35.

Heiser, W. J. (2006). Measurement without copper instruments and experiment without complete control. *Psychometrika*, *71*(3), 457-461.

Henson, R. K. (2004). Expanding reliability generalization: Confidence intervals and Charter's combined reliability coefficient. *Perceptual and Motor Skills*, *99*, 818-820.

Hinton, P. R. (2004). *Statistics explained* (*2nd Ed.*). New York, NY: Routledge.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (*8th Ed.*). Needham Heights, MA: Allyn & Bacon.

Howell, R. T. & Sheilds, A. L. (2008). The file drawer problem in reliability generalization: A strategy to compute a fail-safe N with reliability coefficients. *Educational and Psychological Measurement*, *68*(1), 120-128.

Howell, R. T. & Shields, A. L. (2008). The file drawer problem in reliability generalization: A strategy to compute a fail-safe N with reliability coefficients. *Educational and Psychological Measurement*, *68*(1), 120-128.

Huynh, H. (1986). Reliability of composite measurements based on the m highest of n equivalent components. *Journal of Educational Statistics*, *11*(3), 225-238.

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

Knapp, T. R. & Sawilowsky, S. S. (2001). Constructive criticisms of methodological and editorial practices. *The Journal of Experimental Education*, *70*(1), 94-95.

Knapp, T. R. & Sawilowsky, S. S. (2001). Strong arguments: Rejoinder to Thompson. *The Journal of Experimental Education*, *70*(1), 94-95.

Lautenschlager, G. J. (1989). ALPHATST: Testing for differences in values of coefficient alpha. *Applied Psychological Measurement*, *13*, 284.

Lee, G. & Frisbie, D. A. (2001). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education*, *12*(3), 237-255.

Lee, G. (2000). Estimating conditional standard errors of measurement for tests composed of testlets. *Applied Measurement in Education*, *13*(2), 161-180.

Lee, G. (2002). The influence of several factors on reliability for complex reading comprehension tests. *Journal of Educational Measurement*, *39*(2), 149-164.

Lee, G., Brennan, R. L. & Frisbie, D. A. (2000). Incorporating the testlet concept in test score analysis. *Educational Measurement: Issues and Practice*, *19*(4), 9-15.

Lee, G., Dunbar, S. B. & Frisbie, D. A. (2001). The relative appropriateness of eight measurement models for analyzing scores of tests composed of testlets. *Educational and Psychological Measurement*, *61*(6), 958-975.

Lee, W., Brennan, R. L. & Kolen, M. J. (2006). *Journal of Educational and Behavioral Statistics*, *31*(3), 261-281.

Li, H. (1997). A unifying expression for the maximal reliability of a linear composite. *Psychometrika*, *62*(2), 245-249.

Li, H., Rosenthal, R. & Rubin, D. B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods*, *1*(1), 98-107.

Livingston, S. A. (2004). An interesting problem in the estimation of scoring reliability. *Journal of Educational and Behavioral Statistics*, *29*(3), 333-341.

Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement*, *15*(4), 325-336.

Lord, F. M. (1960). An empirical study of the normality and independence of errors of measurement in test scores. *Psychometrika*, *25*, 91-104.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Menlo Park, CA: Addison-Wesley.

MacCaan, R. G. (2004). Reliability as a function of the number of item options derived from the "knowledge of random guessing" model. *Psychometrika*, *69*(1), 147-157.

Martinussen, M. & Bjørnstad, J. F. (1999). Meta-analysis calculations based on independent and nonindependent cases. *Educational and Psychological Measurement*, *59*(6), 928-950.

Maxwell, S. E., & Cole, D. A. (1995). Tips for writing (and reading) methodological articles. *Psychological Bulletin*, *118*(2), 193-198.

Mehrens, W. A. & Lehmann, I. J. (1987). *Using standardized tests in education*. NY: Longman.

Meier, S. T. & Davis, S. R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. *Journal of Counseling Psychology*, *37*(1), 113-115.

Mental Measurements Yearbook, 11th Ed. (1992). J. J. Kramer & J. C. Conoley (Eds.). Lincoln, NE: The University of Nebraska Press.

