AN EMPIRICAL COMPARISON OF ITEM RESPONSE THEORY AND CLASSICAL TEST THEORY ITEM/PERSON STATISTICS

A Dissertation

by

TROY GERARD COURVILLE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2004

Major Subject: Educational Psychology

AN EMPIRICAL COMPARISON OF ITEM RESPONSE THEORY AND CLASSICAL TEST THEORY ITEM/PERSON STATISTICS

A Dissertation

by

TROY GERARD COURVILLE

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Approved as to style and content by:

Bruce Thompson (Chair of Committee)	Victor L. Willson (Member)
John R. Hoyle (Member)	David J. Martin (Member)
Victor L. Willson (Head of Department)	

August 2004

Major Subject: Educational Psychology

ABSTRACT

An Empirical Comparison of Item Response Theory and Classical Test Theory Item/Person Statistics.

(August 2004)

Troy Gerard Courville, B.S., Louisiana State University-Shreveport;

M.S., Texas A&M University

Chair of Advisory Committee: Dr. Bruce Thompson

In the theory of measurement, there are two competing measurement frameworks, classical test theory and item response theory. The present study empirically examined, using large scale norm-referenced data, how the item and person statistics behaved under the two competing measurement frameworks. The study focused on two central themes: (1) How comparable are the item and person statistics derived from the item response and classical test framework? (2) How invariant are the item statistics from each measurement framework across examinee samples? The findings indicate that, in a variety of conditions, the two measurement frameworks produce similar item and person statistics. Furthermore, although proponents of item response theory have centered their arguments for its use on the property of invariance, classical test theory statistics, for this sample, are just as invariant.

DEDICATION

This dissertation is dedicated to God and his son, Jesus Christ, for being the leading force in my life and that of my family. Also this dissertation is dedicated to Jenny, who put up with me through this, Shane, who could not be here to see this and my kids, who by the grace of God could care less about all this.

ACKNOWLEDGMENTS

First, I would like to acknowledge the staff of Texas A&M University, College of Education, and especially Carol Wagner, for their full support and dedication. Secondly, I am eternally grateful to Dr. John Hoyle, Dr. Victor Willson, and Dr. David Martin, who each, in their own way, provided insight into turning theory into application.

Finally, I owe a heartfelt debt of gratitude to Dr. Bruce Thompson for his patience and ability to not only teach subject matter and produce great research but to care about his student. A rare breed indeed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLE Page

CHAPTER I

INTRODUCTION

Psychological research deals with complex structures that manifest their existence in various situations. Implicit in many situations is the understanding that a complex measurement framework must be employed to generalize beyond the single situation in which a measurement is observed. In psychology, we define the manifestation of structures as *responses*, while the structures are referred to as *constructs*. It is the relationship between the constructs and responses that is of special interest. To represent the relationship, models are developed. When a model is employed, constructs are rendered as latent variables and are expressed as measured response variables.

As models develop, they emerge into theories. As theories develop, divergence will often appear between the established theory and contemporary thinking. While this divergence may evolve into a dramatic alteration of the theory, this divergence, which at the time is portrayed as the bridge over a gulf in theoretical philosophy, can also be little more than a different way of viewing the previously defined theory.

This dissertation follows the style and format of *Educational and Psychological Measurement*.

Currently such a debate rages in the theory of measurement.

In the theory of measurement, there are two competing measurement frameworks, classical test theory and item response theory. It is in the statistical analyses underlying each theory that the differences are most evident.

Classical Test Theory

Classical test theory, just like item response theory, is an attempt to explain measurement error. In classical test theory, the model of measurement error is based on the correlation coefficient. The correlation coefficient, developed by Charles Spearman, attempts to explain error using two components: a true correlation and an observed correlation (Crocker & Algina, 1986; Traub, 1997).

The correlation coefficient, and classical test theory, is based on the theory that the average value of a measurement, taken over all possible measurements, will equal the true measurement in the population (Cochran, 1977). Implicit in the theory is 1) the error is random and 2) a single measurement is comprised of three components: an observed indicator, an hypothetical indicator that represents the true population value, and a hypothetical concept that represents the amount of disagreement between the true indicator and the observed indicator. Therefore, classical

test theory can be depicted as:

$$X = T + E.$$

This equation represents the three components as discussed above, with T being the hypothetical indicator/score, X the observed indicator/score, and E the amount of random disagreement between T and X. The equation can represent the amount of random error (E) as either an addition to or subtraction from the true score. As the random error (E) component approaches 0, the observed score (X) approaches the true scores (T).

Since its inception, classical test theory has been the dominate measurement model, having a significant impact on test-level and item-level information. In the collection of test-level information, one uses classical test theory with the hypothetical indicator (T) as the average score generated from the population of examinees. The observed indicator (X) is average score for the examinees who actually took the test. Reliability, test-level information concerning with consistency of scores across test administrations, is the correlation, or a reliability index, between the observed and true scores. However, because the true score is a hypothetical indicator and the observed indicator is an unbias estimator of the true score, the correlation between the observed scores on parallel tests can be used as an estimate of reliability, or a reliability coefficient.

However, one should note that a reliability coefficient is in a squared metric (i.e., r^2, R^2). Furthermore, reliability coefficients can be negative if 1) the two tests used to compute the reliability coefficient are not parallel or 2) a large amount of random error (E) which is usually the case with small samples and/or a small number of items (Thompson, 2002).

While classical test theory has been successfully applied to test-level information, the symbiotic relationship between reliability and item characteristics magnifies the role classical test theory plays in the development of item-level statistics (item *difficulty* and item *discrimination*).

Classical test theory is a simplistic model. Because of this, classical test theory invokes few assumptions thus allowing the theory to be applied to many testing situations. If a test is dichotomously scored, classical test item *difficulty*, *p*, is the proportion of the total examinees responding to an item correctly. Because, as Fan (1998) noted, *p* is an inverse indicator of item difficulty, as an increasing number of examinees incorrectly answer an item, the *p* value decreases.

Item discrimination statistics focus not on *how many* people correctly answer an item, but on whether the *correct* people get the item right or wrong. Although there are several methods used in classical test theory to assess item

discrimination, often discrimination is expressed as a point-biserial correlation between a dichotomously scored item and the scores on the total test.

Item Response Theory

Classical test theory does have its theoretical weaknesses. Fan (1998) summarized this problem with classical test theory estimators as involving circular dependency. Classical test statistics are *sample dependent* in that as the sample changes, the estimators would change (Cantrell, 1997; Henson, 1999). Therefore, the classical test theory estimators are not generalizable across populations. Because of the criticisms heaped upon classical test theory, many test developers have turned to item response theory.

Item response theory (IRT) is, for some researchers, the answer to the limitations of classical test theory. IRT is a modeling technique that tries to describe the relationship between an examinee's test performance and the latent trait underlying the performance (Cantrell, 1999; Hambleton & Swaminathan, 1985; Henard, 2000).

The most commonly used IRT models are built off a single ability parameter. The ability parameter, θ, is very similar to the classical test theory total-test true score. In fact, the relationship between the observed score and the ability parameter is the same relationship as the observed

score and true score:

$$T = \Sigma_g P_g(\theta), \text{ or}$$

$$X = \Sigma_g P_g(\theta) + E.$$

In contrast to classical test theory, item response models are lauded for their ability to generate invariant estimators. That is, theoretically IRT ability estimates, θ, are "item-free" (i.e., would not change if different items were used) and the item difficulty statistics are "person-free" (i.e., would not change if different persons were used). For single ability, dichotomously scored test items, IRT employs three different models.

A one-parameter model, the simplest of the three models, has the following function:

$$Pg(\theta) = e^{Da(\theta - bg)} / 1 + e^{Da(\theta - bg)}.$$

Looking at the *one-parameter* model, one can see its relationship between this model and the single parameter logistic regression. The value of *D* is usually set to 1.7 (Crocker & Algina, 1986). The *a* parameter is the item *discrimination* parameter with the *b* parameter being the item *difficulty* parameter. However, in the *one-parameter* model the item *discrimination* is assumed to be a constant. The one-parameter model is often called in the Rasch model (Crocker & Algina, 1986; Hambleton & Swaminathan, 1985; Henard, 2000).

The *two-parameter* model has the same function as presented for the one-parameter model. However, in the two-

parameter model, the item discrimination parameter will vary across items, as does the item difficulty parameter.

The *three-parameter* model, the most general model, includes a *psuedo-guessing* parameter especially useful for multiple-choice and true-false testing. In the one and two parameter, the lower asymptote moves toward the probability value of 0 rather quickly. This indicates that examinees in this area have a lower probability of achieving success on an item than they really have because they can "guess" the correct answer. The three-parameter model is expressed as follows:

$$Pg(\theta) = c_g + \{[(1- c_g)\ e^{Da(\theta - bg)}\ ]/\ 1 + e^{Da(\theta - bg)}\}.$$

Despite its assumed advances, item response models are subject to strict assumptions. Two major assumptions of item response theory are *unidimensionality* and *local independence*. Unidimensionality states that there is only one ability being measured. This assumption can never be strictly met. The assumption can be satisfied if a single dominant factor underlies responses. A second assumption is *local independence*, which necessitates that, excluding ability, there is no relationship between the test items and the examinee's responses. If these assumptions are met, an IRT model can be successfully employed.

Purpose of the Study

Over the past twenty-three years, since Lord's 1980's

book *Applications of Item Response Theory to Practical Testing Problems*, item response theory (IRT) has become the jewel of large-scale test construction programs. However, some investigations (Fan, 1998; MacDonald & Paunonen, 2002) have studied the empirical difference between these two models. Fan (1998) noted that "Because IRT differs considerably from CTT in theory, and commands some crucial theoretical advantages over CTT, it is reasonable to expect that there would be appreciable differences between the IRT- and CTT-based item person statistics" (p. 360).

However, articles by Fan (1998), Lawson (1991), MacDonald and Paunonen (2002), Skaggs and Lissitz (1986, 1988) and Stage (1998a, 1998b, 1999) have all pointed to little difference between item response estimates and classical test theory estimates. In Stage's (2000) work with the SweSAT test READ, she noted that "the agreement between results from item-analyses performed within the two different frameworks IRT and CTT was very good. It is difficult to find greater invariance or any other obvious advantages in the IRT based item indices" (pp. 19-20). Furthermore, Fan's (1998) research "failed to support the IRT framework for its ostensible superiority over CTT in producing invariant item statistics" (p. 378). MacDonald and Paunonen (2002) agreed with Fan and Stage, with an important caveat:

When the collection of potential test items in a

> pool possesses a narrow range of item difficulty values (common in personality and interest assessments), then item discrimination estimates should be largely accurate for both IRT and CTT measurement frameworks. In such a situation, item selection decisions based on either framework should result in the selection of roughly the same set of test items. On the other hand if the range of items difficulty statistics exceeds a narrow range of item difficulty values (about -0.5 to .5, common in achievement and ability tests), then the accuracy of item discrimination estimates begins to decrease with CTT methods. (p. 942)

However, findings of this type were first indicated by Nunnally in 1979 when he wrote that "when scores developed by ICC theory can be correlated with those obtained by the more usual approach to simply sum items scores, typically it is found that the two sets of scores correlated .90 or higher; thus it is really hair splitting to argue about any difference between the two approaches or any marked departure from linearity of the measurement obtained from the two approaches" (p. 224).

The present study is designed to replicate the work done by Fan (1998). As Fan (1998) noted, a principle limitation of his study was his use of criterion-referenced test and its inherent tendency toward items that have limited item difficulty ranges. This limitation is especially unsettling considering the results of MacDonald and Paunonen (2002). Considering the results of both Fan and MacDonald and Paunonen (2002), the present study consists of 80,000 examinees drawn from a population of 322,460 examinees who

took the written form of the ACT Assessment, a norm-referenced test. As MacDonald and Paunonen (2002) noted, typical IRT item difficulty values range achievement or ability from -0.5 to .5. For the present study, looking across all subtests, the IRT item difficulty ranges from -3.349 to 3.621.

The present study focused on two central themes: (1) How comparable are the item and person statistics derived from the item response and classical test framework? (2) How invariant are the item statistic from each measurement framework across examinee samples?

Specifically, this study addressed the same five research questions presented by Fan:

1. How comparable are the CTT-based and IRT-based examinee *ability* estimates?
2. How comparable are the CTT-based and IRT-based item *difficulty* estimates?
3. How comparable are the CTT-based and IRT-based item *discrimination* estimates?
4. When compared across different samples, how *invariant* are the CTT-based and IRT-based item *difficulty* estimates?
5. When compared across different samples, how *invariant* are the CTT-based and IRT-based item *discrimination* estimates?

Organization of the Study

The present student consists of three explanatory and three data-related chapters. Chapter II covers classical test theory, from its roots in measurement error and the correlation coefficient to classical test theory's use in reliability and item-level statistics. Chapter III covers item response theory and its basic concepts, use of the normal ogive, and general models employed in single ability, dichotomously-scored tests. Chapter IV develops the differences between the two models. Chapter V, the method section, covers the study's design. The final two chapters cover the data results and summary information.

CHAPTER II

CLASSICAL TEST THEORY

We know that constructs manifest themselves as responses. We also know that responses will change from situation to situation. Measurement is the quantification of the relationship between the responses and the constructs.

In measurement, a response can take several forms, from the analysis of written content to the counting of stimulus responses in a pavlovian experiment. A second component involves an underlying unobservable construct (Carmines & Zeller, 1979). For instance, the scores from an intelligence test represent an observed response, and the theory of intelligence upon which the test was derived is considered an underlying unobservable concept. Carmines and Zeller (1979) combine this into a definition of measurement:

> Measurement focuses on the crucial relationship between the empirically grounded indicator(s)--that is, the observable response-and the underlying unobservable concept(s). When this relationship is a strong one, analysis of empirical indicators can lead to useful inferences about the relationships among the underlying concepts. In this manner, social scientists can evaluate the empirical applicability of theoretical propositions. (p. 11)

Reliability and Validity

Carmines and Zeller's (1979) definition of measurement requires two useful concepts in the evaluation of a measurement: *reliability* and *validity*. Reliability focuses on

the empirical indicator's ability to consistently represent the underlying concept. For an empirical indicator to consistently represent the underlying concept, the empirical indicator must give consistent results. Thompson (2002) explained these considerations using the analogy of a bathroom scale:

> Some days when you step on your bathroom scale you may not be happy with the resulting score. On some of these occasions, you may decide to step off the scale and immediately step back on to obtain another estimate. If the second score is half a pound lighter, you may irrationally feel somewhat happier... But if your second weight measurement yields a score 25 pounds lighter than the initial measurement, rather than feeling happy, you may instead feel puzzled or perplexed. If you then measure your weight a third time, and the resulting score is 40 pounds heavier, you probably will question the integrity of all the scores produced by your scale. It has begun to appear that your scale is exclusively producing randomly fluctuating scores. In essence, your scale measures "nothing." (p. 4)

As Thompson (2002) noted, "When measurements yield scores measuring 'nothing,' the scores are said to be 'unreliable.'"

Validity is the degree to which an empirical indicator measures the intended underlying concept/theory, and only that construct. For an empirical indicator to represent the underlying concept the indicator must not only give consistent results, but the results must have a direct relationship to the underlying concept. The implications in the previous statement leads to an important conclusion about

the relationship between validity and reliability: Reliability is a necessary but not sufficient condition for validity. As Thompson (2002) explained

> Let's presume that upon repeated uses on a given morning your bathroom scale (to your possible disappointment) repeatedly yields the same estimate of your weight: 200 pounds. This evidence suggests that the scores may be reliable. However, if you inferred from your score(s), "Gosh, I must be brilliant, because an IQ of 200 is quite high," questions of score validity might arise! ...Scores can't both measure nothing and measure something. The only time that perfectly unreliable scores could conceivably be valid is if someone was designing a test intended consistently to measure nothing. But people do not ever design tests to measure nothing, because measurements of random fluctuations are already widely available in the form of dice and coin flips and other mechanisms. (p. 6)

Both reliability and validity are population specific. For instance, a commonly given intelligence test is the Wechsler Intelligence Test. The Wechsler has three different versions based on age level. If one were to give the Wechsler Intelligence Scale for Children-Revised to children, the scores would be reliable and valid, assuming the Wechsler accurately represents the theoretical concepts of intelligence.

Measurement Error

In their attempt to define measurement, Carmines and Zeller (1979) directly introduced the concept of measurement error. This definition has its geneses in a definition of

measurement by H.M. Blalock (1968) in which he invoked the notion of a gap between theory and research. Blalock defined the gap as measurement error.

Measurement error is the same error that is often discussed in covering topics such as structural equation modeling. Measurement error can be broken down into two different components, random error and non-random error.

Random error refers to a particular component of error that has no statistical predictability. Ott (1993) defined random error as the component that "takes into account all unpredictable and unknown factors that are not included in the model" (p. 440). As random error increases, reliability decreases. An example of random error's effect can be found in the arithmetic mean. It can be shown that the mean, taken over all possible samples, is an unbiased estimator of the true population value (cf. Cochran, 1977). The fact that a mean is an unbiased estimator of the true population value does not assure that a given mean is accurate. If one were to take a simple random sample of one out of an infinite sampling of means, there is a probability that the mean would not equal the population estimate. The difference between the sampled mean and the population mean is due to particular random error called "sampling error", which is different from measurement error. Random error, however, produces no systematic effects, which is why a mean can be an unbiased

estimator over a large number of samples.

Both random measurement error or unreliability and nonrandom error negatively affect validity. An example of systematic measurement error is a scale that consistently adds five pounds to each person's weight.

