ABSTRACT

Title of Document: A GENERAL METHOD FOR ESTIMATING THE CLASSIFICATION RELIABILITY OF COMPLEX DECISIONS BASED ON CONFIGURAL COMBINATIONS OF MULTIPLE ASSESSMENT SCORES

Karen Mallory Douglas, Ph.D., 2007

Directed By: Professor Robert J. Mislevy, Measurement, Statistics and Evaluation

This study presents a general method for estimating the classification reliability of complex decisions based on multiple scores from a single test administration. The proposed method consists of four steps that can be applied to a variety of measurement models and configural rules for combining test scores:

Step 1: Fit a measurement model to the observed data.

Step 2: Simulate replicate distributions of plausible observed scores based on the measurement model.

Step 3: Construct a contingency table that shows the congruence between true and replicate scores for decision accuracy, and two replicate scores for decision consistency.

Step 4: Calculate measures to characterize agreement in the contingency tables.

Using a classical test theory model, a simulation study explores the effect of increasing the number of tests, strength of relationship among tests, and number of opportunities to pass on classification accuracy and consistency. Next the model is applied to actual data from the GED Testing Service to illustrate the utility of the method for informing practical decisions.

Simulation results support the validity of the method for estimating classification reliability, and the method provides credible estimation of classification reliability for the GED Tests. Application of configural rules results in complex findings which sometimes show different results for classification accuracy and consistency. Unexpected findings support the value of using the method to explore classification reliability as a means of improving decision rules.

Highlighted findings: 1) The compensatory rule (in which test scores are added) performs consistently well across almost all conditions; 2) Conjunctive and complementary rules frequently show opposite results; 3) Including more tests in the decision rule influences classification reliability differently depending on the rule; 4) Combining scores from highly-related tests increases classification reliability; 5) Providing multiple opportunities to pass yields mixed results. Future studies are suggested to explore use of other measurement models, varying levels of test reliability, modeling multiple attempts in which learning occurs between testings; and in-depth study of incorrectly classified examinees.

A GENERAL METHOD FOR ESTIMATING THE CLASSIFICATION RELIABILITY OF COMPLEX DECISIONS BASED ON CONFIGURAL COMBINATIONS OF MULTIPLE ASSESSMENT SCORES

By

Karen Mallory Douglas

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:
Professor Robert Mislevy, Chair
Associate Professor Robert Croninger
Assistant Professor Amy Hendrickson
Professor Robert Lissitz
Professor George Macready

Dedication

This dissertation is dedicated to:

My parents, Ted and Elsie Mallory, who taught me to set high aspirations;

My husband, Doug Douglas, without whose love and support
I could not have completed my doctoral program;

And my daughters, Katie, Laura, and Maggie,
who inspired me to follow my dream.

Acknowledgements

I would like to acknowledge the three "Bobs"
who encouraged me throughout my doctoral program:

Dr. Robert Mislevy, my dissertation advisor, for generous, patient and insightful mentoring throughout the development and execution of this study.

Dr. Robert Lissitz, my academic advisor, for skillful and experienced guidance throughout my doctoral program.

Dr. Robert Croninger, for encouragement in connecting theory to policy.

I would also like to thank the GED Testing Service for
providing access to the test data used in this study,

and to Dr. Carol Ezzelle for providing information
on the psychometric properties of the GED Tests.

Table of Contents

List of Tables vi

List of Figures viii

Chapter 1: Introduction 1
Background 1
Multiple Measures and Complex Decision Rules 4
Reliability As Replicability 9
Reliability of Classification 10
Purpose 12
Significance 13

Chapter 2: Literature Review 14
Studies Related to Traditional Standardized Tests 14
Studies Related to Performance Assessments 19
Classification Reliability Estimates in Cited Studies 24

Chapter 3: Model 27
The Model-Based Approach 28
Application of Model Using Classical Test Theory 30
Classification Reliability of a Single Test 33
Classification Reliability of Two Tests Using a Conjunctive Rule 34
Classification Reliability of Two Tests Using a Complementary Rule 37

Chapter 4: Method 40
Design 40
Simulation Study 40

Chapter 5: Simulation Results 52
Description of Simulated Datasets 52
Individual Tests 55
Research Question 1: Increasing the Number of Tests 58
Fifty Percent Passing Rate 58
Seventy Percent Passing Rate 67
Summary of Results for Research Question 1 70
Research Question 2: Varying the Covariance Among Tests 73
Fifty Percent Passing Rate 73
Seventy Percent Passing Rate 76
Summary of Results for Research Question 2 79
Research Question 3: Allowing Multiple Opportunities to Pass 82
Fifty Percent Passing Rate 82
Seventy Percent Passing Rate 86
Summary of Results for Research Question 3 89

Overview of Simulation Results 91

Chapter 6: Illustration using GED Test Data 96
Description of GED Test Battery 97
Method 98
Overview of GED Test Data 101
Simulated Data 105
Summary of GED Test Illustration 115
Extension to More Complex Configural Rules 118