Metcalf, M., Reid, J., & Cohen, M. (2004). *Fortran 95/2003 explained*. New York, NY: Oxford University Press.

Metropolis, N. (1987). The beginning of the Monte Carlo method. *Los Alamos Science*, 125-130.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156-166.

Micceri, T. (1990). Proportions, pitfalls and pendulums. *Educational and Psychological Measurement*, *50*, 769-774.

Morrow, J. R. & Jackson, A. W. (1993). How "significant" is your reliability? *Research Quarterly for Exercise and Sport*, *64*(3), 352-355.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Research, 23*(2), 5-12.

Nevo, B. (1995). Examinee feedback questionnaire: Reliability and validity measures. *Educational and Psychological Measurement*, *55*, 499-504.

Novick, M. R. & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, *32*, 1-13.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3,* 1-13.

Nugent, W. R. (2009). Construct validity invariance and discrepancies in meta-analytic effect sizes based on different measures: A simulation study. *Educational and Psychological Measurement*, *69*(1), 62-78.

Nunnelly, J. C. (1978). *Psychometric theory* (*2nd Ed.*). New York, NY: McGraw-Hill Book Company.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, *5*(3), 343-355.

Paul, S. R. (1989). Test for the equality of several correlation coefficients. *The Canadian Journal of Statistics*, *17*(2), 217-227.

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

Ponterotto, J. G. & Ruckdeschel, D. E. (2007). An overview of coefficient α and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research methods. *Perceptual and Motor Skills*, *105*(3), 997-1014.

Popham, W. J. (2009). Unraveling reliability. *Educational Leadership*, *66*(5), 77-78.

Press, W. H., et al. (1996). *Numerical recipes in Fortran 90* (*2nd Ed*.). New York, NY: Press Syndicate of the University of Cambridge.

Qualls, A. L. & Moss, A. D. (1996). The degree of congruence between test standards and test documentation within journal publications. *Educational and Psychological Measurement*, *56*, 209-214.

Ramsey, J. & Gilbert, R. (1972). A monte carlo study of some small sample properties of tests for specification error. *Journal of the American Statistical Association*, *67*(337), 180-186.

Ramsey, P. H. (1994). Testing variances in psychological and educational research. *Journal of Educational Statistics*, *19*(1), 23-42.

Reckase, M. D. (1996). Test construction in the 1990s: Recent approaches every psychologist should know. *Psychological Assessment*, *8*(4), 354-359.

Reis, J. P., et al. (2005). Reliability and validity of the occupational physical activity questionnaire. *Medicine & Science in Sports & Exercise*, *39*(3), 416-425.

Revelle, W. & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the GLB: comments on Sijtsma. *Psychometrika*, *74*(1), 145-154.

Rexroade, K. R., Petersen, S. & O'Toole S. (2008). The ways of coping scale: A reliability generalization study. *Educational and Psychological Measurement*, *68*(2), 262-280.

Rodriguez, M. C. & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods*, *11*(3), 306-322.

Romano, J. L. & Kromrey, J. D. (2009). What are the consequences if the assumption of independent observations is violated in reliability generalization meta-analysis studies? *Educational and Psychological Measurement*, *69*(3), 404-428.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, *118*(2), 183-192.

Roth, L. & Sackett, P. R. (1991). Development and Monte Carlo evaluation of meta-analytic estimators for correlated data. *Psychological Bulletin*, *110*(2), 318-327.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement*, *18*(3), 229-244.

Samejima, F. (1997). Departure from normal assumptions: A promise for future psychometrics with substantive mathematical modeling. *Psychometrika*, *62*(4), 471-493.

Saris, W. E., Van Wijk, T. & Scherpenzeel, A. (1998). Validity and reliability of subjective social indicators: The effect of different measures of association. *Social Indicators Research*, *45*, 173-199.

Sawilowsky, S. S., Blair, R. C., & Micceri, T. (1990). A PC FORTRAN subroutine library of psychology and education datasets. *Psychometrika*, *55*(4), 729.

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin*, *111*(2), 352-360.