While every measurement is, in some way, hindered by error, the framework of measurement does have means of reducing its impact. In terms of validity, systematic error has been an issue discussed since the beginning of science. In fact, there seems to be a general consensus concerning procedures to evaluate and reduce its effect. However, methods of dealing with random measurement error are less settled. The discussion seems to have focused on two competing "theories": classical test theory and item response theory.

Classical Test Theory

Classical test theory began as an offspring of Charles Spearman's work on correlation coefficients. In his work, Spearman noted the existence of two types of correlations: a true correlation and an observed correlation (Crocker & Algina, 1986; Traub, 1997).

Spearman's views originates in the theory of unbiased measurement, which states that the average value of the measurement, taken over all possible measurements, will equal the true measurement in the population (Cochran, 1977).

The essence of unbiased measurement is that there are three components of any measure: an observed indicator, an hypothetical indicator that represents the true population value, and a hypothetical concept that represents the amount of disagreement between the true indicator and the observed indicator. Typically, the discussion of unbiased measurement centers on the following equation:

$$T = X - E.$$

This equation represents the three components as discussed above, with T being the hypothetical indicator, X the observed indicator, and E the amount of disagreement between T and X. Classical test theory is equivalently expressed in terms of the observed indicator:

$$X = T + E.$$

The equation might take three different forms. For example, if a student can truly spell 75 of 100 words on a spelling test, then the hypothetical score (T) is 75. However, perhaps the student actually spelled 85 words correctly because on the multiple choice spelling test he was able to guess the spelling of 10 words. Therefore, the model would be as follows: 85 = 75 + 10.

The equation also can represent the amount of random error (E) as either an addition to or subtraction from the true score. The above example represents a positive error component in that the student was able to spell 10 words by

means other than his natural ability. If the student misspelled 10 words because he was distracted by other students, the error component would be negative, and the equation would be expressed as follows: 65 = 75 - 10.

Our main concern in the measurement process is the congruence between the true score and the observed score. However, we do not ever know the hypothetical true score. Only when the observed score is not affected by random error can the hypothetical true score and the observed score be equal.

Observed Scores as Random Variables

If we use the classical test theory model, one can see that the observed scores change as the amount of random error changes. As noted earlier, as the random error component approaches 0, the observed score approaches the true scores. This indicates that random error has a negative effect on the congruence of the true scores and observed scores. However, equally important is that because the error component is random, the observed score is at least in part a random variable unless measurement error is zero.

In their 1986 book on test theory, Crocker and Algina defined a random variable as "a variable that assumes its values according to a set of probabilities" (p. 107). According to Crocker and Algina, there are two noteworthy points in defining a random variable, with the first being

the recognition of random dynamics. We must specify or hypothesize particular dynamics, because we could not take into account all the infinite on at least numerous random errors (such as inattention, loud talking, loud noises, guessing) that might affect the student's score. Once a student has taken the test, he has created a quantifiable estimate of his partially random observed score (X).

The amount of random error found in the student's score leads to the conclusion that his score is one of many possibilities, which together generate an underlying distribution (Crocker & Algina, 1986; Hinkle, Wiserma & Jurs, 1998). An underlying distribution is a hypothetical distribution based on all the possible outcomes of a given event (e.g., for this case, all the possible outcomes of the student's test score). The question is how a student can have more than one test score, with the answer lying in the randomness of the measurement error. As noted earlier, a true score is a *constant*. The student only knows so much about the civil war. Therefore, his true score is his exact ability on the civil war test at a given point in time. The student's observed score is partially *random* (e.g., does not always equal the true score) because measurement error (E) is random. As error becomes a larger component of the classical test equation, the observed score deviates further from the true score. If the civil war test was administered an

infinite number of times, we could get a frequency distribution that could be used to estimate the probability of a particular score, with the mean of the distribution being the student's true score.

Crocker and Algina (1986) stated, "the score of each examinee in a testing situation represents a different random variable. That is, the probability of obtaining a given test score is independently determined from a different distribution for each examinee" (p. 108). To explain this one must remember that the total error component is based on measurement errors that have affected the individual test takers differently. For example, although two students both may have been inattentive, the degree of inattentiveness will vary.

Classical Test Theory as Correlation

In Thompson's (1992) paper on linear regression analysis, he presented four types of linear regression analysis. The first type of linear regression analysis he presented involves only one predictor variable. This case is a bivariate correlation analysis. As mentioned earlier, correlation analysis was developed by Spearman and his contemporaries and has since become an extremely valuable statistical tool. However, it is the use of correlation in measurement context that is of particular interest for our purposes.

Because there is only one predictor variable in correlation analysis, the model would be as follows:

$$Y = X + E.$$

In this equation Y is a hypothetical variable that is defined by the observed variable X and the error component E. Although this is the traditional depiction of correlation analysis, there is no reason why the model could not be depicted as follows:

$$X = Y + E.$$

What is interesting about this depiction of correlation analysis is that the model is strikingly similar to the equation presented for classical test theory. Actually, reliability analysis is a correlational analysis. Dawson (1999) explained the linkages between classical test theory and the statistics general linear model (e.g., regression) in some detail. Using this understanding, one can turn to correlation analysis to better understand reliability.

Correlation is in part a function of the covariance between the hypothetical variable and the observed variable. Covariance is an unstandardized characterization of the amount of shared variance between two variables. Covariance can be depicted as:

$$Cov = [\Sigma(X - Xbar)(Y - Ybar)]/ n-1.$$

The indicates that as the joint deviations from the means increase, the covariance will increase (Henson, 2000). If we

factor out the standard deviation of each score, by dividing COV by ($SD_x[SD_y]$), the covariance is now a correlation coefficient, or a standardized covariance.

When recognizing the relationship between the squared correlation coefficient and reliability, the assumptions of correlation also apply for score reliability estimation. The assumptions for squared correlations are as follows:

- ✓ The population mean of the error component (E) in the population is zero.
- ✓ The correlation between hypothetical indicator (Y) and the error component (E) is zero.
- ✓ The correlations between the two variables' (Y and X) error components (E_x and E_y) are zero.
- ✓ The error scores are normally distributed.

Applied in a measurement context, the first assumption indicates that the population squared correlation coefficient or reliability coefficient, unlike a sample estimate, is not affected by random error. The second assumption indicates that there is no relationship between the hypothetical indicator (X) and the error component (E). The third assumption indicates that there is no relationship between T and X. The fourth assumption has an important impact on the understanding of the relationship between hypothetical indicators (T, Y) and observed scores. Because we expect the

error components to be normally distributed, the distribution of the observed indicator (X) will approximate the hypothetical indictor. Simply stated, the observed indicator (X) approximates the hypothetical indicator (T) (Nunnally, 1978).

There is only one confusing aspect of linking classical test theory and the statistics general linear model (e.g., regression) (Dawson, 1999). Reliability coefficients are always in a squared metric, just like r^2 or R^2 values. However, these coefficients do not have explicit superscripts or "2" (e.g., r_{XX} or α). Furthermore, reliability coefficients are often computed using the formula for the Pearson product-moment correlation coefficient (r), and not the formula for r^2.

As Lord and Novick (1968) explained,

> The square of the correlation between observed scores and true scores is equal to the correlation between parallel measurements. Thus, assuming at least one pair of parallel measurements can be obtained, we have succeeded in expressing an unobservable [universe] quantity ρ_{XT}^2 in terms of $\rho_{XX'}$, a parameter of a (bivariate) observed-score distribution. (pp. 58-59)

Thompson and Vacha-Haase (2000) explained,

> The variance-accounted-for universe reliability coefficient (ρ_{XX}, ρ_r in these various notations) is estimated by computing (or estimating) the unsquared correlation between scores on observed parallel tests, or on a single test administered twice... In other words, often the way we estimate score

> reliability is by computing unsquared r values. But by doing so, nevertheless what we are estimating is variance-accounted-for universe values (i.e., reliability coefficients). (p. 186)

Reliability Coefficient

Formerly, we discussed reliability in a singular context, as if we calculate the reliability of a single person. This is not the case. As Stanley (1971) stated

> the concept reliability coefficient is not applicable to a single individual but only to a group of persons, because that coefficient of correlation involves variation among the scores of different examines. That is, reliability coefficients are measures of interindividual differentiation, where as the variance error of measurement characterizes intraindividual variability for a particular trait, ability, or characteristic. (p. 373)

The magnitude of a reliability coefficient is an indicator of how much of the hypothetical indicator's variation is represented in the observed indicator variability. Therefore, when the variance of the observed indicator is composed primarily of error variance, the reliability coefficient will be near zero. However, as more of the observed indicator's variance becomes composed of hypothetical indicator variance, the reliability coefficient moves toward its upper bound of 1.0.

However, how does one move from an item context to a test context? Nunnally (1978) provided insight into the

topic, noting that

> the correlation of item 1 with the sum of an infinite number of items in a domain would equal the square root of the average correlation among items in the domain. The same could be proved for item 2 or item 3 or any other item. This holds only under the assumption that all items have the same average correlation with other items. (p. 197)

Because the true test score represents the average correlation among items in the domain, the correlation of item and total test scores approaches the correlation of the true score and the item (Nunnally, 1978). Therefore, if we have randomly sampled the items for the given test, "correlations among different tests would tend to be the same. Such randomly sampled collections of items are said to constitute randomly parallel tests, since their means, standard deviations, and correlations with true scores differ only by chance" (Nunnally, 1978, p. 198). Therefore, the reliability coefficient is found by correlating scores from two parallel tests.

For tests to be <u>strictly parallel</u>, all parallel measurements "can be shown to have equal means, equal standard deviations, and equal variances. Furthermore, when there are *k* parallel measurements, the correlation between any pair of these parallel measurements will be equal to the correlation between any other pair" (Crocker & Algina, 1986, p. 118).

Acknowledging the availability of parallel tests is an important concept in the discussion of reliability. Using this concept, we can ascertain the reliability of the scores by giving two different tests in the same population. Conversely, as tests depart from parallelism, the reliability coefficient will decrease (Nunnally, 1978).

Methods of Assessing Reliability

There are four basic forms of assessing test score reliability: (a) test/retest, (b) alternative form, (c) split-half, and (d) internal consistency. In test/retest analysis, the same test is given to same sample over a designated period. The scores from the different administrations are then correlated. However, there are two problems with the test/retest method of reliability assessment. The first problem is that even only one measurement has costs, and that those double with two administrations, given time. This is of greater concern the higher the testing cost is. In addition, if the population being sampled has a high mortality rate, the ability to assess stability reliability will decrease (Crocker & Algina, 1986). The second problem with the test/retest method is that it can cause what Stanley (1971) refers to as reactivity. Reactivity is the phenomenon whereby repeated testing itself causes a substantive change that would have not otherwise occurred. The primary example of reactivity found in testing

is the effect of memory. Memory of the first test administration may appreciably affect performance on the second administration. Because of these problems, alternative form reliability was developed.

Alternative form reliability is calculated by correlating scores from two different tests given to the same sample (Crocker & Algina, 1986). The alternative form method corrects for the reactivity problem found in the test/retest method. However, the alternative form method brings along its own set of problems. The main problem with this method is that there is no guarantee that each test is sampling the same content. Anytime we try to use two tests this problem will occur. This problem led to the development of a single test reliability coefficient.

Single test administration is a method of reliability estimations that uses one administration of a single test. This method is referred to as internal consistency (Crocker & Algina, 1996). There are two methods for performing this method: split-half and item covariances.

The *split-half* method calculation involves giving a test to the sample at the same time and then dividing the test into two parts and correlating the parts. This type of coefficient is called a coefficient of equivalence. However, the splitting of the test brings about a key problem in reliability. As the number of items on a test increase, the

reliability tends to increase. This is the case because as the covariation between items increases the amount of reliable variance increases. While a single test item that is not related to the concept will simply add variance, as an item becomes more related to the construct of interest, it adds both variance and covariance. Stanley (1971) noted, "test length is increased by the addition of parallel or approximate parallel components. True-score variance increases in proportion to the square of the inverse in test length, where as error variance increases only in proportion to the increase in test length" (p. 369). Therefore, as one subtracts items, the reliability tends to decrease. In response to this problem, the Spearman-Brown Formula (Brown, 1910; Spearman, 1910) was developed to estimate the reliability coefficient for the scores on the whole test by correcting for attenuation the correlations of the two halves. The Spearman-Brown formula is:

$$\rho_{xx'n} = 2\rho_{AB} \; / \; 1 + \rho_{AB}$$

The biggest problem with split-half reliability is that a test can be divided in numerous ways. Therefore, split-half reliability does not yield is not a unique estimate of reliability (Crocker & Algina, 1986).

The methods that have come to be depended on the most for reliability estimates are *item covariance* methods. The

main method is called coefficient alpha (Cronbach, 1951). The formula for coefficient alpha is:

$$\alpha = K/K\text{-}1\ [1-(\Sigma\sigma_I^2\ /\ \sigma_x^2)]$$

In the formula, *K* equals the number of items, σ_I^2 is the variance of scores on each item *i*, and σ_x^2 is the total test score variance. There are two important notes regarding the equation. The first note is that the summation of the individual item variances tend to be greater than the total test score variance. The second note is that as the numbers of items increase the *K/K-1* factor moves closer to 1, minimizing its impact. Coefficient alpha is a coefficient of precision but states nothing about stability or equilivance (Crocker & Algina, 1986).

The second analysis that falls into the item covariance analysis is the KR_{20} (Kuder & Richardson, 1937). The KR_{20} is coefficient alpha for dichotomized items:

$$KP_{20} = K/K\text{-}1\ [1-(\Sigma pq\ /\ \sigma_x^2)]$$

The *pq* component is the variance for a dichotomized individual item, because $\sigma^2_I = p_i q_i$.

Hoyt developed the final reliability coefficient in this category (Hoyt, 1941):

$$\rho_{xx'} = MS_{persons}\text{–}MS_{residual}/MS_{persons}$$

The mean square for the persons minus the mean square residual divided by the mean square persons. This equation

honors the very definition of reliability. The subtraction of mean square residual component is the subtraction of random measurement error. This leads to a component that is the amount of true variance. Hoyt's reliability coefficient, as are all reliability coefficients, is the proportion of true score variation found in the observed indicator variation (Crocker & Algina, 1986). Both the numerator variance and the denominator variance are in a squared metric, and thus so too the resulting reliability coefficient is in a square metric.

Each of the discussed methods is designed to assess reliability. However, just as important is the analysis of test content that explicitly deals with the differential functionality of an instrument's constituent items, which is called item analysis.

Item Analysis

In discussing test score reliability coefficients, it was noted that as the number of items increases, the test score reliability coefficients tends to increase. However, from a practical standpoint, there is some limit to a test taker's ability to answer an onslaught of test questions. Therefore, the goal, as in all scientific endeavors, is to measures the phenomenon with the fewest items that still allow for reasonably high test score reliability and validity. Item analysis is a set of statistical procedures that focus on the selection of items that maximizes score reliability.

Item Difficulty

In classical test theory, we define a true score as the average score of a person's distribution of infinity many possible scores. In essence, a true score is a person's true ability. For instance, for the civil war exam, a person's true score represents the amount of knowledge a person at a give time has about the civil war. Typically, the term "true score" is used in reference to a total test score. However, a "true score" could also represent a person's knowledge of a particular item. In the initial discussion concerning classical test theory, it was noted that classical test theory parallels the theory of unbiased estimation in statistics. This holds special importance when discussing the

classical test theory item *difficulty* parameter.

In the theory of unbiased estimation, the sample mean is an unbiased estimator of the population value μ. The average value of sample means, taken over all possible samples, represents the true population value (Cochran, 1977). In testing terms, a student's test score is equal to the true score as long as the average value of the student test scores, taken over infinitely many samples, equals the true score. As one may see from this definition, the one factor that allows an estimator to be unbiased is that the error score are uncorrelated, which is why this is an assumption in classical test theory statistics.

While some measurement instruments are scored in an interval format, such as continuous attitude rating scales, there are many instruments where a right/wrong scoring format is employed. In these cases, the distribution of item responses would form a binomial distribution. A binomial distribution is labeled such because there are exactly two outcomes represented. For a right/wrong scoring format, those two outcomes are usually represented with a 1 or 0. The binomial distribution represents these outcomes as statistically independent events. Furthermore, the binomial distribution is used to assign the probability of these outcomes occurring (Hinkle, Wiersma & Jurs, 1998). Thus, every item on a dichotomously-scored instrument generates a

binomial distribution that can be used to quantify the proportion of responses (the probability of a response occurring/100) answered correctly, which is represented in classical test theory as item difficulty.

Classical test theory item difficulty is termed an item statistic in that it represents an aspect of item functionality. Actually, item difficulty is a central tendency statistic. As noted earlier, a binomial distribution is generated from two independent events. For notational purposes, assume that each event is represented as A_0 or A_1, with A_1 representing the number of correct answers for an item in the population and A_0 as the number of incorrect answers in the population. If we want to know the proportion of items answered correctly in the population, the formula would be $P=A_1/N$, with N being the total number or responses. Cochran (1977) summarized this relationship when he noted that "the problem of estimating A and P can be regarded as that of estimating the total and mean of a population in which every y_i is either 1 or 0" (p. 51). References to the total and mean are, for test purposes, the total number of correct (or incorrect) responses and the item difficulty, respectively. When focusing on the item level, the mean of the population of items represents item difficulty, while at the test level the mean test score represents the test difficulty.

While measurement textbooks often discuss reliability as somewhat distinct from item analysis, item characteristics play a vital role in all the reliability coefficients. It has been noted that item analysis is used to help in the development of tests by maximizing the needed score reliability while minimizing the number of items. In terms of item difficulty, one could minimize the number of items by simply selecting items using an a priori level of difficulty. For instance, on our civil war exam, an a priori difficulty level of .70 (70% correct) might be selected. For the question: "What general in the civil war later became President of the United States?", the *p* was .50. Therefore, we would have to modify or eliminate this question. Utilizing item difficulty in this way is *not* appropriate because while we would minimize the number of questions, we would neglect the reliability of the scores. To minimize the number of questions and maximize score reliability, we must deal with the issue of variance.