Chapter 7: Discussion 124
Future Directions 131

Appendix I: Contingency Tables with Counts 135

Appendix II: Computer Code for R-Software Programs 156

Bibliography 161

List of Tables

1.1	Examples of Application of Decision Rules……..	6
2.1	Summary of Classification Reliability in Studies Using Analytic Methods…….	25
2.2	Summary of Classification Reliability in Studies Using Simulation Methods…….	26
3.1	Comparison of the Accuracy of Decision Outcomes……	33
4.1	Contingency Table for Classification Accuracy and Consistency……	44
4.2	Summary of Simulation Conditions for Research Questions 1, 2, and 3…..	50
5.1	Descriptive Statistics and Covariance for True Scores……..	53
5.2	Descriptive Statistics and Covariance for Six Sets of Replicate Scores, COVAR6……	54
5.3	Descriptive Statistics and Covariance for Three Sets of Replicate Scores, COVAR9……...	55
5.4	Accuracy for Individual Tests, COVAR6, 50% Passing Rate……...	56
5.5	Measures of Consistency for Test 1, COVAR6, 50%.....	57
5.6	Percentage Passing Based on True Score, COVAR6, 50%.....	59
5.7	Percentage Passing Based on Two Replicate Scores, COVAR6, 50%.....	60
5.8	Accuracy for One to Five Tests, COVAR6, 50%.....	64
5.9	Consistency for One to Five Tests, COVAR6, 50%.....	66
5.10	Percentage Passing Based on True Score, COVAR6, 70%.....	67
5.11	Percentage Passing Based on Two Replicate Scores, COVAR6, 70%.....	67
5.12	Accuracy for One to Five Tests, COVAR6, 70%.....	68
5.13	Consistency for One to Five Tests, COVAR6, 70%.....	69
5.14	Effect of Adding More Tests on Classification Reliability, COVAR6, 50%	72
5.15	Effect of Adding More Tests on Classification Reliability, COVAR6, 70%	72
5.16	Passing Rate for COVAR6 and COVAR9, Five Tests, 50%.....	73
5.17	Accuracy for COVAR6 and COVAR9, Five Tests, 50%.....	74
5.18	Consistency for COVAR6 and COVAR9, Five Tests, 50%.....	75
5.19	Passing Rates for COVAR6 and COVAR9, Five Tests, 70%.....	76
5.20	Accuracy for COVAR6 and COVAR9, Five Tests, 70%.....	77
5.21	Consistency for COVAR6 and COVAR9, Five Tests, 70%.....	78
5.22	Effect of Increasing Covariance, Five Tests, 50%.....	81
5.23	Effect of Increasing Covariance, Five Tests, 70%.....	81
5.24	Percentage Passing Multiple Attempts, COVAR6, Five Tests, 50%.....	82
5.25	Accuracy for Multiple Attempts, COVAR6, Five Tests, 50%.....	84
5.26	Consistency for Multiple Attempts, COVAR6, Five Tests, 50%.....	85
5.27	Percentage Passing Multiple Attempts, COVAR6, Five Tests, 70%.....	86
5.28	Accuracy for Multiple Attempts, COVAR6, Five Tests, 70%.....	87
5.29	Consistency for Multiple Attempts, COVAR6, Five Tests, 70%.....	88
5.30	Effect of Multiple Attempts to Pass, COVAR6, Five Tests, 50%.....	90
5.31	Effect of Multiple Attempts to Pass, COVAR6, Five Tests, 70%.....	90

6.1	Descriptive Statistics for Half- and Full-Length GED Tests……..	101
6.2	Correlations Among Half- and Full-Length GED Tests……	103
6.3	Reliability Estimates for GED Test Data……..…	104
6.4	Passing Criteria and Rates for Half- and Full-Length GED Tests…….	104
6.5	Correlations Among Tests for Raw Score Versus Normalized Scores…..	108
6.6	Covariance Matrix for Normalized Scores for Odd Half-Length Test…..	108
6.7	Covariance and Reliability for Half-Length Test Simulated Data…….	110
6.8	Passing Rates for Raw Score and Simulated Data for Half-Length Test...	110
6.9	Consistency Among Split-Half, Half-Length Simulated Scores, and Full-Length Simulated Scores……	111
6.10	Covariance Matrix for Normalized Scores for Full-Length GED Test…..	112
6.11	Covariance and Reliability for Full-Length Test Simulated Data……..	113
6.12	Passing Rates for Raw Score and Simulated Data for Full-Length Test…	113
6.13	Classification Accuracy for Simulated, Full-Length Tests……	114
6.14	Percentage of Candidates Who Pass Various Decision Rules……	117
6.15	Contingency Table for Simple and Complex Rule…….	120
6.16	Classification Accuracy for Simple and Complex Rule…….	121
6.17	Classification Consistency for Simple and Complex Rule…….	121

List of Figures

3.1 Illustration of Individual Observed Score and True Score Distributions…. 32
6.1 Histograms Showing Raw Score Distributions for GED Tests…… 102
6.2 Histograms Showing Normalized Scores for GED Tests……. 107

Chapter 1: Introduction

Background

The desire to improve education in the United States is high on the national agenda, and much of the current discussion relates to raising the rigor of instruction toward increasing the knowledge and skills of students. Assessment of student achievement is a key piece in evaluating whether students, and schools, have met learning objectives. Decisions regarding promotion or high school graduation are increasingly being made on the basis of achievement, attendance, and classroom performance. According to a 2005 report by the Center on Education Policy by Sullivan et al, 26 states either required, or are planning to require, the passage of a series of tests in order to earn a high school diploma. Although less prevalent, some states also require students to pass tests in order to be promoted in elementary and middle school. In light of the consequences of such decisions, the ability of test scores to provide accurate and valid measures is of great importance.

One example of the complex rules in place today is found in the Chicago Public Schools. In order to be promoted in the Chicago public elementary schools, each student must pass district tests in reading and math, earn passing grades in reading and math classes, and have no more than nine unexcused absences. Students who do not pass this rule can be promoted through a special review process, or through substitution of test scores received following summer school. High school promotion standards in the Chicago schools require each student to pass at least three core courses both semesters, as well as earning a prescribed number of credits.

Although the reliability of individual achievement test scores is routinely examined, the reliability of such complex decision rules has not been addressed.