Sawilowsky, S. S. & Fahoome, G. (2003). *Statistics through Monte Carlo simulation with FORTRAN*. Oak Park, MI: JMASM.

Sawilowsky, S. S. (2000). Psychometrics versus datametrics: Comment on Vacha-Haase's "reliability generalization" method and some EPM editorial policies. *Educational and Psychological Measurement*, *60*(2), 157-173.

Sawilowsky, S. S. (2000). Reliability: Rejoinder to Thompson and Vacha-Haase. *Educational and Psychological Measurement*, *60*(2), 196-200.

Sawilowsky, S. S. (2003). Reliability as psychometrics versus datametrics. In Bruce Thompson *Score Reliability: Contemporary Thinking on Reliability Issues*, (103-122). Thousand Oaks, CA: Sage Publications, Inc.

Sawilowsky, S. S. (2003). You think you've got trivials? *Journal of Modern Applied Statistical Methods*, *2*(1), 218-225.

Sawilowsky, S. S. (2004). Teaching random assignment: Do you believe it works? *Journal of Modern Applied Statistical Methods*, *3*(1), 221-226.

Schmidt, F. L. & Le, H. A. (2006). An empirical calibration of the effects of multiple sources of measurement error on reliability estimates for individual differences measures. In Sawilowsky, S. S., *Real Data Analysis*, 287-292. Charlotte, NC: Information Age Publishing.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*(10), 1173-1181.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350-353.

Schumacker, R. E. & Smith, E. V. Jr. (2007). Reliability: A Rasch perspective. *Educational and Psychological Measurement*, *67*(3), 394-409.

Shiell, A., & Hawe, P. (2006). Test-retest reliability of willingness to pay. *European Journal of Health Economics*, *7*, 176-181.

Shipper, F. (1995). A study of psychometric properties of the managerial skill scales of the survey of management practices. *Educational and Psychological Measurement*, *55*, 68-79.

Sijtsma, K. (2009). Correcting fallacies in validity, reliability, and classification. *International Journal of Testing*, *9*, 167-194.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107-120.

Sijtsma, K. (2009). Reliability beyond theory and into practice. *Psychometrika, 74*(1), 169-173.

Sireci, S. G., Thissen, D. & Wainer, H. (1991). On the reliability of testlet based tests. *Journal of Educational Measurement*, *28*(3), 237-247.

Slaney, K. L., Tkatchouk, M., Gabriel, S. M., & Maraun, M. D. (2009). Psychometric assessment and reporting issues: Incongruence between theory and practice. *Journal of Psychoeducational Assessment*, *27*(6), 465-476.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677-680.

Stigler, S. M. (1977). Do robust estimators work with real data? *The Annals of Statistics*, *5*, 1055-1098.

Symonds, P. M. (1928). Factors influencing test reliability. *The Journal of Educational Psychology*, *XIX*(2), 73-87.

Teddlie, C., & Yu, F. (2007). Mixed methods sampling: A typology with examples. *Journal of Mixed Methods Research*, *1*(1), 77-100. Accessed August 31, 2010 from Sagepub.

Ten Berge, J. M. F. & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*(4), 613-625.

Thomas, J. C., & Traux, P. (2008). Assessment and analysis of clinically significant change. In D. McKay (Ed.), *Handbook of research methods in abnormal and clinical psychology*, 317-336. Thousand Oaks, CA: Sage Publications, Inc.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, *54*, 837-847.

Thompson, B. & Vacha-Hasse, T. (2003). Psychometrics is datametrics: The test is not reliable. In Bruce Thompson *Score Reliability: Contemporary Thinking on Reliability Issues*, (123-149). Thousand Oaks, CA: Sage Publications, Inc.

Thompson, B. (2003). Guidelines for authors reporting score reliability estimates. In Bruce Thompson *Score Reliability: Contemporary Thinking on Reliability Issues*, 91-102. Thousand Oaks, CA: Sage Publications, Inc.

Thompson, B. (2003). Understanding reliability and coefficient alpha, really. In Bruce Thompson *Score Reliability: Contemporary Thinking on Reliability Issues*, 3-23. Thousand Oaks, CA: Sage Publications, Inc.