The larger the number of items the higher is the score reliability is a general (not a universal) axiom. An unfavorable scenario occurs when a item is added and the reliability coefficient does not change. This would be the case if the item added zero correlation with the other items. If a new item is negatively correlated with many or all of the initial items, the reliability coefficient can actually

go down. At the extreme, the addition of such items can lead to negative reliability coefficients (Reinhardt, 1996; Thompson, 2002)! This is why scores on short forms of some published tests actually have higher reliability coefficients than the scores on the corresponding long forms (Vacha-Haase, 1998).

Looking at the equation for the item variance of dichotomously-scored items (i.e., $\sigma^2_I = p_i q_i$,), the function of item difficulty (*p*) in the equation indicates this factor's importance in an item's variance. Furthermore, because we are dealing with a binomial distribution, we can find *q* as *1-p*. Therefore, both of the components in the item variance, and by extension, the total test score variance, are representations of item difficulty. Table 1 shows item variances possible from the *pq* equation. One can see that the highest item variances are found when the item difficulty is around .50.

Table 1.

Possible Combination of Item Variances

p	*q*	Item variance
0.00	1.00	0.00
0.10	0.90	0.09
0.20	0.80	0.16
0.30	0.70	0.21
0.40	0.60	0.24
0.50	0.50	0.25
0.60	0.40	0.24
0.70	0.30	0.21
0.80	0.20	0.16
0.90	0.10	0.09
1.00	0.00	0.00

Given the influence item difficulty has in score reliability, one might think that most measurement instruments have an item difficulties close to .50. However, Crocker and Algina (1986) stated, "for most published aptitude and achievement tests designed for norm-referenced score interpretation, item difficulties typically fall in the range of .60 to .80. The reason for this lies in the item format commonly used in such tests" (p. 312). To explain this phenomenon, take the civil war test, and in particular, the question: "What general in the Civil War later became

President of the United States?". If the question format was open-ended, the responses would either be right or wrong, with the likelihood of someone guessing the correct answer would be remote or zero. However, if we were to change the format to multiple choice, then the probability of someone guessing correctly would increase. Under this scenario, the proportion of correct answers (*p*) would actually be comprised of those who knew the answer and *1/m*, with m being the number of response choices, reflecting the proportion who did not know the answer but who simply guessed correctly.

Because we want the item difficulty to optimize the item score variability, for selection-format tests, such as multiple-choice tests, we do not target *p*=.5 as the ideal item difficulty. We know that *1/m* of the proportion of correct answers were from guesses. Therefore, because the optimal level of item variation occurs at .50, we can simply take .50 and divide it by the number of choices for the item. For instance, if we had four choices on our civil war question, we would get .50/4 or .125. Therefore, we can expect 12.5% of the correct answers to our civil war question, assuming *p*=.50, to be guesses. Furthermore, our new item difficulty that would maximize the test/item true score variance can be found by the formula:

$$P_1 = .50 + .50/m.$$

Thus, for the civil war exam, our new optimal item difficulty

would be .625 (.50 + .125) (Crocker & Algina, 1986; French, 2001).

Item Discrimination

Inherent in the discussion concerning item difficulty is the creation of two groups. For item difficulty, we create a group that answered the item correctly and one that did not. Item discrimination statistics focus not on *how many* people correctly answer an item, but on whether the *correct* people get the item right or wrong. In essence, the goal of an item discrimination statistics is to eliminate items that do not function as expected in the tested group. One of the easiest item discrimination statistics to apply is the index of discrimination.

The index of discrimination is used with dichotomously-scored items. A criterion score, usually the total test score, is used to place test takers into an upper and lower group. Division of the test takers into these groups is one of a couple of issues leading to arguments against the index. A natural split would be to place 50% in the upper group and 50% in the lower group. However, it is easier for an item to discriminate between very high scores and very low scores on the criterion of interest. Kelly (1939) suggested that instead of a 50-50, split a 27-27 (omitting 46% of the data on a give item) split would allow the item discrimination statistic to function in a stable and useful manner. However,

others have found that as sample size increases, the group percentages can be gradually expanded with the statistic becoming just as stable and useful as using 27-27 splits (Crocker & Algina, 1996).

Once the group division is decided, the index of discrimination (*D*) can be calculated as : $D = p_u - p_l$, with p_u being the proportion of correct responses for the upper group and p_l being the proportion of correct responses for the lower group. Because a proportion ranges from 0 to 1, the index of discrimination can range from -1 to 1. A positive index indicates that a higher proportion of the upper group answered the item correctly, while a negative item discrimination index (*D*) indicates that a larger proportion of the lower group answered the item correctly.

As noted earlier, there is some subjectivity in the interpretation of *D*. While it is easy to see that a negative *D* is not a desirable result, it is not so clear how a *D* of .20, for example, compares to a *D* of .29. Crocker and Algina (1986) noted that *D* "has no well-known sampling distribution. It is not possible to answer questions such as what *D*-value is significantly greater than zero, or how large a difference between *D*- values is statistically significant" (p. 315). However, Ebel (1965) issued four guidelines to the interpretations for *D* values.

1. If $D \geq .40$: no item revision necessary;

2. If .30 ≤ *D* ≤ .39: little to no item revision is needed;
3. If .20 ≤ *D* ≤ .29: item revision is necessary; and
4. If *D* ≥ .19: either the item should be completely revised or eliminated.

In item discrimination, the key issue is how an item discriminates on a certain criterion. While the index of discrimination provides this information, it is problematic that the index ignores so much data. That is, *D* usually (a) omits the data of a lot of people (e.g., 46% of the respondents), and (b) ignores information regarding the exact scores of persons in the high group and persons in the low group. The product-moment correlation coefficient can be used when the criterion score (e.g., total test score) and the item scores (e.g., 0 or 1) are on an interval scale. However, the point biserial coefficient was created as a computationally friendlier version of the Pearson formula. The point biserial in calculated as $p_{bis} = [(\mu_t - \mu_x)/\sigma_x]*SQRT\ p/q$ with μ_t defined as the mean criterion score for the proportion answering the item correctly and μ_x defined as the entire criterion score mean.

A counterpart to the point biserial is the biserial correlation. In the biserial correlation, we assume that a normally distributed latent variable underlies the item and test performance. The biserial correlation is calculated by

using the formula: $\rho_{bis} = [(\mu t - \mu x)/\sigma x]\ (p/Y)$. The difference between the two formulas is found in the (p/Y). The *Y* is the "standard normal curve at the z-score associated with the *p* value for this item" (Crocker & Algina, 1996 p. 317). As is noted by Crocker and Algina (1996), the *Y* ordinate is always less than *SQRTpg*. Because the mathematical relationship between ρ_{bis} and ρ_{pbis} is $\rho_{bis} = (SQRT\ pq/Y) * \rho_{pbis}$, ρ_{pbis} will always be larger than the ρ_{bis}. Lord and Novick (1968) indicated that the difference would always be at least 20%. Furthermore, Crocker and Algina noted that "difference in magnitude remains fairly moderate for items of medium difficulty; however as p values drop below .25 or increase above .75, the difference between biserial and point biserial increases sharply" (p. 318).

A problem with the formulas for both the point biserial and biseral correlations is that an item contribution is weighted *twice*, once in the μ_t component, and once in the μx component. This can lead to correlations that are too high (in the case of very good items) or too low (in the case of bad items). Crocker and Algina (1986) noted that this problem is not as prevalent as the number of items increase. If the problem is recognized, a "corrected" discrimination coefficient can be computed by simply in turn eliminating the item scores being correlated in turn with the total scores on

the k-1 items.

Limitations to Classical Test Methods

While classical test statistics are still commonly used in test construction process, many researchers have questioned their utility in the modern era. Hambelton and Jones (1993) questioned the use of classical test theory estimators by saying that "classical item statistics such as item difficulty (i.e., proportion correct) and item discrimination (i.e., point biserial correlations) and test statistics such as test reliability are dependent on the examinee sample in which they are obtained" (p. 38). Fan (1998) summarized this problem with classical test theory estimators as involving circular dependency. Classical test statistics are sample dependent in that as the sample changes, the estimators would change (Cantrell, 1997; Henson, 1999). As MacDonald and Paunonen (2002) explained:

> examinee ability scores are dependent on the difficulty of test items. Thus, if the test is composed of relatively easy items, the person statistics (i.e., observed test scores) will be relatively high, giving the impression that the examinees possess high level of ability. If the test is composed of relatively difficulty items, however, the person statistics will be relatively low, giving the impression that the examinees possess low levels of ability. As such, estimates of examinee ability are dependent on the difficulty of the test items. (p. 922)

Therefore, the classical test theory estimators are not generalizable across populations.

Traub and Rowley (1991) wrote that classical test reliability is "an indicator of the quality of a set of test scores; hence, reliability is dependent on characteristics of the group of examinees who take the test, in addition to being dependent on characteristics of the test and the test administration" (p. 41). Another limitation of classical test theory is that to compare the performance of different examinees, the examinees must be given either the same or parallel items. The problem is further accented by a third limitation of classical test theory in that parallel forms are difficult to achieve. A fourth problem of classical test theory is "that it provides no basis for determining how an examinee might perform when confronted with a test item" (Hambelton & Swaminathan, 1985, p. 3). Finally, classical test theory assumes that the measurement error is the same for all examinees (Hambelton & Swaminathan, 1985). Because of the criticisms heaped upon classical test theory, some test developers have turned to item response theory.

CHAPTER III

ITEM RESPONSE THEORY

Item response theory (IRT) is, for some researchers, the answer to the limitations of classical test theory. Item response theory (IRT) looks at the examinee's performance by using item distributions based on the examinee's probability of success on a latent variable. In essence, IRT is a modeling technique that tries to describe the relationship between an examinee's test performance and the latent trait underlying the performance (Cantrell, 1999; Hambleton & Swaminathan, 1985; Henard, 2000).

Basic Concepts of IRT

In the earlier discussion of classical test theory, it was noted that there are two general factors in measurement, an observed response and an underlying unobservable construct. In classical test theory, we define this relationship as $X=T+E$. This is a *theoretical* model. In item response theory the models employed are *mathematical* functions. Thus, both models are fallible in that they are dependent on the assumptions a researcher is willing to posit with given data (Hambleton & Swaminathan, 1985).

Hambleton and Swaminathan (1985) summarized the characteristics of an item response model as involving four ideas. First, an IRT model must specify the relationship between the observed response and the underlying unobservable

construct. Secondly, the model must provide a way to estimate scores on the ability. Third, the examinee's scores will be the basis for the estimation of the underlying unobservable construct. Finally, an IRT model assumes that the performance of an examinee can be completely predicted or explained from one or more abilities.

The most commonly used IRT models are built off a single ability, a parameter in the model. The ability parameter, θ, is very similar to the classical test theory total-test true score. Crocker and Algina (1986) noted that "the relationship between T and θ is not statistical. The true score is a nonlinear transformation of the latent trait. The relationship between the observed score (X) and latent trait scores (θ) is statistical" (pp. 351-352). In fact, the relationship between the observed score and the ability parameter is the same relationship as the observed score and true score:

$$T = \Sigma_g P_g(\theta), \text{ or}$$

$$X = \Sigma_g P_g(\theta) + E.$$

The four characteristics of an item response model at first glance do not seem to set an IRT model appreciably apart from the classical test model. However, there are major differences between the two models. Item response models are lauded for their ability to generate invariant estimators. Theoretically, in item response theory, while

item parameter estimates (i.e., item difficulty and discrimination) are *not* dependent on the characteristics of the examinees, the ability estimates are also *not* dependent on the items. That is, theoretically IRT ability estimates, θ, are "item-free" (i.e., would not change if different items were used) and the item difficulty statistics are "person-free" (i.e., would not change if different persons were used). As noted earlier, a major criticism of the classical test theory is that estimators in that model may not have these proprieties. Also, each IRT ability estimate has a separate error estimate, while the classical test model assumes equal error variances across a measurement instrument.

Despite all its assumed advances, item response models are subject to strict assumptions. In classical test theory, the assumptions are relatively easy to meet because the assumptions do not have to be met exactly. Therefore, classical test theory is said to have weak assumptions.

Two major assumptions of item response theory are *unidimensionality* and *local independence*. Unidimensionality states that there is only one ability being measured. This assumption can never be strictly met. The assumption can be satisfied if a single dominant factor underlies responses. In our civil war exam, for our examinee to answer an item correctly he must know the particular element of history

under assessment. However, if the items required knowledge of the chronology of historical events, a new ability is introduced. A second assumption is local independence, which necessitates, that excluding ability, there is no relationship between the test item responses other than the relationship determined by ability or other model parameters. For example, if the responses to one item structurally constrain the possible answers to other items, then the items are not locally independent. If these assumptions are met, an IRT model can be successfully employed.

IRT Models

All IRT models are derived to generate item characteristic curves. An item characteristic curve plots the probability that an examinee will respond correctly to an item solely as a function of the test's latent trait (Crocker & Algina, 1986). Hambelton and Swaminathan (1985) noted that "The main difference to be found around currently popular item response models is in the mathematical form of $P_i(\theta)$, the ICC. It is up to the test developer or IRT user to choose one of the many mathematical functions to serve as the form of the ICCs" (p. 26).

The values on the X-axis of an ICC represent the latent trait, usually ranging from -3 to +3. The Y-axis represents the probability of an examinee's success. As the latent trait

increases, the probability of the examinee responding correctly will increase but with diminishing returns. In their discussion of item characteristic curves, Crocker and Algina (1996) discussed two interpretations that they consider acceptable. The first interpretation for a correct response is "the probability that a randomly chosen member of a homogeneous subpopulation will respond correctly to an item" (p. 341). A second interpretation is that the probability represents the probability of a specific examinee responding correctly for a subpopulation of items.

The first IRT model was built off a normal ogive, which is a standardized form of an ogive. An ogive, or a cumulative frequency polygon, "is the graph of a cumulative frequency distribution. It is useful for determining the various percentile points in a distribution of scores" (Hinkle, Wiersma and Jurs, 1998, p. 347). A normal ogive is a monotonically increasing curve, increasing from left to right. The normal ogive has a lower and upper asymptote, indicating that it will never equal 0 or 1 at any point. Figure 1 illustrates an ogive and Figure 2 illustrates a normal ogive.

Figure 1. Ogive

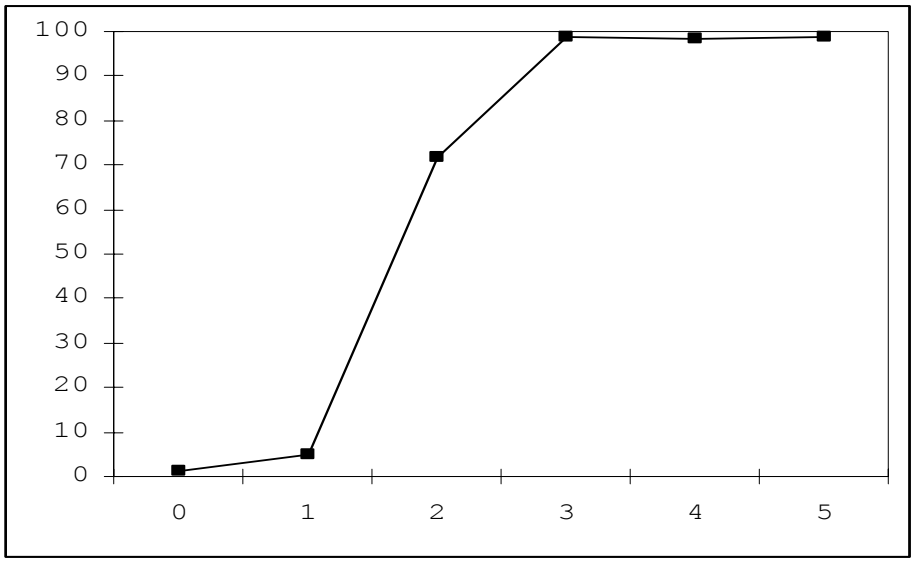
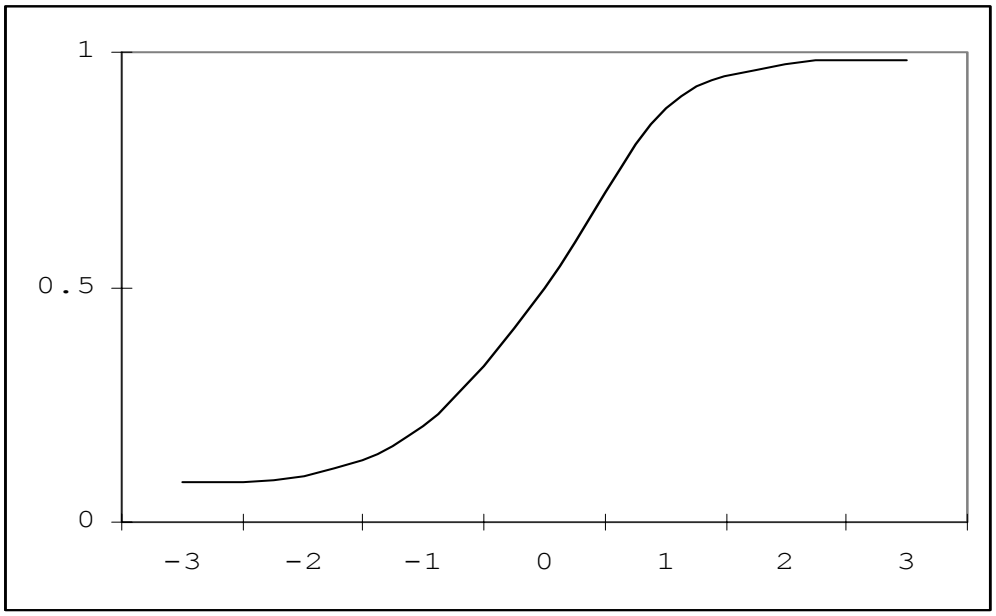


Figure 2. Normal Ogive



The focus of item response theory on item level information can be seen in the normal ogive equation. In latent trait models, the *b* parameter is the item difficulty parameter. The item difficulty parameter represents *the point on the ability scale θ, the horizontal axis, where there is a 50 percent probability the item is answered correctly.*

In recent times, the use of the normal ogive IRT modeling has all but vanished. While theoretically tantalizing, the normal ogive's calculations are tedious (Crocker & Algina, 1986). However, a complementary procedure to the normal ogive is found in the logistic item response models.