The call for responsible use of test scores comes from measurement experts and educational policy analysts as well as the general public. It is well recognized by all these constituencies that no assessment score is without error, and therefore important decisions are best made using several scores or sources of information. The Committee on Appropriate Test Use, formed by the National Research Council, articulated the importance of the use of more than one measure in making decisions in *High Stakes: Testing For Tracking, Promotion, and Graduation* (Heubert & Hauser (Eds.), 1999, pg. 3): "An educational decision that will have a major impact on a test taker should not be made solely or automatically on the basis of a single test score." Other relevant information about the student's knowledge and skills should also be taken into account." This recommendation was also expressed in Standard 13.7 in *Standards for Educational and Psychological Testing* (AERA, 1999, pg. 147): "In educational settings, a decision or characterization that will have major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision." The importance of this recommendation was underscored in 2003 by the National Council on Measurement in Education which devoted an entire issue of *Educational Measurement: Issues and Practice* (Vol. 22, No. 2) to an exploration of the use of multiple measures in decision making.

From a validity standpoint it is logical to suggest that using more than one test score will result in better decisions: common sense tells us that more evidence,

assuming it is credible, should improve our decisions. If a student passes both the essay portion of a test as well as multiple-choice questions on grammar we are more confident that the student can write at the desired level. Using both classroom grades in math and a score on a mathematics achievement test should provide a better estimate of a student's math abilities. However, Cronbach, Linn, Brennan and Haertel (1997) suggested caution in adopting this logic. They used generalizability analysis to investigate various sources of error in performance assessment, and issued a warning on combining assessment scores based on a complex decision rule (pg. 381), saying "Those who endorse such a rule should be aware that measurement error is likely to make decisions highly fallible." The reason for this admonition is that measurement error compounds under certain multiple decision strategies. The laws of probability state that the joint probability of two independent events[1] is equal to the product of the individual probabilities. Consider a student whose true[2] score is above the criterion on two tests and therefore should pass a conjunctive rule (that is, the student must pass both tests) but due to measurement error has a probability of passing Test 1 equal to 0.6, and probability of passing Test 2 equal to 0.8. The probability of the student passing **both** tests is the product of 0.6 and 0.8, or 0.48, which is lower than the probability of passing either test alone. In this case, more information does not always increase the accuracy of the decision.

[1] In most situations, test scores are most likely correlated. This discussion assumes independence for simplicity.

[2] "True" score is used throughout to denote the score an examinee would receive on the test if there were no measurement error, rather than as a measure of true ability on the construct of interest. The second usage is more closely related to the concept of validity.

Multiple Measures and Complex Decision Rules

The term "multiple measures" has been applied to a wide number of decision strategies and purposes. Gong and Hill (2001) listed twelve different situations in which the use of multiple measures may be indicated. These uses range from combining test scores from several content areas to make an overall decision (such as promotion or graduation), combining scores from several different types of assessments in the same content area to increase the validity of the decision (such as a multiple-choice test on grammar and a writing sample), combining scores for students who substitute an alternative test for the usual test (such as Advanced Placement Tests for high school exit exams), and allowing students who fail the opportunity to retake tests to reduce the impact of measurement error and/or to allow for remediation.

The purpose of each decision leads to different rules for combining multiple measures, which in turn influences the estimation of classification reliability. Examples of complex rules are readily found in examples of high-stakes testing throughout the United States. The following policy regarding requirements for promotion to fifth grade is excerpted from the Louisiana State Department of Education:

> "The LEAP 21 tests measure your knowledge and skills in English language arts, math, science and social studies to see whether you know enough to move to the next grade. Students must pass the English Language Arts and Math tests to be considered for promotion to the next grade. … Current policy

states that a student who scores at the *Approaching Basic* level in Mathematics must score at the *Basic* level or above in English Language Arts to pass the LEAP 21. Alternatively, a student who scores at the *Approaching Basic* level in English Language Arts must score at the *Basic* level or above in Mathematics to pass the LEAP 21."

Louisiana has a similar policy for promotion to ninth grade.

At the high school level, an increasing number of states are requiring passage of exit exams in order to receive a high school diploma. Maryland requires students to pass tests in Algebra, Biology, English, and Government, and applies the following criteria in evaluating the results of these tests:

> "The passing scale scores for the High School Assessments are: Algebra 412, Biology 400, English 396, and Government 394. If a student does not pass an HSA, he or she can still fulfill the HSA requirement for the Maryland diploma by earning at least the minimum score on each test **and** a combined score of 1602. The combined score is the total of all HSA test scores. The minimum HSA scale scores are Algebra 402, Biology 391, English 386, and Government 387."

Chester (2003) provides a clear overview of the different ways in which any measure (e.g., test scores, assessments, attendance records, teacher ratings, other indicators) can be combined in making a decision about a student. He outlines three basic combination rules and provides some insight into the situations in which each type of rule is applicable:

- conjunctive (AND)
- complementary (OR)
- compensatory (+)

A conjunctive rule requires the student to pass all measures; a complementary rule requires the student to pass only one of a number of measures. The compensatory rule adds scores together, and therefore allows a higher performance on one measure to counterbalance a lower score on another measure. Chester further outlines four general situations in which multiple measures are used: measures of different constructs; different measures of the same constructs; multiple opportunities to pass; and allowance for accommodations and alternate assessments. He suggests that it is usually most appropriate to combine measures of different constructs using a conjunctive rule, whereas multiple opportunities and alternate assessments lend themselves to a complementary rule. Different measures of the same construct may be combined with any of the types of rules. Table 1.1 presents examples offered by Chester of application of each of the types of rules in educational settings.