Thye, S. R. (2000). Reliability in experimental sociology. *Social ForcesI 78*(4), 1277-1309.

Traub, R. E. & Rowley, G. L. (1981). Reliability of test scores and decisions. *Applied Psychological Measurement*, *4*(4), 517-545.

Traub, R. E. & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues and Practice*, *10*, 37-45.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, *16*(4), 8-14.

Traub, R. E. (1994). *Measurement methods for the social sciences – Reliability for the social sciences: Theory and application* (*Vol. 3*). Thousand Oaks, CA: Sage.

Troped, P. J., et al. (2007). Reliability and validity of YRBS physical activity items among middle school students. *Medicine & Science in Sports & Exercise*, *39*(3), 416-425.

Tzeng, O. C. S. & Welch, J. (1999). Inconsistencies across three contextual meanings of reliability. *Quality & Quantity*, *33*, 117-133.

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

Vacha-Haase, T. (2003). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. In Bruce Thompson *Score Reliability: Contemporary Thinking on Reliability Issues*, 203-218. Thousand Oaks, CA: Sage Publications, Inc.

Vacha-Haase, T., Henson, R.K., & Caruso, J. C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement*, *62*(4), 562-569.

Vacha-Haase, T., Kogan, L. R., Tani, C. R. & Woodallo, R. A. (2001). Reliability generalization: Exploring variation of reliability coefficients of MMPI clinical scales scores. *Educational and Psychological Measurement*, *61*(1), 45-59.

Vassar, M. & Crosby, J. W. (2008). A reliability generalization study of coefficient alpha for the UCLA loneliness state. *Journal of Personality Assessment*, *90*(6), 601-607.

Vassar, M. (2008). A note on the score reliability for the satisfaction with life scale: an RG study. *Social Indicators Research*, *86*, 47-57.

Verhelst, J. D. (1998). *Estimating the reliability of a test from a single test administration* (Measurement and Research Department Report 98-2). Arnhem, The Netherlands, CITO National Institute for Educational Measurement. Accessed from http://www.cito.com/research_and_development/pyschometrics/~/media/cito_com/research_and_development/publications/cito_report98_2.ashx

Vetterling, W. T., et. al. (1985). *Numerical recipes: Example book (FORTRAN)*. New York, NY: Press Syndicate of the University of Cambridge.

Viswesvaran, C. & Ones, D. S. (2003). Measurement error in 'big five factors' personality assessment: Reliability generalization across studies and measures. In Bruce Thompson *Score Reliability: Contemporary Thinking on Reliability Issues*, 245-258. Thousand Oaks, CA: Sage Publications, Inc.

Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185–201.

Wainer, H. & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement*, *27*, 1–14.

Wainer, H. & Lewis, C. (1991). Toward a psychometrics for testlets. *Journal of Educational Measurement*, *27*(1), 1-14.

Wainer, H. & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational and Psychological Measurement*, *57*, 741-758.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement*, *30*(1), 1-21.

Wainer, H., Sireci, S. G. & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement*, *28*(3), 197-219.

Wang, J. (2002). Reliability generalization: An HLM approach. *Journal of Instructional Psychology*, *29*(3), 213-218.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, *19*(1), 231-240.

Weller, L. D. Jr. (2001). Building validity and reliability into classroom tests. *National Association of Secondary School Principals, BASSP Bulletin*, *85*(622), 32-37.

Whitely, S. E. (1978). Individual inconsistency: Implications for test reliability and behavioral predictability. *Applied Psychological Measurement*, *13*(4), 571-579.

Wilcox, R. R. (1990). Comparing the variances of two dependent groups. *Journal of Educational Statistics*, *15*(3), 237-247.

Willson, V. L. (1980). Research techniques in *AERJ* articles: 1969-1978. *Educational Researcher*, *9*(6), 5-10.

Worthen, B. R., Borg, W. R., & White, K. R. (1993). *Measurement and evaluation in the school*. NY: Longman.

Woodhouse, G, et al. (1996). Adjusting for measurement error in mulitlevel analysis. *Journal of the Royal Statistical Society*, *159*(2), 201-212.