Logistic item response models are simply a form of logistic regression. The theory behind logistics regression is that when the dependent variable is a set of dichotomized scores, one can set the probability of a particular score. For a singe independent variable, one can write the probability equation as

$$P(\theta) = e^{B0 + B1\theta} / 1 + e^{B0 + B1\theta}.$$

Algebraically, the above equation can be expressed in the equation:

$$P(\theta) = 1 / 1 + e^{-(B0 + B1\theta)}.$$

Regardless of the equation used, the B_0 is the *Y* intercept and B_1 is the slope of the function produced by the mathematical relationship between the independent variable

and the dependent variable. *e* represents the base of the natural logarithm approximated at 2.178 (NORUSIS, 1990). Although the above equations represent one independent variable, the following equation represents multiple independent variables where θ equals $B_0 + B_1\theta_1 + \ldots + B_p\theta_p$:

$$P(\theta) = 1/1 + e^{-\theta}.$$

Using the logistic equations presented above, similar item characteristic curves can be developed for logistic models. The logistic curve has the same S shape curve as the normal ogive and the θ parameter ranges from -3 to +3. The relationship between the latent ability variable (θ) and the probability of a particular score is a nonlinear function. However, one should note that the top and bottom on the S shaped curve have an asymptotic relationship with the probability values of 0 and 1.

In item response theory, a one-parameter model will have the following function:

$$Pg(\theta) = e^{Da(\theta - bg)} /1 + e^{Da(\theta - bg)}.$$

Looking at the one-parameter model, one can see its relationship between this model and the single parameter logistic regression. The *D* in the one-parameter item response model represents a constant adjustment to the model to reduce the differences between the logistic IRT model and the normal ogive model to less than .01 (Crocker & Algina, 1986). The value of *D* is usually set to 1.7 (Crocker & Algina, 1986).

The *a* parameter is the item *discrimination* parameter with the *b* parameter being the item *difficulty* parameter. However, in the *one-parameter* model the item discrimination is assumed to be a constant. The one-parameter model is often called in the Rasch model (Crocker & Algina, 1986; Hambleton & Swaminathan, 1985; Henard, 2000).

The *two-parameter* model has the same function as presented for the one-parameter model. However, in the two-parameter model, the item discrimination parameter will vary across items, as does the item difficulty parameter.

The third model of interest in this research is the *three-parameter* model. The third parameter is a *psuedo-guessing* parameter especially useful for multiple choice and true-false testing. In the one and two parameter, the lower asymptote moves toward the probability value of 0 rather quickly. This indicates that examinees in this area have a lower probability of achieving success on an item than they really have because they can "guess" the correct answer. The three-parameter model is expressed as follows:

$$Pg(\theta) = c_g + \{[(1- c_g)\, e^{Da(\theta - bg)}]/\, 1 + e^{Da(\theta - bg)}\}.$$

In each of the above equations, the latent parameter θ was given considerable attention. This parameter is termed the ability parameter. The ability parameter has the same type of functional relationship with test items on a given test as does the dependent variable have with the independent variable(s) in linear regression. Both the dependent variable and the latent variable represent a concept that inconceivable without the use of independent variables or test items. Latent ability can represent anything that the test represents. In fact, the latent ability parameter is "dependent" on the test items to define its numerical functionality. Item response theory, like any other form of reliability evaluation, cannot prove the viability of this relationship but must depend on construct validation.

Unlike classical test theory, item response theory uses a maximum likelihood statistical theory to estimate the ability parameter. This estimator, as Hambleton and Swaminathan (1985) noted, can be interpreted as the "value of the examinee's ability that generates the greatest 'probability' for the observed pattern" (p. 77). Hambelton and Swaminathan's use of an "observed pattern" indicates that item response theory is a modeling technique that

requires some subjectivity. Hambleton and Swaminathan also noted that the maximum likelihood estimates for students who have prefect scores on the test or get nothing correct are not estimated well.

For test takers with zero correct responses, we can estimate their *maximum* possible ability. But we have no way to determine how much lower their abilities are, unless we give these persons a series of easier items to find the boundaries of their abilities. There are *infinitely many* reasonable ability estimates below the maximum ability for these test takers. The converse situation arises for persons with all items correct. We can estimate their *minimum* ability, but there are infinitely many reasonable estimates above this maximum.

CHAPTER IV

ITEM RESPONSE THEORY VS CLASSICAL TEST THEORY

When the measurement community turned to item response theory, the item response model was heralded as "one of the more important methodological advances in psychological measurement in the past half-century" (McKinley & Mills, 1989, p. 71). In fact, nine years earlier, Lord (1980) noted "nothing in this book will contradict either the assumptions or the basic conclusions of classical test theory. Additional assumptions will be made; these will allow us to answer questions that classical test theory cannot answer. Although we will supplement rather than contradict classical test theory, it is surprising how little we will use classical theory explicitly" (p. 7).

In analyzing the differences between item response theory and classical test theory, Embretson and Reise (2000) listed ten rules of measurement that will change due to the item response movement. Four are of particular interest. The first rule change is that the standard error of measurement which applies to all scores in a particular population in classical test theory would apply differently across response patterns but would generalize across populations. In classical test theory, the raw score transformation is linear (i.e., $T = X + E$). Thus, we must assume that the variances are approximately equal. Because the variances are assumed to

be approximately equal, an assumption known as homogeneity of variance, measurement errors for individual scores are assumed to be distributed normally and equally at each score level. In other words, standard errors of measurement become *conditional* on ability levels. Conversely, IRT modeling is a nonlinear modeling technique, which does not require the homogeneity assumption. Consequently, standard errors of measurement in the IRT framework can differ across score levels. However, Embretson and Reise (2000) noted that standard errors for the IRT framework can be averaged to provide a generalized standard error of measurement for the population.

A second rule change is that longer tests would no longer mean better reliability. The larger the number of items the higher the reliability is a general (not universal) axiom in classical test theory. The IRT framework can achieve maximum reliability with fewer items because ability score estimates, being item-free, can be based on giving different items to different test takers, and matching item difficulties to person abilities so as to obtain the most information form the fewest items.

A third change is that one no longer needs to have parallelism for test equating. Comparing different test forms often requires some form of equating to enable score compatibility (Embretson & Reise, 2000). In CTT, the

prevalent equating methods all require parallelism to some extent. As noted earlier, parallelism is rarely met. Embretson and Reise (2000) noted that "various equating methods can be applied when the test forms have different means, variances, and reliabilities; equating error is influenced by differences between the test forms. Equating error is especially influenced by differences in test difficulty length" (p. 21). When tests have varying difficulties, linear equating methods underestimate some test scores while overestimating others. IRT, a nonlinear method of equating, was shown by Embretson and Reise (2000) to be a better equating method when equating tests.

A fourth noted change is depicted in the ability of the IRT framework to generate item parameter estimates that are unbiased even across different samples. The ability of the IRT framework to generate unbiased item parameter estimates is termed invariance. Hambleton, Swaminathan and Rogers (1991) noted:

> The property of invariance of item and ability parameters is the corner stone of IRT and its major distinction from classical test theory. This property implies that the parameters that characterize an item do not depend on the ability distribution of the examinees and the parameter that characterize and examinee does not depend on the set of items. (p. 18)

The property of invariance has been considered an accepted benefit of the IRT framework (Drasgow & Parsons,

1983; Embretson, 1999; Embretson & Reise, 2000; Fan, 1998; Hambleton, Swaminathan & Rogers, 1991; MacDonald & Paunonen, 2002).

Looking at Embretson's (2000) rules, one might formulate the notion that item response theory produces tests that have vastly improved ability estimates. Fan (1998) noted that "Because IRT differs considerably from CTT in theory, and commands some crucial theoretical advantages over CTT, it is reasonable to expect that there would be appreciable differences between the IRT- and CTT-based item person statistics" (p. 360).

However, articles by Fan (1998), Lawson (1991), MacDonald and Paunonen (2002), Skaggs and Lissitz (1986, 1988) and Stage (1998a, 1998b, 1999) have all pointed to little difference between item response estimates and classical test theory estimates. In Stage's (2000) work with the SweSAT test READ, she noted that "the agreement between results from item-analyses performed within the two different frameworks IRT and CTT was very good. It is difficult to find greater invariance or any other obvious advantages in the IRT based item indices" (pp. 19-20). Furthermore, Fan's (1998) research "failed to support the IRT framework for its ostensible superiority over CTT in producing invariant item statistics" (p. 378). MacDonald and Paunonen (2002) agreed with Fan and Stage, with an important caveat:

> When the collection of potential test items in a pool possesses a narrow range of item difficulty values (common in personality and interest assessments), then item discrimination estimates should be largely accurate for both IRT and CTT measurement frameworks. In such a situation, item selection decisions based on either framework should result in the selection of roughly the same set of test items. On the other hand if the range of items difficulty statistics exceeds a narrow range of item difficulty values (about -0.5 to .5, common in achievement and ability tests), then the accuracy of item discrimination estimates begins to decrease with CTT methods. (p. 942)

However, findings of this type were first indicated by Nunnally in 1979 when he wrote that "when scores developed by ICC theory can be correlated with those obtained by the more usual approach to simply sum items scores, typically it is found that the two sets of scores correlated .90 or higher; thus it is really *hair splitting* [italics added] to argue about any difference between the two approaches or any marked departure from linearity of the measurement obtained from the two approaches" (p. 224).

The present study is designed to replicate the work done by Fan (1998). However, as Fan (1998) noted, a principle limitation of his study is the use of criterion-referenced test and its inherent tendency toward items that have limited item difficulty ranges. This limitation is especially unsettling considering the results of MacDonald and Paunonen (2002). Considering the results of both Fan and MacDonald and Paunonen (2002), the present study consists of 80,000 examinees drawn from a population of 322,460 examinees who took the written form of the ACT Assessment, a norm-referenced test. As MacDonald and Paunonen (2002) noted, typical item difficulty values range achievement or ability from -0.5 to .5. For the present study, looking across all subtests, the item difficulty ranges from -3.349 to 3.621.

The present study focused on two central themes: (1) How comparable are the item and person statistics derived from the item response and classical test framework? (2) How invariant are the item statistic from each measurement framework across examinee samples?

Specifically, this study addressed the five research questions presented by Fan:

1. How comparable are the CTT-based and IRT-based examinee *ability* estimates?
2. How comparable are the CTT-based and IRT-based item *difficulty* estimates?
3. How comparable are the CTT-based and IRT-based item *discrimination* estimates?
4. When compared across different samples, how *invariant* are the CTT-based and IRT-based item *difficulty* estimates?
5. When compared across different samples, how *invariant* are the CTT-based and IRT-based item *discrimination* estimates?

CHAPTER V

METHOD

Chapter V, the method section, covers the study's design. The data source section introduces information on the data used, the instrument, and any constraints to the data or instruments. The participant sampling section covers the three participant sampling plans employed to study the behavior of the examinee's scores under the CTT and IRT measurement frameworks. A discussion of the comparability and invariance of IRT and CTT item statistics and a correction bias in sample correlation coefficient (employed for small samples) is also included.

Data Source

The data used in this study are from the ACT Assessment Test. The ACT is typically taken by college-bound students in the eleventh and twelfth-grades. The ACT is taken by over one million students each year. Nearly 3,000 postsecondary institutions require or recommend that applicants submit ACT results. The ACT Assessment is given via written and computer adaptive testing formats. For the present study, only examinees (a) given the written format (b) in the same ACT administration were considered.

The ACT Assessment is composed of four tests: English, Mathematics, Reading, and Science. The English Test is a composed of 75 four-option multiple-choice items with a 45

minute time limit. The test is designed to measure the examinee's understanding of the conventions of standard written English and rhetorical skills.

The Mathematics Test is composed of 60 four-option multiple-choice items, with a 60 minute time limit. The test is designed to assess the mathematical reasoning skills typically acquired in math courses such as pre-algebra, algebra/elementary algebra, intermediate algebra/coordinate geometry, and plane geometry/trigonometry.

The Reading Test is composed of 40 four-option multiple-choice items, with a 35-minute time limit. The test is designed to measure reading comprehension as defined in skills of referring and reasoning.

The Science Test is composed of 40 four-option multiple-choice items, with a 35-minute time limit. The test is designed to measure the interpretation, analysis, evaluation, reasoning, and problem solving sills in the natural sciences.

For the present study, a sample of 80,000 examinees was randomly drawn from an examinee population of the specified administration consisting of 322,460 test takers. The sample of 80,000 was composed of 40,000 males and 40,000 females. The male and female examinees were further subdivided into *mutually exclusive* subsamples for each test. Therefore, each test sample is comprised of mutually exclusive subsamples of

10,000 males and 10,000 females. The random sampling was compiled by the ACT Corporation.

To facilitate comparisons across the four tests, a random sampling of items was conducted to restrict the longer English and Math tests to 40 items. The 40 items were randomly selected.

Participant Sampling

To replicate Fan's (1998) article, three sampling plans were employed to study the behavior of the examinee's scores under the CTT and IRT measurement frameworks. Each sampling plan was employed for each test. The sampling plans allow for the comparability of each framework across progressively less comparable samples.

According to Chang, Hanson and Harris (2001), stable estimates of CTT item difficulty and discrimination can be found with a sample size of 150 to 200. Wright and Stone (1979) found that sufficient sample sizes for CTT stability would allow for stable estimates of one-parameter IRT item indices. To investigate the functionality of CTT and IRT estimates under different conditions, two different sample size conditions was employed. To replicate functionality of the two measurement theories in large scale measurement situations, one set of samples were randomly selected with n=1,000. Conversely, to replicate clinical situations were tests are often constructed with small sample sizes, a second

set of samples were randomly selected n=100 (Skaggs & Lissitz, 1986).

One set of 400 random samples, each consisting of 1,000 examinees, were drawn from the 80,000 examinees. The 400 random samples represent 100 random samples for each of the four tests.

A second set of random samples was drawn to look at the effect of small samples. Eight hundred random samples, each consisting of 100 examinees, were drawn from the 80,000 examinees. That is, 100 random samples of n=1,000 were drawn for each of the four tests.

For gender, 100 random samples of each gender group were drawn from the four tests, equaling 1,600 gender samples (4 subtests * 100 random samples * 2 gender groups * 2 different sample size conditions = 1,600). The same process was employed to generate the small sample replicates. As Fan (1989) noted, because the gender samples are subpopulations of the total population, theoretically, disparity between statistics calculated from different samples will be *larger* than that found in random sampling plan.

A third sampling involved truncated high-ability and low ability group samples. For this sampling plan, there were 1,600 samples. The low-ability sample was comprised of students whose total test score fell in the 0 to 40th percentile range while the high-ability group fell in the

60^{th} to 100^{th} percentile range. One-hundred samples were randomly drawn from both the low and high ability group for each test. These truncated high-ability and low-ability group samples should theoretically display the *greatest* dissimilarity between the CTT and IRT statistics, because "these two groups were defined in terms of test performance, not in terms of a demographic variable" (Fan, 1998, p. 363).

Comparability of IRT and CTT Statistics

Person Statistics

Correlating the two parameters will assess the comparability of the IRT ability score and the CTT estimated true score. The ability parameter was assessed using Bilog-MG's marginal-maximum likelihood method (Windows Version 3.0.2327.2 for one-, two-, and three parameter IRT models). The CTT true score estimate is the obtained total number of right answers. For each sampling plan, both the CTT- and IRT-based (one-, two- and three-parameter) ability estimates were obtained. Therefore, each sample generated three correlation coefficients: CTT-based ability estimate with the IRT-based ability estimates for the one- two- and three-parameter models.

Two Item Statistics

The compatibility of item statistics for both methods was obtained by correlating (a) the item difficulty and (b) the item discrimination methods. The IRT item *difficulty* parameter, denoted by b in IRT models but referred to as the threshold parameter in Bilog-MG, was compared to the item difficulty (*p)* value generated using the CTT technique. The CTT item *discrimination* technique, the corrected item-test point bi-serial correlation, was compared to the IRT item slope parameter, a.

Degree of Invariance between IRT and CTT

As noted by Fan (1998), the three sample techniques employed here will generate *progressively dissimilar* samples, when looking across the three sample techniques. The three sampling frames used to evaluate invariance were: (a) random samples, (b) gender group sampling, and (c) truncated high-ability and low-ability group samples. By correlating the item parameters from different samples, within the same sampling plan, within the same measurement framework (i.e., IRT to IRT, CTT to CTT), the degree of estimated invariance, a commonly cited advantage of IRT, was evaluated.

Transformations for CTT *P* Value and Item-Test Correlations

In CTT, the item difficulty statistic is expressed on an ordinal scale. In an ordinal measurement scale, one is able to discern whether one item is more difficult than other item. However, it can not tell us whether the differences in various item difficulties are the same across the different comparisons. For instance, if items 1, 2 and 3 have an item difficulty of .25, .20, and .15, just because the difference between 1 and 2 and 2 and 3 equals .05 does not indicate that the difference in difficulty is the same in these two comparisons.