Table 1.1: Examples of Application of Decision Rules

Rule	Example
Conjunctive	• Diploma awarded based on passage of a set of exit tests. • Passage of course based on minimum teacher's grade as well as passage of district-wide exam.
Complementary	• Providing multiple opportunities to pass a test. • Allowing students to pass any of a set of alternate assessments, such as is the case for English Language Learners or students in accelerated classes (such as Advanced Placement).
Compensatory	• Class grades calculated by adding together scores from midterm and final exam. • Test score calculated by adding the number of correct items. Items could be of a similar type, or mixture of multiple-choice and constructed response.

In practice, school and licensing policies frequently combine several of these basic rules. Chester describes promotion and graduation policies in the Philadelphia School System in 1999 that combined several of these types of decision rules. For example, in order to be promoted to fifth grade a student must:

- receive a passing mark ('D' or higher) in reading, mathematics, science, *and* social studies; and
- achieve the Below Basic III standard on the Stanford Achievement Test, Ninth Edition (SAT-9) *or* achieve the third-grade level on the citywide test in both reading and mathematics; and
- successfully complete a multidisciplinary project.

Furthermore, students in bilingual programs could substitute Spanish-language versions of the standardized and/or citywide test. This policy represents a complex combination of conjunctive and complementary rules. The decision rule is even more complex because students are allowed multiple opportunities to pass each measure, can take accommodated versions of the tests, and/or can substitute versions in other languages.

A search of the literature and other likely sources, such as state department of education websites, did not yield an exhaustive summary of the prevalence of different types of decision rules used for high school graduation and promotion decisions in use today. According to a 2005 report by the Center on Education Policy, most states use a conjunctive rule in which the student must earn a passing score on each test for high school exit decisions; however, because the student is allowed multiple attempts to pass each test this rule adds a complementary layer to the

conjunctive rule. One state, Maryland, uses a conjunctive-compensatory rule for high school exit exams in which the student can pass either by earning a criterion score on each of five tests *or* by earning a somewhat lower score on some tests *and* a prescribed overall score (Maryland State Department of Education website). Of course, since students are allowed multiple opportunities to pass, and may substitute accommodated or alternate test versions, in actuality the rule is even more complex. The GED Testing Service requires examinees to earn a passing score on five tests, as well as a total overall score, in order to be eligible for a high school equivalency credential (GED Technical Manual, 2006). Although each state sets its own rules in regard to retesting for the GED, most states allow examinees to retest as many as three times each year.

Complex decision rules are also used to combine different types of test items. For example, the CLAS tests in California incorporated performance-rated and multiple-choice items. Overall proficiency ratings on the tests allowed for several different combinations of the two types of scores[3]. A student could be designated as 'Proficient" by earning a rating of 4 on the performance item and a total score of 3, 4, or 5 on the multiple-choice items, *or* by earning a rating of 3 on the performance item and a total multiple-choice score of 5. This is an example of a complex combination of conjunctive and complementary rules.

The examples offered above illustrate the complexity of decisions in use in a variety of school and testing programs. Chester (2003) suggests that in estimating the validity and reliability of complex decisions, the choice of decision rule may be as important as the measures on which the decision is based. Established methods for

[3] Personal communication: David Wiley, 9/20/2004.

estimating the classification reliability for decisions based on single tests do not allow for the estimation of reliability of such complex rules. This study suggests a method that has the flexibility to evaluate classification reliability for many of the complex decisions commonly used in educational settings today.

Reliability As Replicability

An investigation of the impact of measurement error on classification reliability must first include a definition of what is meant by the term "reliability." Entire books have been written about reliability, but in its most general sense it refers to consistency of measurement. If we obtained a second score (or replicate) for a group of examinees, how similar would it be to the first score obtained? The conditions that are varied in obtaining the replicate score depend on the intended use of the test score. For example, test-retest reliability estimates the consistency of measurement for the same test form given under similar test conditions; alternate-form reliability investigates the consistency of scores obtained from two forms of the same test; and inter-rater reliability assesses the similarity between scores given by two independent raters of the same performance. It is important to remember that reliability, however it is defined, is specific to a set of test scores used for a given population of examinees for a prescribed purpose.

The approach to reliability taken in this study is described in Brennan (2001), who frames the basic question of reliability in terms of replicability and emphasizes the importance of delineating the conditions under which we want to assess the replicability in scores. More specifically, the replicability of classifications based on test scores is of central importance in this project.

In most cases, the ideal method for measuring the reliability of test scores would be to obtain two sets of scores for each examinee and then calculate the correlation between the sets of scores. A second best approach is to calculate the correlation between two split-half scores if only one set of scores is available. Such approaches are informative for norm-referenced tests, in which the relative ranking of scores is of interest, but are less informative for criterion-referenced tests in which interest lies in classifying scores based on a criterion. The critical issue in classification reliability is not the absolute amount of measurement error, but whether the measurement error results in inconsistent classification. Two sets of scores can be highly correlated, but result in low classification reliability if the tests vary in difficulty. In such a case a high correlation indicates that examinees have similar rankings on both sets of scores, but fewer examinees would pass the criterion on the more difficult test. Ultimately, the decision about acceptable levels of measurement error in classification situations rests on the consequences for decisions based on the scores. Different uses of test scores allow for different levels of consistency depending on the nature of the decision to be made, the population of interest, and the criterion applied.

Reliability of Classification

Crocker and Algina (1976) discuss the estimation of decision consistency for parallel forms of mastery tests from the standpoint of generalizability theory and highlight four factors that affect consistency.

1. Test length. Increasing the number of items is a well accepted technique for increasing test reliability, which in turn increases decision consistency.

2. Location of the cut score in the score distribution. Scores near to the cut score are most at risk for misclassification, and therefore when the cut score is located near many scores classification error increases.
3. Test score generalizability. Higher generalizability is associated with higher consistency.
4. Similarity of the score distributions for the two forms. Higher similarity is associated with higher consistency.