Wu, A. W., et al. (1997). Evidence for reliability, validity and usefulness of the medical outcomes study HIV health survey (MOS-HIV). *Quality of Life Research*, *6*, 481-493.

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck depression inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement*, *60*(2), 201-223.

Yuan, K. & Bentler, P. M. (2002). On robustness of the normal-theory based asymptotic distributions of three reliability coefficient estimates. *Psychometrika*, *67*(2), 251-259.

Zimmerman, Donald, W. & Williams, R. H. (1986). Note on the reliability of experimental measures and the power of significance tests. *Psychological Bulletin*, *100*(1), 123-124.

Zimmerman, Donald, W. (1994). A note on interpretation of formulas for the reliability of differences. *Journal of Educational Measurement*, *31*(2), 143-147.

Zimmerman, Donald, W., Williams, R. H. & Zumbo, B. D. (1993). Reliability of measurement and power of significance tests based on differences. *Applied Psychological Measurement*, *17*(1), 1-9.

Zinbarg, R. E., et al. (2005). Chronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123-133.

ABSTRACT

RELIABILITY GENERALIZATION: LAPSUS LINGUAE

by

JULIE M. SMITH

December 2011

Advisor: Dr. Shlomo S. Sawilowsky

Major: Education Evaluation and Research

Degree: Doctor of Philosophy

This study examines the proposed Reliability Generalization (RG) method for studying reliability. RG employs the application of meta-analytic techniques similar to those used in validity generalization studies to examine reliability coefficients. This study explains why RG does not provide a proper research method for the study of reliability, including describing how reliability is not a singular metric but a family of coefficients that are not interchangeable, along with other issues, such as sample and test administration. This research used Monte Carlo simulations designed to illustrate how the same instrument, administered repeatedly, can result in different reliability coefficients and to show that variation in reliability coefficients is due to sampling error; results illustrate that the reliability of a test will vary across test administrations based on the size and composition of the sample and how the sample was selected (randomly versus non-randomly).

AUTOBIOGRAPHICAL STATEMENT

JULIE M. SMITH
ax7955@wayne.edu

Education

Doctor of Philosophy – Educational Evaluation and Research
Wayne State University, School of Education, Detroit, Michigan
December 2011

Master of Public Health – Environmental and Industrial Health
University of Michigan, School of Public Health, Ann Arbor, Michigan
December 1996

Bachelor of Science – Physical Science (Major), Biology (Minor)
Michigan State University, College of Natural Science, East Lansing, Michigan
June 1991

Professional Experience

Wayne State University, Detroit Michigan
Adjunct Faculty, EER Department
Editorial Assistant, *Journal of Modern Applied Statistical Methods* (EER)
Accreditation Assistant (EDA, Counseling)
Grant Evaluation and Reporting Assistant (TED)

Cranbrook Institute of Science, Bloomfield Hills, Michigan
Public Program Presenter and Program Evaluator

Research Consultant, Various Companies
Quantitative and qualitative research design consulting
Data analysis, statistical testing, reporting

Certification

State of Michigan Secondary Provisional Teaching Certificate (DX, DA), Grades 7-12

Publications

Wahab, S., Baker, L., Smith, J. M., & Cooper, K. (2011). Exotic dance research: A review of the literature from 1970 to 2008. *Sexuality and Culture*, *15*(*1*), 56-79.

Smith, Julie M. (2009). Intermediate r Values for Use in the Fleishman Power Method. *Journal of Modern Applied Statistical Methods*, *8*(*2*), 610-612.

Smith, Julie M. (2008). Amyotrophic lateral sclerosis association. *Journal of Consumer Health on the Internet, 12*(*2*), 143-156.

Smith, Julie M. Nine nations, one Nile. (1996). *Population-Environment Dynamics: Ten Case Studies*. University of Michigan, EIH575. Monograph, Fall 1996. Ann Arbor, Michigan: Ann Arbor Press.

Awards

Graduate-Professional Scholarship, 2007-2008, 2008-2009, 2009-2010
Wayne State University Graduate School