However, if the trait being measured is normally distributed, the CTT item difficulty statistic can be expressed as equal interval normal curve units (Anastasi, 1988). The transformation is achieved by finding the *z* score that corresponds to the proportion of examinees who answer an item correctly. The present study correlated *both* the CTT item difficulty (*p*) and the normalized CTT item difficulty statistics with IRT item difficulty estimates.

An item-test point bi-serial correlation, identified as the CTT item discrimination statistic, is not linearly scaled. As Hinkle, Wiserma and Jurs (1998) explained, "the sampling distribution of the correlation coefficient changes its shape as a function of both the magnitude and the sign of

the coefficients" (p. 231). R.A. Fisher developed a transformation that in large samples allows the transformed correlation coefficient to be distributed approximately normal (Hinkle, Wiserma & Jurs, 1998). Therefore, the assessment of the invariance of CTT item discrimination statistic is based on the correlation analysis between *both* the original and the Fisher z transformed point bi-serial for the differing samples of examinees. For each test, an average correlation coefficient was obtained by (a) transforming the individual correlation coefficients to Fisher *Zs*, (b) averaging the Fisher *Zs*, and (c) transforming the average Fisher *Zs* back to correlation coefficients (Fan, 1998).

Correcting for the Bias in Sample Correlation Coefficients

Because the sample correlation coefficient, *r*, is a ratio, it is a biased estimator of the population correlation coefficient. Zimmerman, Zumbo, and Williams (2003) noted that *r* can be biased as much as .03 or .04, which, as Zimmerman et al. indicated, may be vital when investigating the accuracy of the magnitude of *r* in measurement studies.

To correct for the bias in the sample correlation coefficient, R.A. Fisher developed a procedure to approximate the population correlation coefficient:

$$E[r] = r[1+\{(1-r^2)/2n\}]$$

Later, Olkin and Pratt (1958) indicated that the following approximation is a more nearly unbiased estimator of *r*:

$$E[r] = r[1+\{(1-r^2)/2(n-3\}]$$

The bias is greatest in the .500/-.500 range and decreases as the sample correlation coefficient moves out of this range. As the sample size decreases, the effect of bias increases. The present study used *both* the Fisher and the Olkin and Pratt corrections to compare model parameters across CTT and IRT procedures.

CHAPTER VI

RESULTS AND DISCUSSION

Chapter VI covers the results and discussion. The assessment of model-data fit is discussed. The results for both different sample size conditions are discussed under each research question in the first section.

IRT Assessment of Model-Data Fit

Every statistical model requires assumptions about the data to obtain viable parameter estimates. In some instances, such as classical test theory, these assumptions are *weak*, meaning that most data will be able to meet these assumptions. Conversely, IRT models have *strong* assumptions. In fact, Hambleton and Swaminathan (1985) concluded that IRT assumptions are so strong that no data set will ever be able to meet fully the assumptions. The violation of IRT assumptions can not only eliminate the possible advantages of test score interpretation (Hambleton & Swaminthan, 1985) but will lead to erroneous or unstable IRT estimates.

Because all IRT models require a unidimensional latent space, unidimensionality is viewed as the most important IRT model assumption. Hambelton and Swaminthan (1985) noted four different approaches to assessing unidimensionality. For the present study, factor analysis was used to assess the unidimensionality, using the eigenvalues to identify the number of dominate factors that exist among the test items.

For the English and Math test, the analysis was conducted using the same 40 test items used in subsequent analysis. The *population data* consisting of 80,000 cases (20,000 cases for each subtest) was the basis for this data. For the English test, the top three eigenvalues were 6.55, 1.34, and 1.24. For the 40 math items, the top three eigenvalues were 7.30, 1.69, and 1.26. The first three eigenvalues for the Reading items were 6.85, 1.90, and 1.20. Finally, the first three eigenvalues for the Science items were 5.79, 1.63, and 1.20. Based on these results, the unidimensionality assumption appeared to hold for the data. This result was expected considering the amount of measurement research that has been done in developing and maintaining the ACT Assessment.

To assess overall model-data fit, individual item misfit was assessed using *population data* consisting of 80,000 cases (20,000 cases for each subtest) and the subtest items used in the subsequent analysis. In BILOG-MG (Windows Version 3.0.2327.2), a like-hood ratio chi-square test is supplied. Tests of statistical significance, like the likelihood-ratio chi-square, are heavily influenced by sample size. For this analysis two corrections were enlisted to restrict the possibility of misinterpretation due to sample size was used. Table 2 summarized the number of items, the number of misfitting items and the percentage of items that were misfitting.

Table 2.

Number of Misfitting Items (α = .01)

Test	Items	1P	%	2P	%	3P	%
English	40	6	15	0	0	0	0
Math	40	22	55	6	15	4	10
Reading	40	8	20	1	3	6	15
Science	40	14	35	2	5	1	3

Note: 40 items were randomly sampled from the larger items pools for the Math and English tests. "1P" = one-parameter IRT model; "2P" = two-parameter IRT model; "3P" = three-parameter IRT model

The English test items fit the best of all the tests. Only six of the forty items (15%) were designated as misfitting items in the one-parameter model and none were labeled as such in the two and three-parameter models. Conversely, fifty-five percent of the Math test items were identified as misfitting in the one-parameter model, with a considerable decrease in model-data misfit when the two and three-parameter models were employed. A curious result was found in the comparison of Reading test results across the three models. While the results indicated a small number of misfitting items, the three-parameter model had more misfitting items than the two-parameter model. This seems to

run contrary to the thought that as the number of explanatory variables increases (in this case the inclusion of a pseudo-guessing parameter) the expanded model will fit at least as well as the previous model. However, here lies one of the quintessential problems with interpretation based on statistical significance. In the two-parameter model, several items had probabilities ranging in the .02 to low .01 range, indicating that these items could have easily been classified as misfitting items if a different alpha were chosen, the sample size increased slightly, or the sample dynamics were changed slightly. Therefore, the question is whether or not these items are, indeed, misfitting items.

Overall, the two-parameter and three-parameter models seemed to fit well across tests while the one-parameter fit for the Math and Science items might be suspect. As Hambelton and Swaminthan (1985) noted, the robustness of IRT models to these departures is not entirely clear. Therefore, the results for the one-parameter model should be interpreted with some caution.

Research Question 1

Table 3 and 4 present the results addressing the first research question, "How comparable are the CTT-based and IRT-based examinee ability estimates?", by analyzing the comparability of the average correlations between the CTT- and IRT- based person *ability* estimates. Table 3 presents the results for the *n*=1000 data, Table 4 presents the results for the *n*=100 data, and Table 5 presents the results for the n-100 data using the Fisher and Olkin and Pratt's corrected sample correlations. To obtain the entries in Table 3 and 4, the following three steps were invoked: (a) for each of the 100 samples, the IRT one-, two-, and three-parameter model estimates and the CTT estimate were obtained; (b) for each sample the CTT- and IRT-based ability estimates were correlated; (c) the correlations were averaged across the 100 samples for the same sampling plan and test. Consequently, each table entry is the average of 100 correlations. The exception is when the IRT model did not converge. To obtain the average correlation for these and all subsequent tables, all the individual correlations coefficients were transferred to Fisher Zs, averaging the Fisher Zs, and then transforming the average Fisher Z back to the Pearson correlation coefficient.

Table 3.

Comparability of Person Statistics from the Two Measurement Frameworks: Average Correlations between CTT- and IRT-Based Person Ability Estimates (n=1000)

		IRT Models		
Sampling Frame	Tests	1P	2P	3P
Random Samples	English	0.982 (.001)	0.983 (.002)	0.984 (.002)
	Math	0.995 (.000)	0.984 (.001)	0.978 (.001)
	Reading	0.994 (.000)	0.987 (.003)	0.981 (.002)
	Science	0.994 (.000)	0.981 (.002)	0.976 (.002)
Gender group sampling				
Female	English	0.989 (.001)	0.982 (.001)	0.984 (.001)
	Math	0.997 (.001)	0.984 (.001)	0.974 (.002)
	Reading	0.994 (.001)	0.987 (.001)	0.981 (.002)
	Science	0.996 (.000)	0.982 (.001)	0.973 (.002)
Male	English	0.991 (.001)	0.983 (.001)	0.984 (.001)
	Math	0.993 (.001)	0.982 (.001)	0.977 (.001)
	Reading	0.994 (.001)	0.988 (.001)	0.981 (.002)
	Science	0.991 (.001)	0.987 (.001)	0.979 (.002)
Truncated ability group sampling				
High-ability	English	0.999 (.000)	0.978 (.001)	0.930 (.020)
	Math	0.999 (.001)	0.998 (.002)	0.947 (.005)
	Reading	0.999 (.000)	0.967 (.010)	0.945 (.010)
	Science	1.000 (.000)	0.969 (.008)	0.903 (.010)
Low-ability	English	1.000 (.000)	0.976 (.003)	NC
	Math	1.000 (.000)	0.922 (.030)	NC
	Reading	0.999 (.000)	0.955 (.010)	NC
	Science	1.000 (.000)	0.938 (.020)	NC

Note: Standard deviations are presented in parentheses. “NC” are models where all the items did not converge.

Table 4.

Comparability of Person Statistics From the Two Measurement Frameworks: Average Correlations Between CTT- and IRT-Based Person Ability Estimates (n=100)

		IRT Models		
Sampling Frame	Tests	1P	2P	3P
Random Samples	English	0.992 (.000)	0.981 (.004)	0.983 (.002)
	Math	0.995 (.000)	0.983 (.002)	0.975 (.001)
	Reading	0.994 (.002)	0.984 (.003)	0.971 (.006)
	Science	0.994 (.002)	0.981 (.003)	0.965 (.007)
Gender group sampling				
Female	English	0.989 (.001)	0.979 (.003)	0.979 (.009)
	Math	0.988 (.001)	0.979 (.001)	0.980 (.001)
	Reading	0.994 (.002)	0.984 (.003)	0.971 (.005)
	Science	0.997 (.003)	0.984 (.003)	0.961 (.003)
Male	English	0.991 (.001)	0.982 (.001)	0.982 (.002)
	Math	0.981 (.001)	0.973 (.001)	0.979 (.001)
	Reading	0.985 (.002)	0.985 (.003)	0.971 (.009)
	Science	0.992 (.003)	0.979 (.004)	0.970 (.007)
Truncated ability group sampling				
High-ability	English	1.000 (.000)	0.970 (.008)	0.915 (.020)
	Math	0.998 (.000)	0.965 (.005)	0.950 (.005)
	Reading	NC	NC	NC
	Science	1.000 (.000)	0.983 (.005)	0.899 (.027)
Low-ability	English	1.000 (.000)	0.975 (.003)	NC
	Math	1.000 (.000)	0.883 (.030)	NC
	Reading	NC	NC	NC
	Science	1.000 (.000)	0.979 (.009)	NC

Note: Standard deviations are presented in parentheses. “NC” are models where all the items did not converge.

Table 3 shows, as Fan (1998) also did, that the CTT- and IRT-based examinee ability estimates correlated highly across IRT models for every test. All but a few of the correlations ranged above .95 and *all the correlations were above .90*. The sampling plans showed a pattern of progressively lower correlations as the number of model parameter increased. The comparability worsened somewhat as the sampling plans became increasing dissimilar but *sample dissimilarity had only minor effects on the estimates*. Therefore, based on these results, we can safely answer research question one by concluding that the CTT-based and IRT-based examinee ability estimates are very comparable, indicating that an analysis of the ability level of individual examinees will lead to similar results across the different measurement theories.

The Table 4 results were strikingly similar to the Table 3 results, even though sample size was reduced from 1,000 to 100. The CTT- and IRT-based examinee ability estimates correlated highly across IRT models for every test, with most correlations above .90 and the vast majority above .95. Again, the sampling plans indicated an increasing pattern of progressively lower correlations across the IRT models as more parameters were estimated with the condition worsening as the sampling plan became increasing dissimilar.

Overall, an analysis of the ability level of individual examinees, even in small sample (n=100) clinical situations, will lead to similar results across the different measurement frameworks.

Table 5 shows the results of Table 4 (*n* = 100) except the sample correlations from Table 4 have been corrected for bias using both the Fisher and Olkin and Pratt correction. All of the correlations generated using the Fisher correction matched those generated using the Olkin and Pratt correction. The correlations found in Table 5 matched those found in Table 4 expect in 5 cases (9.80% of the correlations did not match). However, the largest disagreement in the five correlations was .001. These results indicated that, despite the small sample size used to compute the Table 4 correlations, the sample correlations are a good estimate of what would be found in the population. This result was expected because the Table 4 correlations are large correlations and the bias is greatest in the .500/-.500 range.

Table 5.

Comparability of Average Correlations Between CTT- and IRT-Based Person Ability Estimates (n=100) Using Fisher and Olkin and Pratt's Unbiased Estimators

		IRT Models					
		Fisher Correction			Olkin and Pratt Correction		
Sampling Frame	Tests	1P	2P	3P	1P	2P	3P
Random Samples							
	English	0.992	0.981	0.984	0.992	0.981	0.984
	Math	0.995	0.983	0.975	0.995	0.983	0.975
	Reading	0.994	0.984	0.971	0.994	0.984	0.971
	Science	0.994	0.981	0.965	0.994	0.981	0.965
Gender group sampling							
Female	English	0.990	0.980	0.979	0.990	0.980	0.979
	Math	0.988	0.979	0.980	0.988	0.979	0.980
	Reading	0.994	0.984	0.971	0.994	0.984	0.971
	Science	0.997	0.984	0.962	0.997	0.984	0.962
Male	English	0.991	0.982	0.982	0.991	0.982	0.982
	Math	0.981	0.973	0.979	0.981	0.973	0.979
	Reading	0.985	0.985	0.971	0.985	0.985	0.971
	Science	0.992	0.980	0.970	0.992	0.980	0.970
Truncated ability group sampling							
High-ability	English	1.000	0.970	0.915	1.000	0.970	0.915
	Math	0.998	0.966	0.951	0.998	0.966	0.951
	Reading	NC	NC	NC	NC	NC	NC
	Science	1.000	0.983	0.900	1.000	0.983	0.900
Low-ability	English	1.000	0.975	NC	1.000	0.975	NC
	Math	1.000	0.884	NC	1.000	0.884	NC
	Reading	NC	NC	NC	NC	NC	NC
	Science	1.000	0.979	NC	1.000	0.979	NC

Note: Standard deviations are presented in parentheses. "NC" are models where all the items did not converge.

Research Question 2

Table 6 and 7 present the results addressing the second research question, "How comparable are the CTT-based and IRT-based item difficulty estimates?", by analyzing the comparability of average correlations between the CTT- and IRT-based item *difficulty* estimates. Table 6 presents the *n*=1000 data while Table 7 presents the *n*=100 data. To obtain the entries in Table 6 and 7, the following three steps were invoked: (a) for each of the 100 samples, the IRT one-, two, and three-parameter models estimates and CTT estimates were obtained; (b) for each sample the CTT- and IRT-based difficulty estimates were correlated; (c) the correlations were averaged across the 100 samples for the same sampling plan and test. Consequently, each of the table values is the average of 100 correlations, expect where the IRT model did not converge. The IRT-based item difficulty estimates were correlated with *both* the CTT-based item difficulty estimate *p* and the CTT-based normalized *p* values. Because the CTT *p* values were not reversed so that the higher the value the more difficult the item, the correlations between the IRT-based item difficulty estimates and the CTT-based *p* values are negative. However, these differences in scaling direction of the difficulty estimates are arbitrary.

Table 6.

Comparability of Item Statistics from the Two Measurement Frameworks: Average Correlations between CTT- and IRT-Based Item Difficulty Indexes (n=1000)

		IRT Models					
		CTT *P* VALUES			CTT NORMALIZED *P* VALUES		
Sampling Frame	Tests	1P	2P	3P	1P	2P	3P
Random Samples	English	-0.992 (.001)	-0.960 (.020)	-0.937 (.020)	1.000 (.001)	0.956 (.020)	0.946 (.010)
	Math	-0.999 (.000)	-0.975 (.008)	-0.909 (.020)	1.000 (.000)	0.973 (.007)	0.909 (.020)
	Reading	-0.998 (.000)	-0.986 (.010)	-0.913 (.020)	1.000 (.001)	0.984 (.010)	0.917 (.020)
	Science	-0.988 (.002)	-0.963 (.010)	-0.948 (.010)	1.000 (.000)	0.961 (.010)	0.964 (.008)
Gender group sampling							
Female	English	-0.992 (.001)	-0.952 (.010)	-0.932 (.010)	1.000 (.001)	0.949 (.010)	0.948 (.009)
	Math	-0.997 (.000)	-0.964 (.006)	-0.914 (.010)	1.000 (.000)	0.966 (.005)	0.912 (.010)
	Reading	-0.998 (.000)	-0.984 (.003)	-0.920 (.010)	1.000 (.001)	0.983 (.003)	0.923 (.010)
	Science	-0.989 (.001)	-0.965 (.009)	-0.963 (.008)	1.000 (.000)	0.958 (.001)	0.974 (.001)
Male	English	-0.993 (.001)	-0.966 (.008)	-0.938 (.010)	1.000 (.001)	0.962 (.008)	0.950 (.008)
	Math	-0.999 (.000)	-0.978 (.004)	-0.917 (.010)	1.000 (.000)	0.980 (.004)	0.920 (.010)
	Reading	-0.998 (.000)	-0.987 (.003)	-0.898 (.020)	1.000 (.001)	0.986 (.003)	0.904 (.020)
	Science	-0.989 (.001)	-0.955 (.010)	-0.953 (.010)	1.000 (.000)	0.954 (.001)	0.972 (.010)
Truncated ability group sampling							
High-ability	English	-0.938 (.008)	-0.812 (.040)	-0.652 (.060)	0.998 (.001)	0.775 (.040)	0.665 (.070)
	Math	-0.969 (.003)	-0.889 (.020)	-0.612 (.050)	0.998 (.002)	0.902 (.020)	0.672 (.050)
	Reading	-0.978 (.002)	-0.844 (.030)	-0.767 (.090)	0.999 (.000)	0.834 (.030)	0.801 (.080)
	Science	-0.936 (.008)	-0.908 (.020)	-0.616 (.030)	0.997 (.001)	0.892 (.030)	0.700 (.040)
Low-ability	English	-0.998 (.002)	-0.949 (.010)	NC	1.000 (.001)	0.951 (.010)	NC
	Math	-0.984 (.002)	-0.909 (.010)	NC	0.999 (.000)	0.913 (.020)	NC
	Reading	-0.997 (.000)	-0.934 (.010)	NC	1.000 (.001)	0.940 (.010)	NC
	Science	-0.997 (.000)	-0.928 (.010)	NC	1.000 (.000)	0.926 (.020)	NC

Note: Standard deviations are presented in parentheses. "NC" are models where all the items did not converge.