The question has been raised as to whether all misclassifications are equally important. For example, is misclassifying an examinee who is just below the cut score more serious than misclassifying an examinee who is further away from the cut score? Crocker and Algina (1976) describe two reliability indices designed to incorporate information about distance of scores from the cut-score -- Livington's K^2 and Kane and Brennan's *M(C)* – citing Kane and Brennan's index as preferred because it incorporates a measure of variance due to item difficulty. Crocker and Algina compare K^2, *M(C)*, percent agreement, and Cohen's Kappa at a number of cut-scores imposed on a set of hypothetical data consisting of two scores for eight examinees. All four indices yielded similar estimates at extreme cut scores, and were most disparate when the cut score was in the middle of the distribution. The decision of whether to use K^2 or *M(C)* rests with the researcher and his or her decision regarding whether misclassifications are equally important regardless of the proximity of observed score to the cut score. Both of these measures address the reliability of test scores rather than the reliability of classification decisions (i.e.,

master or non-master), and are therefore not of primary interest in this study given its focus on the reliability of classification decisions.

Purpose

Given current recommendations to use multiple measures in making decisions, and the admonition issued by Cronbach et al (1997) about the potential for increased error in such decisions, a method for estimating decision reliability for complex decisions is warranted. The purposes of this study are to (1) present a general method for estimating decision reliability for multiple measures using data from a single test administration; (2) investigate factors that affect decision reliability of complex decision rules; and (3) demonstrate the utility of the method using actual data.

Ideally, decision reliability would be demonstrated by administering two or more versions of an assessment to the same examinees and assessing the consistency of the decision outcome based on each of the assessment scores. Such an approach is often not practical, and much attention has been focused on estimating reliability from a single administration of a test or assessment. Because the underlying scales for test and assessment scores can be either continuous (as is the case for many standardized tests) or categorical (as is typical for performance tasks), a general method that accommodates a variety of response scales and ability distributions would prove most useful. In addition, proficiency levels can be thought to represent segments along an underlying continuous distribution, or they can be conceptualized as representing qualitative differences in ability such as in latent class analysis. In many practical applications, the continuous distribution is most commonly assumed

with the imposition of a number of cut-scores used to assign students to proficiency levels. Latent class analysis, however, provides an alternative to what can be viewed as an arbitrary process in dividing the latent trait distribution into proficiency classes. This study will present a general method that could be used for any type of response scale or latent trait distribution, and illustrate this method through application using a classical test theory measurement model.

Significance

This study will expand existing theory about classification reliability for a single test to address classification reliability in the common situation of imposing complex decision rules on sets of multiple measures. Complex decision rules may be motivated by the desire to increase validity, reliability, or both. The estimation of classification reliability provided by the method lays the necessary groundwork for investigating important questions concerning the validity of such rules since it is commonly thought that reliability is a necessary condition for validity. A general method is outlined that is suitable for estimating classification accuracy and consistency from a single administration for a wide variety of assessment types and measurement models. This method is demonstrated using simulation techniques for the specific case of multiple scores obtained from standardized tests. Finally, the application of the method to a real world application provides evidence of the utility of the approach.

Chapter 2: Literature Review

Although no examples are found in the measurement literature of methods to estimate classification reliability for complex decision rules, a number of studies present such methods for single standardized test scores and/or performance assessment scores. A selection of these studies is reviewed to illustrate different approaches used in estimating classification reliability for single scores as a starting point in developing an approach for use with multiple scores, as well as to shed light on factors that influence classification reliability. The emphasis in this review is on the methods utilized, rather than specific findings in regard to level of reliability. The amount of potential misclassification estimated in these studies is of interest, however, in supporting the need to investigate potential misclassification as it applies to multiple measures. Results for studies cited in this review appear at the end of this chapter in Tables 2.1 and 2.2.

Studies Related to Traditional Standardized Tests

Increased interest in criterion-referenced testing in the 1970's spawned investigation of the reliability of classifications based on these test scores. Subkoviak (1980) reviewed five analytic approaches to estimating classification reliability for tests comprised of dichotomously-scored items. Two of these methods (Carver, 1970; Swaminathan, Hambleton & Algina, 1974) require two sets of scores for each examinee to estimate classification reliability, and are therefore not useful for most practical problems. Three other methods, however, present approaches to estimating classification reliability on the basis of a single score for each examinee (Huynh,

1976; Marshall-Haertel, 1976; Subkoviak, 1976). Huynh (1976) modeled a set of test scores using a beta binomial distribution, and based on the desired mean, variance, and Kuder-Richardson 21 coefficient for a set of scores then calculated the proportion of the joint distribution that represented consistent classification on both scores for a given criterion score. Due to the computational complexity of this approach, Huynh (1976) also demonstrated a simpler approximation method by assuming that scores are normally distributed, and asserted that such an assumption is warranted in cases where the number of items exceeds 8, and the ratio of mean score to the number of items is between .15 and .85.

Subkoviak's method (1976) uses only the mean and Kuder-Richardson 20 coefficient to calculate the probability of getting an item right for an examinee at each observed score. Assuming a binomial distribution, the probability of getting a specific number of items right (for example, 8 items on a 10 item test) is calculated for each observed score. Then the probability of consistent classification is obtained for the entire group.

The Marshall-Haertel method (1976) uses the binomial distribution to estimate scores on a hypothetical test with twice as many items. This hypothetical test is then divided into two split-half tests and the consistency of classification between the half tests is calculated. Since there are many ways to split a test, the Marshall-Haertel method calculates the consistency for each possible split-half and then takes the average.