Table 7.

Comparability of Item Statistics From the Two Measurement Frameworks: Average Correlations Between CTT- and IRT-Based Item Difficulty Indexes (n=100)

		IRT Models					
		CTT *P* VALUES			CTT NORMALIZED *P* VALUES		
Sampling Frame	Tests	1P	2P	3P	1P	2P	3P
Random Samples	English	-0.991 (.004)	-0.964 (.013)	-0.894 (.095)	1.000 (.000)	0.963 (.013)	0.914 (.093)
	Math	-0.998 (.000)	-0.980 (.006)	-0.892 (.102)	1.000 (.000)	0.980 (.005)	0.896 (.102)
	Reading	-0.998 (.001)	-0.980 (.010)	-0.860 (.063)	1.000 (.000)	0.980 (.010)	0.869 (.062)
	Science	-0.986 (.008)	-0.969 (.011)	-0.878 (.150)	1.000 (.000)	0.969 (.012)	0.905 (.144)
Gender group sampling							
Female	English	-0.989 (.006)	-0.948 (.019)	-0.876 (.113)	1.000 (.000)	0.945 (.020)	0.904 (.113)
	Math	-0.997 (.002)	-0.975 (.008)	-0.864 (.170)	1.000 (.000)	0.977 (.007)	0.869 (.170)
	Reading	-0.993 (.001)	-0.979 (.008)	-0.867 (.085)	1.000 (.000)	0.979 (.007)	0.876 (.083)
	Science	-0.987 (.006)	-0.972 (.011)	-0.886 (.143)	1.000 (.000)	0.969 (.013)	0.907 (.136)
Male	English	-0.993 (.005)	-0.967 (.014)	-0.889 (.100)	1.000 (.000)	0.966 (.014)	0.909 (.097)
	Math	-0.998 (.001)	-0.979 (.009)	-0.901 (.059)	1.000 (.000)	0.979 (.007)	0.905 (.057)
	Reading	-0.998 (.001)	-0.981 (.009)	-0.832 (.102)	1.000 (.000)	0.980 (.008)	0.841 (.101)
	Science	-0.988 (.005)	-0.965 (.010)	-0.890 (.128)	1.000 (.000)	0.965 (.010)	0.918 (.121)
Truncated ability group sampling							
High-ability	English	-0.939 (.016)	-0.907 (.025)	-0.568 (.063)	0.998 (.000)	0.909 (.027)	0.553 (.074)
	Math	-0.962 (.010)	-0.960 (.010)	-0.638 (.096)	0.999 (.000)	0.966 (.009)	0.664 (.104)
	Reading	NC	NC	NC	NC	NC	NC
	Science	-0.943 (.013)	-0.996 (.014)	-0.593 (.076)	0.997 (.001)	0.995 (.013)	0.591 (.085)
Low-ability	English	-0.998 (.001)	-0.982 (.006)	NC	1.000 (.000)	0.982 (.006)	NC
	Math	-0.981 (.013)	-0.971 (.008)	NC	0.999 (.001)	0.973 (.011)	NC
	Reading	NC	NC	NC	NC	NC	NC
	Science	-0.996 (.001)	-0.976 (.012)	NC	1.000 (.000)	0.974 (.016)	NC

Note: Standard deviations are presented in parentheses. "NC" are models where all the items did not converge.

Table 8.

Comparability of Item Statistics From the Two Measurement Frameworks: Average Correlations Between CTT (*P*) - and IRT-Based Item Difficulty Indexes Using Fisher and Olkin and Pratt's Unbiased Estimators (n=100)

		IRT Models					
		CTT P VALUES					
		Fisher Correction			Olkin and Pratt Correction		
Sampling Frame	Tests	1P	2P	3P	1P	2P	3P
Random Samples							
	English	-0.991	-0.964	-0.895	-0.991	-0.964	-0.895
	Math	-0.998	-0.980	-0.893	-0.998	-0.980	-0.893
	Reading	-0.998	-0.980	-0.861	-0.998	-0.980	-0.861
	Science	-0.986	-0.969	-0.879	-0.986	-0.969	-0.879
Gender group sampling							
Female	English	-0.989	-0.949	-0.877	-0.989	-0.949	-0.877
	Math	-0.997	-0.975	-0.865	-0.997	-0.975	-0.865
	Reading	-0.993	-0.979	-0.868	-0.993	-0.979	-0.868
	Science	-0.987	-0.972	-0.887	-0.987	-0.972	-0.887
Male	English	-0.993	-0.967	-0.890	-0.993	-0.967	-0.890
	Math	-0.998	-0.979	-0.902	-0.998	-0.979	-0.902
	Reading	-0.998	-0.981	-0.833	-0.998	-0.981	-0.833
	Science	-0.988	-0.965	-0.891	-0.988	-0.965	-0.891
Truncated ability group sampling							
High-ability	English	-0.940	-0.908	-0.569	-0.940	-0.908	-0.569
	Math	-0.963	-0.961	-0.639	-0.963	-0.961	-0.639
	Reading	NC	NC	NC	NC	NC	NC
	Science	-0.943	-0.996	-0.595	-0.943	-0.996	-0.595
Low-ability	English	-0.998	-0.982	NC	-0.998	-0.982	NC
	Math	-0.981	-0.971	NC	-0.981	-0.971	NC
	Reading	NC	NC	NC	NC	NC	NC
	Science	-0.996	-0.976	NC	-0.996	-0.976	NC

Note: Standard deviations are presented in parentheses. "NC" are models where all the items did not converge.

Table 9.

Comparability of Item Statistics From the Two Measurement Frameworks: Average Correlations Between CTT (*Normalized P*) - and IRT-Based Item Difficulty Indexes Using Fisher and Olkin and Pratt's Unbiased Estimators (n=100)

		IRT Models					
		CTT NORMALIZED *P* VALUES					
		Fisher Correction			Olkin and Pratt Correction		
Sampling Frame	Tests	1P	2P	3P	1P	2P	3P
Random Samples							
	English	1.000	0.963	0.915	1.000	0.963	0.915
	Math	1.000	0.980	0.896	1.000	0.980	0.896
	Reading	1.000	0.980	0.870	1.000	0.980	0.870
	Science	1.000	0.969	0.906	1.000	0.969	0.906
Gender group sampling							
Female	English	1.000	0.946	0.905	1.000	0.946	0.905
	Math	1.000	0.977	0.870	1.000	0.977	0.870
	Reading	1.000	0.979	0.877	1.000	0.979	0.877
	Science	1.000	0.969	0.908	1.000	0.969	0.908
Male	English	1.000	0.966	0.910	1.000	0.966	0.910
	Math	1.000	0.979	0.906	1.000	0.979	0.906
	Reading	1.000	0.980	0.842	1.000	0.980	0.842
	Science	1.000	0.965	0.919	1.000	0.965	0.919
Truncated ability group sampling							
High-ability	English	0.998	0.910	0.555	0.998	0.910	0.555
	Math	0.999	0.966	0.666	0.999	0.966	0.666
	Reading	NC	NC	NC	NC	NC	NC
	Science	0.997	0.995	0.593	0.997	0.995	0.593
Low-ability	English	1.000	0.982	NC	1.000	0.982	NC
	Math	0.999	0.973	NC	0.999	0.973	NC
	Reading	NC	NC	NC	NC	NC	NC
	Science	1.000	0.974	NC	1.000	0.974	NC

Note: Standard deviations are presented in parentheses. "NC" are models where all the items did not converge.

In Table 6, the IRT-based one-parameter item difficulty estimates had very high correlations with the CTT-based *p* values. Almost all the CTT-based and IRT-based one-parameter estimates are correlated around the -.98 range. The only departures are the English and Science estimates in the 60 percentile ability group, but both are above -.90.

The IRT-based difficulty estimates for the two- and three-parameter models both had a high correlation with the normalized and non-normalized CTT-based difficulty estimates. The IRT-based two-parameter model correlations were, generally in the -.95 to -.96 range, with the lowest correlation, -.812, found in the English test 60 percentile group (the normalized value was .775). The IRT three-parameter difficulty was highly correlated with the CTT-based difficulty estimates in the random and gender sampling plans. The truncated ability sampling plan indicated only moderate correlations (-.652 to -.612) on the Math test.

Table 7 shows strong correlations in the -.98 to -.99 range between the IRT-based one-parameter and CTT-based item difficulty estimates for n = 100. As has been seen in previous tables, the two- and three-parameter IRT models produced lower correlations than the one-parameter Rash model. The two-parameter, overall, still showed quite strong correlations, with the large percentage of the correlations in the -.96 range and all the correlation above -.90. The

three-parameter IRT-based difficulty had a high degree of correlation, in the -.85 to -.89 range, with the CTT-based estimates.

Tables 8 and 9 show the results of Table 7 (n = 100) except the sample correlations from Table 7 have been corrected for bias using both the Fisher and Olkin and Pratt correction. All of the correlations generated using the Fisher correction matched those generated using the Olkin and Pratt correction. The correlations found in Table 8 matched those found in Table 7 expect for 19 cases (37.54% of the correlations did not match) in the three-parameter model. However, the largest disagreement in the 19 correlations was only .002. The correlations found in Table 9 matched those found in Table 7 expect for 16 cases (31.37% of the correlations did not match) in the three-parameter model. However, the largest disagreement in the 16 correlations was, again, only .002. These results indicated that, despite the small sample size used to compute the Table 7 correlations, the sample correlations are a good estimate of what would be found in the population. This result was expected because the Table 7 correlations are large correlations and the bias is greatest in the .500/-.500 range.

Overall, concerning the correlations between the CTT-based item difficulty estimates and the IRT-based estimates, the one- and two-parameter IRT item difficulty estimate

provided results very similar to their CTT counterparts. Unless the IRT estimates show a higher degree of invariance, as proponents suggest, there seems to be little value to the IRT estimates above what CTT provides

Research Question 3

Table 10 and 11 present the results addressing the third research question "How comparable are the CTT-based and IRT-based item discrimination estimates?", by analyzing the comparability of average correlations between the CTT- and IRT- based item *discrimination* estimates. Table 10 presents the results for the n=1000 data while Table 11 presents the results for the n=100 data. To obtain the entries in Table 10 and 11, the following three steps were invoked: (a) for each of the 100 samples the IRT one-, two-, and three-parameter models estimates and CTT estimates were obtained; (b) for each sample the CTT- and IRT-based discrimination estimates were correlated; (c) the correlations were averaged across the 100 samples for the same sampling plan and test. Consequently, each of the tabled values is the average of 100 correlations, expect where the IRT model did not converge. Note that the one-parameter IRT model does not estimate item discrimination, as so results for this model are indicated to be "not applicable" ("N/A").

Table 10.

Comparability of Item Statistics From the Two Measurement Frameworks: Average Correlations Between CTT- and IRT-Based Item Discrimination Indexes (n=1000)

		IRT Models					
		Point-Biserial			Fisher Z Transformed Point-Biserial		
Sampling Frame	Tests	1P	2P	3P	1P	2P	3P
Random Samples	English	N/A	0.829 (.050)	0.706 (.070)	N/A	0.833 (.050)	0.703 (.080)
	Math	N/A	0.892 (.030)	0.375 (.090)	N/A	0.898 (.030)	0.385 (.090)
	Reading	N/A	0.913 (.020)	0.582 (.100)	N/A	0.918 (.020)	0.587 (.100)
	Science	N/A	0.795 (.040)	0.293 (.115)	N/A	0.797 (.040)	0.301 (.113)
Gender group sampling							
Female	English	N/A	0.820 (.030)	0.749 (.080)	N/A	0.823 (.030)	0.746 (.080)
	Math	N/A	0.907 (.010)	0.294 (.106)	N/A	0.914 (.010)	0.283 (.106)
	Reading	N/A	0.914 (.010)	0.508 (.102)	N/A	0.920 (.010)	0.512 (.102)
	Science	N/A	0.832 (.030)	0.229 (.127)	N/A	0.833 (.030)	0.241 (.125)
Male	English	N/A	0.840 (.020)	0.691 (.080)	N/A	0.843 (.020)	0.689 (.070)
	Math	N/A	0.878 (.020)	0.443 (.090)	N/A	0.875 (.020)	0.449 (.090)
	Reading	N/A	0.919 (.010)	0.634 (.090)	N/A	0.924 (.010)	0.638 (.090)
	Science	N/A	0.794 (.030)	0.347 (.090)	N/A	0.797 (.030)	0.354 (.090)
Truncated ability group sampling							
High-ability	English	N/A	0.487 (.070)	0.734 (.060)	N/A	0.486 (.070)	0.738 (.060)
	Math	N/A	0.651 (.060)	0.829 (.040)	N/A	0.651 (.060)	0.832 (.040)
	Reading	N/A	0.762 (.050)	0.838 (.050)	N/A	0.762 (.050)	0.841 (.050)
	Science	N/A	0.579 (.080)	0.581 (.040)	N/A	0.872 (.080)	0.874 (.040)
Low-ability	English	N/A	0.956 (.009)	NC	N/A	0.957 (.009)	NC
	Math	N/A	0.892 (.164)	NC	N/A	0.894 (.164)	NC
	Reading	N/A	0.895 (.030)	NC	N/A	0.896 (.030)	NC
	Science	N/A	0.876 (.156)	NC	N/A	0.881 (.157)	NC

Note: Standard deviations are presented in parentheses. “NC” are models where all the items did not converge. The one-parameter IRT model does not estimate item discrimination, as so results for this model are indicated to be “not applicable” (“N/A”).

Table 11.

Comparability of Item Statistics From the Two Measurement Frameworks: Average Correlations Between CTT- and IRT-Based Item Discrimination Indexes (Point-biserial and Fisher Z Transformed (n=100)

		IRT Models					
		Point-Biserial			Fisher Z Tranformed Point-Biserial		
Sampling Frame	Tests	1P	2P	3P	1P	2P	3P
Random Samples	English	N/A	0.857 (.050)	0.726 (.098)	N/A	0.828 (.052)	0.722 (.095)
	Math	N/A	0.879 (.050)	0.626 (.171)	N/A	0.893 (.050)	0.634 (.169)
	Reading	N/A	0.911 (.033)	0.611 (.144)	N/A	0.922 (.032)	0.613 (.143)
	Science	N/A	0.821 (.083)	0.647 (.155)	N/A	0.832 (.085)	0.646 (.153)
Gender group sampling							
Female samples	English	N/A	0.844 (.050)	0.783 (.090)	N/A	0.853 (.050)	0.776 (.090)
	Math	N/A	0.888 (.030)	0.673 (.146)	N/A	0.902 (.030)	0.683 (.142)
	Reading	N/A	0.916 (.028)	0.611 (.128)	N/A	0.926 (.027)	0.613 (.125)
	Science	N/A	0.838 (.083)	0.684 (.147)	N/A	0.847 (.084)	0.685 (.156)
Male samples	English	N/A	0.869 (.050)	0.715 (.104)	N/A	0.880 (.046)	0.711 (.103)
	Math	N/A	0.868 (.040)	0.625 (.152)	N/A	0.882 (.040)	0.632 (.150)
	Reading	N/A	0.919 (.026)	0.623 (.125)	N/A	0.930 (.026)	0.628 (.125)
	Science	N/A	0.828 (.065)	0.631 (.121)	N/A	0.840 (.067)	0.632 (.120)
Truncated ability group sampling							
High-ability samples	English	N/A	0.653 (.080)	0.826 (.050)	N/A	0.654 (.070)	0.827 (.070)
	Math	N/A	0.594 (.090)	0.827 (.074)	N/A	0.595 (.090)	0.825 (.073)
	Reading	N/A	NC	NC	N/A	NC	NC
	Science	N/A	0.534 (.112)	0.802 (.097)	N/A	0.539 (.112)	0.802 (.096)
Low-ability samples	English	N/A	0.908 (.135)	NC	N/A	0.912 (.135)	NC
	Math	N/A	0.835 (.114)	NC	N/A	0.839 (.115)	NC
	Reading	N/A	NC	NC	N/A	NC	NC
	Science	N/A	0.863 (.134)	NC	N/A	0.868 (.070)	NC

Note: Standard deviations are presented in parentheses. "NC" are models where all the items did not converge. The one-parameter IRT model does not estimate item discrimination, as so results for this model are indicated to be "not applicable" ("N/A").

Table 12.