Subkoviak (1980) compares results for each of these approaches (along with the Swaminathan et al method that uses two observed test scores) using data for

which 1,586 students took parallel forms for each of several tests. The test lengths were 10, 30, and 50 items. He also applied four different mastery criteria to each test. Subkoviak found that all three single-test methods were difficult to compute and yielded biased estimates for short tests. Estimates using the Huynh method provided underestimates at all criterion levels, whereas the other two methods provided either underestimates or overestimates depending on the criterion level. Subkoviak summarized this exercise by recommending the Huynh method because it offers conservative estimates of consistency, and computation can be simplified through use of a normal distribution. In all approaches, classification reliability increased as test reliability increased, and also as the location of the cut-score became more extreme in relation to the score distribution. In a later study, Subkoviak (1988) uses Hyunh's method to estimate classification consistency for a wide range of cut-scores for tests with reliability ranging from 0.10 to 0.90. Results demonstrate that classification consistency is lowest when the cut-score is in the middle of the distribution and test reliability is also low (cut-score at z = 0.0; reliability =.10; exact agreement = .53), and highest when the cut-score is at the extreme of the test score distribution and test reliability is highest (cut-score at z = 2.00; reliability = .90; exact agreement = .98).

Studies of classification reliability reviewed thus far were based on tests comprised of dichotomous, exchangeable items sampled from well-specified domains. Interest in this topic continued to spawn related studies in the more recent literature, perhaps for several reasons: the increased prevalence of high-stakes decisions based on performance assessment scores with constructed response items, and the arrival of new computer capabilities that greatly expand the possibility for

both analytic solutions and simulation studies. Studies by Wainer and Thissen (1996), Rogosa (1999), and Klein and Orlando (2000) examined decision reliability for dichotomous test items using a classical test theory model. Wainer and Thissen (1996) used a simulation approach to generate two scores for each hypothetical student by constructing two distributions with mean of 500 and standard deviation of 100 in which the scores are correlated to reflect the level of test reliability. They reported three simulations in which the level of the test reliability was varied, and compared actual differences between the two scores. The impact of these differences on classification reliability (i.e., in regard to a potential criterion) was not explored, but the differences between the two scores increased as test reliability decreased.

Rogosa (1999) extended the investigation to examine the reliability of decisions based on percentile ranks (rather than standard scores) both in terms of accuracy and consistency. A particularly useful contribution of Rogosa's study was his straight-forward conceptualization of reliability. Rogosa compared test reliability to the desired accuracy with which a Tomahawk missile hits a target (i.e., "hit rate"). In testing, the hit rate describes how often a percentile score falls within an acceptable target range of the percentile that would be obtained with error-free measurement. Rogosa used a simulation approach in which he generated a hypothetical distribution of true scores, and a distribution of plausible scores around each true score. He then drew two plausible scores for each true score and calculated the probability that the scores were in the same score range (defined as "tolerance") in several conditions in which desired tolerance and test reliability were varied. Rogosa did not, however, apply a decision criterion so no estimates were provided concerning consistent

classification. His findings support Wainer and Thissen (1996) in that higher test reliability was associated with a higher hit rate.

A RAND report by Klein and Orlando (2000) evaluating the City College of New York (CUNY) testing program reported the results of a simulation approach to classification reliability. Findings based on a Monte Carlo simulation study with 10,000 replications reported misclassification as a function of passing rate and test reliability. Given a normal distribution of scores, a higher percentage of examinees will pass when the mean score is increasingly higher than the cut-score. More consistent classification was found as the passing rate moved in either direction away from 50 percent and as test reliability increased.

A study by Rudner (2001) departed from the classical test theory approach in earlier studies, and investigated misclassification from an item response theory (IRT) approach. Using a three parameter IRT model, Rudner generated data for a 50 item test and calculated a standard error of measurement (based on the information function) for each true score. He then calculated the proportion of the distribution of plausible scores for each true score that fell above and below a criterion, and constructed a 2 X 2 contingency table summarizing the two types of correct classification (true masters and true non-masters) and two types of incorrect classification (false master and false non-masters). He reported the potential benefit of including more items toward increasing the reliability of the test, and presumably classification reliability. Rudner explored an interesting, and highly relevant, issue in high-stakes testing by estimating classification consistency in the common case of allowing multiple opportunities to pass a test. His calculations are inaccurate,

however, because he inappropriately used a conditional probability (i.e., the probability of misclassification for a particular true score) as though it represented the probability for the entire group of masters who initially failed the test.

Studies Related to Performance Assessments

Recent trends in the use of performance and constructed-response items present the need to model categorical as well as continuous scores. Livingston and Lewis (1995) developed a complex analytic method that can be used with any type of assessment score, and can also be used for assessments comprised of both dichotomous and polytomous items. Their approach allows for different types of scores by first calculating an estimate of effective test length using the mean, variance, and reliability coefficient for an assessment score. Livingston and Lewis define effective test length as the number of dichotomous, equally difficult, locally independent items necessary to obtain a prescribed reliability estimate. Use of effective test length allows the method to be used with a variety of types of scores, such as essay ratings and performance measures, or a combination of both. Next, a distribution of true scores is estimated based on a four parameter beta distribution and several hypothetical observed scores based on this distribution and the effective test length. From these distributions both classification accuracy and consistency can be summarized. Classification accuracy describes how often the same classification would be made based on the true score and observed score. Classification consistency describes how often the same classification would be made based on two observed scores. Livingston and Lewis evaluated the effectiveness of their approach by

applying it to real data, and comparing the results to those obtained through calculating split-half scores for the tests. The seven tests varied in reliability, location of cut-scores, and consistency of classification. Support for the method was seen in high agreement between results using the analytic method and those based on the split-half scores (most of which yield estimates within .01 of each other) across all conditions. Livingston and Lewis suggested that the method should be further explored with a variety of types of assessments, and with real test-retest data. The Livingston and Lewis method is useful in its flexibility to address several types of scores for a single assessment but given the complexity of the calculations, extension to multiple measures is quite difficult. Their findings generally support other studies in which higher consistency is found for more extreme cut-scores and for higher test reliability.