Comparability of Item Statistics From the Two Measurement Frameworks: Average Correlations Between CTT- and IRT-Based Item Discrimination (Point-biserial) Indexes with Fisher and Olkin and Pratt's corrections for bias (n=100)

		IRT Models					
		Point-biserial					
		Fisher Correction			Olkin and Pratt Correction		
Sampling Frame	Tests	1P	2P	3P	1P	2P	3P
Random Samples							
	English	N/A	0.858	0.728	N/A	0.858	0.728
	Math	N/A	0.880	0.628	N/A	0.880	0.628
	Reading	N/A	0.911	0.612	N/A	0.911	0.612
	Science	N/A	0.822	0.649	N/A	0.822	0.649
Gender group sampling							
Female	English	N/A	0.845	0.784	N/A	0.845	0.784
	Math	N/A	0.889	0.675	N/A	0.889	0.675
	Reading	N/A	0.917	0.613	N/A	0.917	0.613
	Science	N/A	0.839	0.686	N/A	0.839	0.686
Male	English	N/A	0.870	0.717	N/A	0.870	0.717
	Math	N/A	0.869	0.627	N/A	0.869	0.627
	Reading	N/A	0.920	0.625	N/A	0.920	0.625
	Science	N/A	0.829	0.633	N/A	0.829	0.633
Truncated ability group sampling							
High-ability	English	N/A	0.655	0.827	N/A	0.655	0.827
	Math	N/A	0.596	0.828	N/A	0.596	0.828
	Reading	N/A	NC	NC	N/A	NC	NC
	Science	N/A	0.536	0.803	N/A	0.536	0.803
Low-ability	English	N/A	0.909	NC	N/A	0.909	NC
	Math	N/A	0.836	NC	N/A	0.836	NC
	Reading	N/A	NC	NC	N/A	NC	NC
	Science	N/A	0.864	NC	N/A	0.864	NC

Note: Standard deviations are presented in parentheses. "NC" are models where all the items did not converge. The one-parameter IRT model does not estimate item discrimination, as so results for this model are indicated to be "not applicable" ("N/A").

Table 13.

Comparability of Item Statistics From the Two Measurement Frameworks: Average Correlations Between CTT- and IRT-Based Item Discrimination (Fisher Z Transformed Point-biserial) Indexes with Fisher and Olkin and Pratt's corrections for bias (n=100)

		IRT Models					
		Fisher Z Transformed Point-biserial					
		Fisher Correction			Olkin and Pratt Correction		
Sampling Frame	Tests	1P	2P	3P	1P	2P	3P
Random Samples							
	English	N/A	0.830	0.723	N/A	0.830	0.723
	Math	N/A	0.894	0.636	N/A	0.894	0.636
	Reading	N/A	0.923	0.615	N/A	0.923	0.615
	Science	N/A	0.833	0.648	N/A	0.833	0.648
Gender group sampling							
Female	English	N/A	0.854	0.778	N/A	0.854	0.778
	Math	N/A	0.903	0.685	N/A	0.903	0.685
	Reading	N/A	0.927	0.615	N/A	0.927	0.615
	Science	N/A	0.848	0.687	N/A	0.849	0.687
Male	English	N/A	0.881	0.713	N/A	0.881	0.713
	Math	N/A	0.883	0.634	N/A	0.883	0.634
	Reading	N/A	0.931	0.629	N/A	0.931	0.629
	Science	N/A	0.841	0.634	N/A	0.841	0.634
Truncated ability group sampling							
High-ability	English	N/A	0.655	0.828	N/A	0.655	0.828
	Math	N/A	0.597	0.826	N/A	0.597	0.826
	Reading	N/A	NC	NC	N/A	NC	NC
	Science	N/A	0.541	0.803	N/A	0.541	0.803
Low-ability	English	N/A	0.913	NC	N/A	0.913	NC
	Math	N/A	0.840	NC	N/A	0.840	NC
	Reading	N/A	NC	NC	N/A	NC	NC
	Science	N/A	0.869	NC	N/A	0.869	NC

Note: Standard deviations are presented in parentheses. "NC" are models where all the items did not converge. The one-parameter IRT model does not estimate item discrimination, as so results for this model are indicated to be "not applicable" ("N/A").

Table 10 presents statistics for the n = 1,000 data that, barring a few exceptions, demonstrated strong relationships of discrimination coefficients across measurement models, regardless of sampling plan or test. The CTT- and IRT-based estimates of item discrimination were highly correlated for the two-parameter model. However, the relationships weakened for the three-parameter IRT models.

The different sampling frameworks also had considerable effect on the results for Table 10. For the random sample framework, across the two different IRT models in Table 10, the English test estimates had the greatest degree of stability. The Reading test estimates generated the highest correlations. The Math test estimates had the biggest in the CTT-based and IRT-based two- and three-parameter models. For the random sample sampling frame, the IRT-based item discrimination correlation with the CTT item discrimination statistic drop from .892 in the two-parameter model to .392 in the three-parameter model.

Table 11 addressed the third research questions ("How comparable are the CTT-based and IRT-based item discrimination estimates?") as it relates to clinical tests (e.g., *n*=100). As has been the case across the n=1,000 data, across every sampling plan expect the 60 percentile group, the two-parameter item discrimination estimates correlated higher, on average, with the CTT-based item discrimination

estimates than did the three-parameter IRT-based estimates.

In the random sampling plan, the average correlation was highest for the Reading test at .911. However, each of the other tests showed fairly strong correlations. The three-parameter IRT-based item discrimination estimates, while weaker, were still in the .61 to .73 range. Interestingly, the Reading test, while having the highest average correlation between the two-parameter IRT-based item discrimination estimates and the CTT-based item discrimination estimates, had the lowest average correlation in the three-parameter IRT model.

Like the random sample plan, the gender samples produced fairly strong correlations for the average correlation between the CTT-based item discrimination and IRT-based item discrimination statistics. As was the case for the previous data, the average correlations between the CTT-based item discrimination and three-parameter IRT-based item discrimination estimates were low. The average correlations for each test were stable within the gender plan and comparable to what was found in the random sample plan.

Tables 12 and 13 shows the results of Tables 11 (n = 100) except the sample correlations from Tables 12 and 13 have been corrected for bias using both the Fisher and Olkin and Pratt correction. Nearly all of the correlations generated using the Fisher correction matched those generated

using the Olkin and Pratt correction. Most of the correlations found in Tables 12 and 13 did not match those found in Tables 11. However, the largest disagreement between the correlations was .001.

Overall, comparing the n=100 versus the n=1000 samples, both samples produced very strong correlations between the CTT-based and IRT-based two-parameter item discrimination estimates. But both produced lower, albeit strong correlations between the three-parameter IRT-based and CTT-based item discrimination estimates. Correspondence of the IRT-based and the CTT-based item discrimination estimates was actually higher, albeit it slightly, for the n=100 samples. However, these results should be evaluated in the light of (a) the CTT item discrimination estimates and IRT item discrimination estimates being more invariant in the *n*=1,000 than in the *n*=100 samples and (b) the standard deviations being larger in the *n*=100 samples.

Research Question 4

Table 14 and 15 present the results addressing the fourth research question "When compared across different samples, how invariant are the CTT-based and IRT-based item difficulty estimates?" by analyzing the comparability of average correlations between item difficulty estimates from two different samples sizes derived from the *same measurement framework* (i.e., CTT vs CTT, or IRT vs IRT). Table 14

presents the *n*=1000 data while Table 15 presents the *n*=100 data.

To obtain the entries in Table 14 and 15, the following three steps were invoked: (a) for each of the 100 samples, the IRT one-, two-, and three-parameter models estimates and CTT estimates were obtained; (b) for each sample the CTT- and IRT-based difficulty estimates were correlated with opposing estimates within the sample sampling plan (e.g., males vs females); (c) the correlations were averaged across the sampling plan for the same test. For example, the entry under the IRT one-parameter between the *p* female-male samples for the science test is the average of the correlations between the CTT *p* values obtained from a female sample and the CTT *p* values obtained from a male sample. Each of the 100 female samples was correlated with the corresponding male sample, generating 100 correlations. The average of these correlation coefficients were .987 (SD=.002).

Table 14.

Invariance of Item Statistics from the Two Measurement Frameworks: Average Between-Sample Correlations of CTT and IRT Item Difficulty Indexes (n=1000)

		CTT Models		IRT Models		
Sampling Frame	Tests	*p* values	Normalized *p* values	1P	2P	3P
Random Samples						
	English	0.992 (.004)	0.990 (.004)	0.991 (.002)	0.923 (.015)	0.954 (.013)
	Math	0.996 (.004)	0.995 (.004)	0.995 (.001)	0.987 (.011)	0.984 (.005)
	Reading	0.989 (.004)	0.988 (.003)	0.989 (.003)	0.978 (.002)	0.959 (.012)
	Science	0.996 (.030)	0.975 (.030)	0.995 (.002)	0.975 (.004)	0.980 (.007)
Female- Male samples						
	English	0.946 (.004)	0.973 (.004)	0.973 (.004)	0.957 (.003)	0.927 (.003)
	Math	0.980 (.003)	0.975 (.004)	0.975 (.004)	0.962 (.010)	0.968 (.009)
	Reading	0.939 (.009)	0.937 (.008)	0.937 (.009)	0.927 (.012)	0.911 (.018)
	Science	0.987 (.002)	0.987 (.002)	0.986 (.003)	0.968 (.010)	0.965 (.011)
High-low ability samples						
	English	0.856 (.011)	0.887 (.011)	0.882 (.012)	0.890 (.022)	NC
	Math	0.814 (.008)	0.853 (.008)	0.480 (.009)	0.883 (.002)	NC
	Reading	0.833 (.010)	0.844 (.020)	0.842 (.014)	0.878 (.024)	NC
	Science	0.845 (.007)	0.912 (.008)	0.918 (.009)	0.919 (.017)	NC

Note: Standard deviations are presented in parentheses. "N/A" are models where all the items did not converge. "1P" = one-parameter IRT model; "2P" = two-parameter IRT model; "3P" = three-parameter IRT model

Table 15.

Invariance of Item Statistics from the Two Measurement Frameworks: Average Between-Sample Correlations of CTT and IRT Item Difficulty Indexes (n=100)

		CTT Models		IRT Models		
Sampling Frame	Tests	*p* values	Normalized *p* values	1P	2P	3P
Random Samples						
	English	0.924 (.023)	0.920 (.023)	0.916 (.022)	0.864 (.043)	0.781 (.121)
	Math	0.957 (.012)	0.953 (.012)	0.951 (.012)	0.933 (.019)	0.844 (.119)
	Reading	0.904 (.022)	0.904 (.023)	0.890 (.023)	0.865 (.033)	0.768 (.079)
	Science	0.959 (.031)	0.950 (.052)	0.946 (.013)	0.912 (.028)	0.705 (.177)
Female- Male samples						
	English	0.360 (.432)	0.314 (.458)	0.301 (.375)	0.304 (.349)	0.269 (.335)
	Math	0.954 (.013)	0.935 (.014)	0.932 (.015)	0.907 (.023)	0.752 (.166)
	Reading	0.853 (.039)	0.850 (.039)	0.845 (.039)	0.811 (.059)	0.676 (.145)
	Science	0.953 (.011)	0.946 (.012)	0.943 (.012)	0.910 (.027)	0.748 (.151)
High-low ability samples						
	English	0.632 (.271)	0.622 (.338)	0.791 (.043)	0.779 (.051)	NC
	Math	0.756 (.087)	0.733 (.145)	0.797 (.034)	0.819 (.041)	NC
	Reading	0.762 (.040)	0.766 (.045)	NC	NC	NC
	Science	0.618 (.180)	0.584 (.279)	0.859 (.031)	0.829 (.028)	NC

Note: Standard deviations are presented in parentheses. "NC" are models where all the items did not converge. "1P" = one-parameter IRT model; "2P" = two-parameter IRT model; "3P" = three-parameter IRT model

Table 16.

Invariance of Item Statistics from the Two Measurement Frameworks: Average Between-Sample Correlations of CTT and IRT Item Difficulty Indexes with Fisher and Olkin and Pratt's Corrections for Bias (n=100)

		CTT Models				IRT Models					
		Fisher Correction		Olkin and Pratt Correction		Fisher Correction			Olkin and Pratt Correction		
Sampling Frame	Tests	*p* values	Normalized *p* values	*p* values	Normalized *p* values	1P	2P	3P	1P	2P	3P
Random Samples											
	English	0.926	0.921	0.926	0.921	0.917	0.866	0.784	0.917	0.866	0.785
	Math	0.958	0.954	0.958	0.954	0.952	0.934	0.846	0.952	0.934	0.846
	Reading	0.906	0.906	0.906	0.906	0.892	0.867	0.771	0.892	0.867	0.771
	Science	0.960	0.951	0.960	0.951	0.947	0.913	0.708	0.947	0.913	0.708
Female- Male samples											
	English	0.361	0.315	0.361	0.315	0.304	0.307	0.272	0.303	0.306	0.270
	Math	0.955	0.935	0.955	0.935	0.933	0.909	0.755	0.933	0.908	0.753
	Reading	0.854	0.851	0.854	0.851	0.848	0.814	0.679	0.847	0.812	0.678
	Science	0.954	0.946	0.954	0.947	0.944	0.912	0.751	0.943	0.911	0.750
High-low ability samples											
	English	0.633	0.624	0.633	0.624	0.794	0.782	NC	0.792	0.781	NC
	Math	0.757	0.735	0.757	0.735	0.799	0.821	NC	0.798	0.820	NC
	Reading	0.764	0.768	0.764	0.768	NC	NC	NC	NC	NC	NC
	Science	0.620	0.585	0.620	0.585	0.862	0.831	NC	0.861	0.830	NC

Note: "NC" are models where all the items did not converge. "1P" = one-parameter IRT model; "2P" = two-parameter IRT model; "3P" = three-parameter IRT model

Table 14 indicates that both the transformed CTT *p* and the CTT *p* were strong invariant for the *random sampling* plan, with correlations ranging from .989 to .996. The IRT-based item difficulty estimates for the one-parameter also indicated strong signs of invariance, with correlations ranging from .989 to .995. The two-parameter IRT-based item difficulty estimates were lower, but still strong, with the correlations ranging from .923 to .987. A further drop in the strength of the correlations was found in the three-parameter IRT-based item difficulty estimates, with correlations ranging from .954 to .984.

For the *gender* sample plan, both the transformed CTT *p* and CTT *p* showed signs of *strong* invariance with correlations ranging from .939 to .987. The IRT-based item difficulty estimates for the one-parameter model also indicated strong signs of invariance, with correlations ranging from .937 to .986. The two-parameter IRT-based item difficulty estimates were lower, but still strong, with the correlations ranging from .927 to .968. A further drop in the strength of the correlations was found in the three-parameter IRT-based item difficulty estimates, with correlations ranging from .968 to .911.

The *ability* sample plan yielded results that ran contrary to the other sampling plans. The transformed CTT *p* and CTT *p* showed signs of *strong* invariance, with

correlations ranging from .845 to .856. However, these correlations were weaker than had been found in the previous sampling plans. The IRT-based item difficulty estimates, while still showing a decrease in invariance from the previous sampling plans, showed a higher degree of invariance than did the CTT-based item difficulty estimates.

Table 15 (the clinical samples *n*=100) indicated that, for the *random sample* plan, both the normalized and non-normalized CTT-based item difficulty estimates produced strong correlations (correlations ranged from .904 to .959), indicating that invariance held for the CTT-based estimates. The results from the one-parameter IRT item difficulty estimates indicated that the Rash model item difficulty estimates are virtually identical. However, as was seen in Table 14, the two- and three-parameter item difficulty estimates demonstrated weaker correlations, especially for the English and Reading tests.

The *gender sampling* plan, as was the case in Table 14, indicated a continued degeneration of the correlations found in the random sample plan. Like the previous sampling plan, the math and science tests had greater invariance than did estimates for the other two tests.

Table 16 shows the results of Tables 15 (n = 100) except the sample correlations from Table 16 have been corrected for bias using both the Fisher and Olkin and Pratt

correction. Nearly all of the correlations generated using the Fisher correction matched those generated using the Olkin and Pratt correction. None of the correlations found in Table 16 matched those found in Tables 15. However, the largest disagreement between the correlations was .003.

Overall, although the two measurement frameworks both produced correlations suggesting strong invariance, the CTT item difficulty estimates for the random sampling plan had a higher degree of invariance than did the IRT-based item difficulty estimates, especially the two- and three-parameter models. The large scale measurement samples (n=1,000; Table 14) and the clinical samples (n=100; Table 15) produced very comparable results. Both the IRT and CTT estimates had stronger average correlations when the n=1000 samples were employed. However, the trends, such as greater invariance for the math and science tests, the progressive decay of strength of the average correlations as the sampling frameworks became more dissimilar, and the greater invariance for the IRT-based item difficulty estimates in the one- and two-parameter model, were more pronounced in the clinical sample results.

Research Question 5

Tables 17 and 18 present the results addressing the fourth research question, "When compared across different samples, how invariant are the CTT-based and IRT-based item discrimination estimates?" by analyzing the comparability of average correlations between item *discrimination* estimates from two different samples derived from the same measurement framework. Table 17 presents the *n*=1000 data while Table 18 presents the *n*=100 data. Because the IRT one-parameter *(Rash)* model assumes fixed item discrimination for all items, no correlations could be produced for this model. Therefore, the one-parameter IRT estimates are listed as N/A in the following tables.

Table 17.

Invariance of Item Statistics from the Two Measurement Frameworks: Average Between-Sample Correlations of CTT and IRT Item Discrimination Indexes (n=1000)

		CTT Models		IRT Models		
Sampling Frame	Tests	*Point-biserial*	*Transformed Point-biserial*	1P	2P	3P
Random Samples						
	English	0.862 (.050)	0.865 (.050)	N/A	0.857 (.049)	0.713 (.095)
	Math	0.937 (.060)	0.938 (.060)	N/A	0.927 (.062)	0.791 (.069)
	Reading	0.905 (.060)	0.906 (.060)	N/A	0.880 (.175)	0.782 (.064)
	Science	0.856 (.090)	0.857 (.090)	N/A	0.859 (.010)	0.750 (.060)
Female- Male samples				N/A		
	English	0.836 (.040)	0.839 (.040)	N/A	0.835 (.038)	0.712 (.080)
	Math	0.879 (.030)	0.883 (.030)	N/A	0.909 (.022)	0.756 (.064)
	Reading	0.835 (.050)	0.838 (.050)	N/A	0.862 (.034)	0.740 (.074)
	Science	0.802 (.040)	0.803 (.040)	N/A	0.824 (.047)	0.752 (.065)
High-low ability samples				N/A		
	English	-0.351 (.080)	-0.351 (.080)	N/A	0.346 (.095)	NC
	Math	-0.627 (.050)	-0.625 (.050)	N/A	-0.018 (.118)	NC
	Reading	-0.663 (.080)	-0.659 (.080)	N/A	0.240 (.122)	NC
	Science	-0.508 (.050)	-0.508 (.050)	N/A	-0.257 (.149)	NC

Note: Standard deviations are presented in parentheses. "NC" are models where all the items did not converge. "1P" = one-parameter IRT model; "2P" = two-parameter IRT model; "3P" = three-parameter IRT model. The one-parameter IRT model does not estimate item discrimination, as so results for this model are indicated to be "not applicable" ("N/A").

Table 18.

Invariance of Item Statistics from the Two Measurement Frameworks: Average Between-Sample Correlations of CTT and IRT Item Discrimination Indexes (n=100)

		CTT Models		IRT Models		
Sampling Frame	Tests	*Point-biserial*	*Fisher Transformed Point-biserial*	1P	2P	3P
Random Samples						
	English	0.399 (.128)	0.404 (.129)	N/A	0.396 (.143)	0.387 (.133)
	Math	0.593 (.088)	0.594 (.086)	N/A	0.575 (.108)	0.467 (.152)
	Reading	0.375 (.154)	0.378 (.154)	N/A	0.393 (.161)	0.292 (.202)
	Science	0.430 (.127)	0.438 (.124)	N/A	0.450 (.115)	0.354 (.131)
Female- Male samples				N/A		
	English	0.148 (.458)	0.146 (.225)	N/A	0.035 (.264)	0.087 (.226)
	Math	0.558 (.097)	0.563 (.096)	N/A	0.589 (.099)	0.350 (.159)
	Reading	0.331 (.124)	0.333 (.124)	N/A	0.367 (.135)	0.300 (.154)
	Science	0.473 (.119)	0.475 (.120)	N/A	0.471 (.140)	0.369 (.150)
High-low ability samples				N/A		
	English	0.130 (.151)	0.131 (.151)	N/A	0.178 (.161)	NC
	Math	0.376 (.128)	0.376 (.127)	N/A	0.158 (.095)	NC
	Reading	0.196 (.169)	0.197 (.169)	N/A	NC	NC
	Science	0.301 (.153)	0.301 (.153)	N/A	0.259 (.159)	NC

Note: Standard deviations are presented in parentheses. "NC" are models where all the items did not converge. "1P" = one-parameter IRT model; "2P" = two-parameter IRT model; "3P" = three-parameter IRT model. The one-parameter IRT model does not estimate item discrimination, as so results for this model are indicated to be "not applicable" ("N/A").

Table 19.

Invariance of Item Statistics From the Two Measurement Frameworks: Average Between-Sample Correlations of CTT and IRT Item Discrimination Indexes with Fisher and Olkin and Pratt's corrections for bias (n=100)

	CTT Models				IRT Models					
	Fisher Correction		Olkin and Pratt Correction		Fisher Correction			Olkin and Pratt Correction		
Tests	*Point-biserial*	*Transformed Point-biserial*	*Point-biserial*	*Transformed Point-biserial*	1P	2P	3P	1P	2P	3P
ples										
English	0.402	0.407	0.402	0.407	N/A	0.399	0.390	N/A	0.399	0.391
Math	0.597	0.598	0.597	0.598	N/A	0.579	0.471	N/A	0.579	0.471
Reading	0.378	0.381	0.378	0.381	N/A	0.397	0.294	N/A	0.397	0.295
Science	0.433	0.442	0.434	0.442	N/A	0.454	0.357	N/A	0.454	0.357
le samples					N/A			N/A		
English	0.149	0.147	0.149	0.147	N/A	0.035	0.088	N/A	0.035	0.088
Math	0.560	0.565	0.560	0.565	N/A	0.593	0.353	N/A	0.591	0.352
Reading	0.332	0.334	0.332	0.334	N/A	0.370	0.302	N/A	0.368	0.301
Science	0.475	0.477	0.475	0.477	N/A	0.475	0.372	N/A	0.473	0.370
bility samples					N/A			N/A		
English	0.130	0.131	0.130	0.131	N/A	0.180	NC	N/A	0.179	NC
Math	0.378	0.377	0.378	0.377	N/A	0.160	NC	N/A	0.159	NC
Reading	0.197	0.198	0.197	0.198	N/A	NC	NC	N/A	NC	NC
Science	0.303	0.303	0.303	0.303	N/A	0.262	NC	N/A	0.260	NC

Note: Standard deviations are presented in parentheses. "NC" are models where all the items did not converge. "1P" = one-parameter IRT model; "2P" = two-parameter IRT model; "3P" = three-parameter IRT model. The one-parameter IRT model does not estimate item discrimination, as so results for this model are indicated to be "not applicable" ("N/A").

Looking at Tables 17 and 18, the CTT-based and IRT-based item difficulty estimates were more invariant than the item discrimination estimates. For the *random sample* plan in Table 17, the average correlation of CTT-based item discrimination estimates ranged from .856 to .937. For the same sampling plan, the IRT-based estimates for the two-parameter model were very similar to the CTT-based estimates. However, the three-parameter model average correlations compared with CTT-based estimates were much lower.

The *gender sampling* plan indicated a continued degeneration of the correlations found in the random sample plan. Like the previous sampling plan, the IRT-based estimates for the two-parameter model were very similar to the CTT-based estimates. Also, as in the previous sampling plan, the three-parameter IRT-based correlations, ranging from .712 to .756, were much lower than invariance correlations for the CTT-based item discrimination estimates.

The CTT-based item discrimination estimates, for the ability sampling plan, were appreciably lower than the other sampling plans. However, the CTT-based item discrimination estimates were appreciably higher than the IRT-based item discrimination estimates. In fact, the two- and three-parameter IRT-based item discrimination estimates invariance totally collapsed.

The clinical trial samples (n=100) showed a near total

collapse of invariance across both the CTT and IRT item discrimination estimates. In the *random sample* plan, the Math test maintained moderate invariance (.593). The other tests ranged from .375 (Reading test) to .430 (Science test). The IRT-based item discrimination estimates were very similar for the IRT-based two-parameter and CTT-based estimates. However, as seen in other results, the three-parameter estimates were appreciably lower. Furthermore, the results from Table 13 indicated, as has all the previous results, that as the dissimilarity between sample plans increased, the invariance decreased.

Table 19 shows the results of Tables 18 (n = 100) except the sample correlations from Table 19 have been corrected for bias using both the Fisher and Olkin and Pratt correction. Nearly all of the correlations generated using the Fisher correction matched those generated using the Olkin and Pratt correction. None of the correlations found in Table 19 matched those found in Tables 18. However, the largest disagreement between the correlations was .003. This result was expected because the Table 18 correlations are large correlations and the bias is greatest in the .500/-.500 range.

CHAPTER VII

SUMMARY AND CONCLUSION

In the theory of measurement, there are two competing measurement frameworks, classical test theory and item response theory. The present study empirically examined how the item and person statistics behaved under the two competing measurement frameworks. The study was designed to replicate the work done by Fan (1998). This study focused on two central themes: (1) How comparable are the item and person statistics derived from the item response and classical test framework? and (2) How invariant are the item statistic from each measurement framework across examinee samples?

The data used in this study were from the ACT Assessment Test. The ACT Assessment is composed of four tests: English, Mathematics, Reading, and Science. A sample of 80,000 examinees, each taking the written form and in the same test, were randomly drawn from an examinee population of 322,460. The sample of 80,000 was composed of 40,000 males and 40,000 females. Therefore, each of the four test samples consisted of 10,000 males and 10,000 females. The four test item pools each consisted of 40 items.

To replicate the functionality of the two measurement theories in large scale measurement situations, one set of samples were randomly selected to equal with an n=1,000.

Conversely, to replicate clinical situations where tests are often constructed with small sample sizes, a second set of samples were randomly selected with an n=100. Each of the random samples was drawn under three sampling plans, each progressively dissimilar, thus enabling theoretically greater disparity between the statistics calculated from the different samples.

The major findings were:

1. For the clinical and large-scale samples, the CTT-based and IRT-based examinee *ability* estimates were very comparable, indicating that an analysis of the ability level of individual examinees will lead to similar results across the different measurement theories.
2. The CTT-based item *difficulty* estimates and the one- and two-parameter IRT item difficulty estimate provided very similar results.
3. The investigation of the item *discrimination* statistics marked a downturn in the comparability of estimates across the two measurement models. Both samples produced very strong correlations between the CTT-based and IRT-based two-parameter item discrimination estimates but produced lower, albeit strong correlations between the three-parameter IRT-based and CTT-based item discrimination estimates. The three-parameter IRT-based and CTT-based item discrimination

estimates were actually higher for the n=100 samples.

4. Although the two measurement frameworks produced estimates that were strongly correlated, the CTT item difficulty estimates, for the random sampling plan, had a higher degree of invariance than the IRT-based item difficulty estimates, especially for the two- and three-parameter models.
5. For the large-scale samples, the IRT-based estimates for the two-parameter model were highly correlated with the CTT-based estimates. However, the three-parameter model average correlations were much lower than the CTT-based estimates. Conversely, the clinical trial samples (*n*=100) showed a near total collapse of invariance across both the CTT and IRT item discrimination estimates.
6. All the statistics indicated a progressive decay in the average correlations as the sampling frameworks became more dissimilar.
7. Across all samples, the IRT-based item and person estimates in the one- and two-parameter model were much more similar to the CTT-based item and person estimates. Further, IRT-based item estimates in the one- and two-parameter model were much more invariant than the three-parameter estimates.

Overall, the results of this study indicate that CTT-based and IRT-based estimates, at least for the one-parameter and two-parameter models, are quite similar. This result holds for either small sample clinical trials or large sample assessment situations. The findings indicate that, in a variety of conditions, the two measurement frameworks produce similar item and person statistics.

Proponents of item response theory have centered their arguments for its use on the property of invariance. CTT and IRT may produce very similar results in a single test administration. But because CTT estimates are theoretically sample dependent, across different samples item response theory should yield results that are more invariant. However, as has been shown, classical test theory statistics, for this sample, were just as invariant as their item response theory counterparts.

These results corroborate results reported by Lawson (1991), Fan (1998), Stage (1998a, 1998b, 1999), and MacDonald and Paunonen (2002) all indicating that the two measurement theories often produce quite similar results. The results of this study are part of a growing body of literature that supports Nunnally's (1979) assertion that "when scores developed by ICC theory can be correlated with those obtained by the more usual approach to simply sum items scores, typically it is found that the two sets of scores correlated .90 or higher; thus it is really hair splitting to argue about any difference between the two approaches or any marked departure from linearity of the measurement obtained from the two approaches" (p. 224).

REFERENCES

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Blalock, H. (1968). The measurement problem. In H.M. Blalock & A. Blalock (Eds.), *Methodology in social research* (pp. 5-27). Washington, DC: McGraw Hill

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296-322.

Cantrell, C.E. (1999). Item response theory: Understanding the one-parameter rasch model. In B. Thompson, *Advances in social science methodology* (Vol. 5, pp. 171-192). Stamford, CT: JAI Press.

Carmines, E. & Zeller, R. (1979). *Reliability and validity assessment*. Newbury Park, CA: Sage.

Chang, S., Hanson, B., & Harris, D. (2000, April). *A Standardization Approach to Adjusting Pretest Item Statistics.* Paper presented at the annual meeting of the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA. (ERIC Document Reproduction Service No. ED 442 838).

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

Cochran, W. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley & Sons.

Crocker, L. & Algina, J. (1986). *Introduction to classical & modern test theory*. New York: Holt, Rinehart and Winston.

Dawson, T.E. (1999). Relating variance partitioning in measurement analyses to the exact same process in substantive analyses. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 101-110). Stamford, CT: JAI Press.

Drasgow, F. & Parsons, C. (1983) Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, 7, 189-199.

Ebel, R. L. (1965). *Measuring educational achievement*. Englewood Cliffs, NJ: Prentice-Hall.

Embretson, S. & Reise (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Associates.

Embretson, S. (1999). The new rules of measurement. *Psychological Assessment, 8,* 341-49.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58,* 357-381.

French, C. (2001, January). *A review of classical methods of item analysis*. Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 450 152).

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

Hambelton, R., Swaminathan, H. & Rogers, H. (1991) *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hambleton, R. & Swaminathan, H. (1985). *Item response theory*. Boston: Kluwer Nijhoff.

Henard, D.H. (2000). Item response theory. In L. Grimm & P. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 67-97). Washington, DC: American Psychological Association.

Henson, R.K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34,* 177-189.

Henson, R. (1999, January) . *Understanding the one-parameter Rasch model of item response theory*. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX. (ERIC Document Reproduction Service No. ED 428 078).

Hinkle, D., Wiersma, W. & Jurs, S. (1998). *Applied statistics for the behavioral sciences* (4th ed.). Boston: Houghton Mifflin.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6,* 153-160.

Kelly, T. L. (1939). Selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30,* 17-24.

Kuder, G. & Richardson, C. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151-160.

Lawson, S. (1991). One Parameter latent trait measurement: Do the results justify the effort?. In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, pp. 159-168). Greenwich, CT: JAI Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum Associates.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

MacDonald, P. & Paunonen, S. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62,* 921-943.

McKinley, R. & Mills, C. (1989). Item response theory: Advances in achievement and attitude measurement. In B. Thompson, *Advances in social science methodology* (Vol. 1, pp. 71-135). Stamford, CT: JAI Press.

Norusis, MJ. (1990) *SPSS Advanced Statistics Student Guide*. Chicago: SPSS Inc.

Nunnally, J. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Olkin, I. & Pratt, J.W. (1985). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, 29, 201-211.

Ott, R. (1993). *An introduction to statistical methods and data analysis* (4th ed.). Belmont, CA.: Duxbury.

Reinhardt, B. (1996). Factors affecting coefficient alpha: A mini Monte Carlo study. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 4, pp. 3-20). Greenwich, CT: JAI Press.

Stage, C. (1998a). *A comparison between item analysis based on item response theory and classical test theory: A study of the SweSAT ERC.* (Educational Measurement No 30). Umea University, Department of Educational Measurement.

Stage, C. (1998b). *A comparison between item analysis based on item response theory and classical test theory: A study of the SweSAT test WORD.* (Educational Measurement No 29).Umea University, Department of Educational Measurement.

Stage, C. (1999). *A comparison between item analysis based on item response theory and classical test theory: A study of the SweSAT test READ.* (Educational Measurement No 31). Umea University, Department of Educational Measurement.

Skaggs, G. & Lissitz, R. (1986). An exploration of the robustness of four test equating models. *Applied Psychological Measurement, 10,* 303-317.

Skaggs, G. & Lissitz, R. (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement, 12,* 69-82.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271-95.

Stanley, J. (1971). Reliability. In R.Thorndike(Ed.), *Educational measurement* (3rd ed., pp. 356-442). Washington, DC: American Council on Education.

Thompson, B. (Ed.). (2002). *Score reliability: Contemporary thinking on reliability issues.* Newbury Park, CA: Sage.

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

Thompson, B. (1992, April). *Interpreting regression results: beta weights and structure coefficients are both important.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED

344 897).]

Traub, R. & Rowley, G.(1991). Understanding reliability. *Educational Measurement: Issues and Practice, 8,* 8-14.

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

Wright, B. & Stone M. (1979). *Best test design*. Chicago, IL: Mesa Press.

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

Zimmerman, D., Zumbo, B., & Williams, R. (2003). Bias in estimation and hypothesis testing of correlation. *Psicologica, 24,* 133-158.

VITA

Name: Troy Gerard Courville

Address: 761 Elkins Lake
Huntsville, Texas 77340

Education: Ph.D., Educational Psychology – Research, Measurement and Statistics
Texas A&M University

M.S., Educational Psychology
Texas A&M University

B.S., Psychology
Louisiana State University-Shreveport

Presentations:

Courville, T. & Thompson, B. (2000, April). *Utility of structure coefficient in published reports of regression analysis*. Paper presented at the annual meeting of the American Education Research Association.

Amado, A., Courville, T., George, C., McGee, J., O'Neill, K., Tanguma, J., Walker, D., Willson, V. (2000, January). *An evaluation of the college of education graduate admissions process: A non-registrant perspective*. Paper presented at the annual meeting of the Southwest Educational Research Association.

Courville, T. (1998, February). *Exploratory and confirmatory rotation techniques in exploratory factor analysis*. Paper presented at the annual meeting the Southwestern Educational Research Association.

Publications:

O'Neil, K., George, C., Willson, V., Courville, T., McGee, J., Amado, A., Tanguma, J., & Walker, D. (2002). An evaluation of the college of education graduate admissions process: A non-registrant perspective. *College and University, 77,* 23-26.

Courville, T. & Thompson, B. (2001). Use of structure coefficients in published multiple regression articles: β is not enough. *Educational and Psychological Measurement, 61,* 229-248.