Another study relevant to the scoring of performance assessments was presented in Bradlow and Wainer (1998), in which a simulation approach was used to examine the potential effect of rescoring on the accuracy of decisions based on performance tasks. Since multiple rating of subjective items is a laborious and expensive task, the question of the accuracy of such decisions is highly pertinent to practitioners. The simulation design was based on the classical test theory (CTT) model, and included a total of 54 conditions with 250,000 examinees in each condition: 3 rescoring rules, 2 levels of initial probability of passing criterion, 3 levels of standard error of measurement, and 3 levels of ratio of true score variance to total variance. Bradlow and Wainer reported the estimated percentage of initial scoring error, as well as the percentage of errors created or ameliorated by rescoring.

Fewer initial scoring errors were found in conditions with smaller standard errors of measurement, and also for conditions with higher ratios of true score to total variance.

Studies reviewed so far all rely on measurement models and assumptions about the distribution of true and observed scores. A simulation study by Brennan and Wan (2004) avoided problems in specifying a distribution by using a bootstrap approach at the item level (termed the "boot-*i*" method) to estimate classification consistency. Similar to the Livingston and Lewis method, the boot-*i* method can model assessments comprised of both dichotomous and polytomous items. Each examinee's item responses serve as the pool of potential responses for the examinee's simulated test. After hypothetical scores are generated, decision rules are applied to both the observed score and hypothetical score, and classification consistency is assessed. An interesting question raised in this approach is whether the sampling of items should be the same for all examinees, or allowed to differ among examinees. Brennan and Wan concluded that sampling the same items for all examinees was most consistent with a test-retest approach to reliability. Sampling different items for examinees would be equivalent to each examinee taking a different form of the test. Brennan and Wan offered an example of a complex assessment comprised of 80 multiple-choice items and 8 constructed response items scored on a four-point scale to which a pass/fail criterion is applied. Using 1000 boot-*i* replications, the average proportion of consistent decisions was .936.

As noted earlier, the fact that the boot-*i* approach does not require any assumptions about distributional form is appealing. It does require other assumptions,

however, that may or may not be justified. In the boot-*i* approach all items are sampled with replacement, and it is assumed that items are exchangeable (within dichotomous or polytomous item types). This may be a questionable assumption if items are not equally difficult or discriminating. In addition, the relatively small sample sizes for items, particularly the constructed response items, may provide questionable estimates. Brennan and Wan recognized that the boot-*i* method may not introduce enough "noise" and may over-estimate the similarity between observed and simulated scores.

Which of the methods reviewed holds the most promise for investigating complex rules for multiple measures? Given the additional complexity introduced by increasing the number of measures on which a decision is based, analytic methods become difficult if not impossible to utilize. Simulation approaches, when well conceptualized, facilitate the modeling of complex distributions and relationships. Although the lack of distributional assumptions in the boot-*i* method is attractive, the number of items in many tests (particularly performance tests) makes this approach questionable. Bayesian methods present an attractive option due to their utility in estimating complex relationships that are not readily calculated analytically (Gelman et al, 1995) and because distributional assumptions are not made about the posterior distributions.

An example of a Bayesian approach was presented in a recent study by Wainer, Wang, Skorupski, and Bradlow (2005) in which the efficacy of a polytomously scored test at a variety of passing scores is investigated using Samejima's graded response model and Markov chain Monte Carlo (MCMC)

procedures. In the Bayesian approach, a prior distribution is combined with observed data to estimate a posterior distribution. Draws can then be made from the posterior distribution and can simply be counted to provide an estimate of the probability of passing. Wainer et al counted draws from the posterior distribution that were above the prescribed passing score, and then constructed graphs showing the probability of passing at a number of levels of proficiency. An example of the method is provided using scores for 6,066 examinees on the Integrated Clinical Encounter Score of Clinical Skills Assessment (CSA). This is a lengthy assessment that includes 10 scores on simulated encounters with patients. These 10 scores are constructed by counting the number of times the examinee solicits information from the patient. In addition, a communication subscore is also constructed based on the effectiveness with which the examinee communicates with the patient. Wainer et al used MCMC procedures to obtain estimates of ability for each examinee, and then constructed a curve depicting the probability of passing by proficiency. Based on this curve they calculated that the probability of incorrectly passing was .028, and the probability of incorrectly failing was .014. Therefore, the probability of a consistent decision was .958.

The closing to the Wainer et al paper provides encouragement for the application of Bayesian methods to investigate complex problems saying, “Using a Bayesian approach requires care and computer time, but not genius.” (p. 278) Given the complexity of estimating and performing calculations on joint distributions for more than two variables, the Bayesian approach seems most appropriate in estimating classification reliability for multiple measures.

Classification Reliability Estimates in Cited Studies

A summary of highlighted classification reliability estimates for studies reviewed appears in Tables 2.1 and 2.2. Studies are listed with abbreviated information about the design, methods, and estimates of classification reliability found in the study. These studies vary greatly in the focus and conditions investigated, such as whether a criterion was applied, and the properties of the distribution that were varied. The purpose of this table is to illustrate the range of classification reliability estimates obtained in these studies as a starting point for investigating such estimates in more complex studies. The lowest estimate, .56, was found for a difficult test with low reliability (Bradlow & Wainer, 1998). The highest reliability, .97, was found for a moderately easy test with 50 items (Subkoviak, 1980).

Table 2.1: Summary of Classification Reliability in Studies Using Analytic Methods

Authors	Assessment	Type of Congruence	Test Reliability	Classification Reliability
Huynh's method (Subkoviak, 1980)	dichotomous items	consistency	.50	.83
			.80	.96
Marshall-Haertel's method (Subkoviak, 1980)	dichotomous items	consistency	.50	.84
			.80	.96
Subkoviak (1980)	dichotomous items	consistency	.50	.91
			.80	.97
Subkoviak (1988)	dichotomous items	consistency	.90	.86 (cut-score set at z=0)
Livingston & Lewis (1995)	dichotomous items	accuracy	Not specified	.90 - .97 (depending on cut-score)
	Advanced Placement Test scores : mostly MC, a few CR	consistency	.64	.75 to .88 (depending on cut-score)
			.85	.92 to .98 (depending on cut-score)
Rogosa (1999)	dichotomous items	accuracy	.70	.36 at 50th percentile and tolerance = .10
			.95	.75 at 50th percentile and tolerance = .10

Table 2.2: Summary of Classification Reliability in Studies Using Simulation Methods

Authors	Assessment	Type of Congruence	Test Reliability	Classification Reliability
Bradlow & Wainer (1998)	essay	accuracy	.90	.88 when probability of passing=.5
			.40	.56 when probability of passing = .5
			.90	.95 when probability of passing = .9
			.40	.90 when probability of passing = .9
Brennan and Wan (2004)	licensure exam –MC and CR items	accuracy	Not provided	.94
Klein & Orlando (2000)	CUNY tests; dichotomous items	accuracy	.30	.60 when probability of passing = .5
			.90	.86 when probability of passing = .5
Rudner (2001)	dichotomous items	accuracy	.92	.89
Wainer & Thissen (1996)	dichotomous items	consistency	.40	65% differ by more than 50 points
			.85	36% differ by more than 50 points
Wainer, Wang, Skorupski & Bradlow (2005)	certification test; polytomous items	consistency	Not provided	.96

Chapter 3: Model

A general framework for investigating classification reliability for a variety of complex decisions and assessment scores follows. This model is presented from the standpoint of accuracy; that is, determining the congruence between decisions based on true score and those based on observed scores. The same model can be applied in estimating classification consistency by comparing the congruence of decisions based on two observed scores.

- Observable scores are written as $X = (X_1, \ldots, X_n)$. Elements of X can be discrete or continuous to reference test scores, performance assessments, or ratings. Elements of X can be from one time point, as when multiple tests must be passed in order to pass a decision rule, or from multiple time points, as occurs when students are allowed multiple attempts to pass a criterion.
- A configural scoring rule is a partition of the X space, into exhaustive and mutually exclusive categories. Examples of such rules include: Pass/Fail; Below Basic, Basic, Proficient, & Advanced; Remedial Math, Algebra only, Geometry only, Algebra + Geometry, Exempt from Math. Coarser partitions can be made by collapsing equivalence classes together, such as combining Proficient and Advanced into Above Basic.
    - A configural scoring rule g that partitions students into K equivalence classes is denoted by $g(X)=(g_1(X), \ldots, g_K(X))$. Each element in this K-

dimensional vector is either 0 or 1: $g_k(X)=1$ if X is in class K and 0 if not.

- A configural rule is defined by the description of the equivalence classes of the X space. Examples of such rules include:
  - CR_1: { $g(X) = 0: X_1 < 3$ OR $X_2 < 3$}
  - CR_2: { $g(X) = 1: X_1 > 3$ OR $X_2 > 3$ AND $X_1 + X_2 \geq 10$}
  - CR_3: { $g(X) = 1: X_1 > 3$ AND $X_2 > 3$ AND $X_1 + X_2 \geq 10$}

In addressing the motivating question concerning the reliability of $g(X)$, note that a specific context and population must be specified. As noted previously, reliability of scores is specific to a particular group of students and a particular decision rule. This population is characterized by a distribution of observed scores. Such reliability could be obtained empirically by administering the test or assessment twice, or through estimation of a model that estimates the conditional distribution of X given proficiency θ from the perspective of any one of a number of measurement models, as best suits the data at hand. Throughout this model, θ represents the score that an individual would earn on a given assessment if measurement could be accomplished without error.

The Model-Based Approach

Given the practical challenges posed by an empirical approach to accuracy, a model-based approach is proposed. Examples of potential measurement models include classical test theory (CTT), item response theory (IRT; both unidimensional

and multidimensional), and latent class analysis (LCA). For each measurement model, the following steps are required:

- Apply the appropriate measurement model to data to estimate $p(X|\theta)$ and the population distribution of proficiency $p(\theta)$.
- Characterize the accuracy of $g(X)$ conditional on θ, or marginal with respect to some distribution $p(\theta)$ of θ in a given population.

There are two cases in the model-based approach:

- *Case 1.* One-to-one relationship between elements of θ and elements of X.
    - g can be applied directly to θ as if it were X in order to get the true value of g corresponding to this θ. Call it $g^t(\theta)$. This is the case in CTT, or Haertel & Wiley's (1993) "true skill vector" LCA model.
- *Case 2.* One-to-many relationship between elements of θ and elements of X, as is the case for multidimensional assessments. That is, a given true value of θ is not associated with a single specific true value of X; it can give rise to many possible values X. For example, in factor analysis there can be 3 factors and 7 X_js. A question arises as to what to use as the true value $g^t(\theta)$ of g in this case, for characterizing accuracy. Propose the category with the highest expected value of g given θ:

$$\boldsymbol{g}^t(\theta) = \max_k E\left[g_k(\boldsymbol{X}) \middle| \theta \right]. \tag{3.1}$$

In all cases, the product of interest is a contingency table comparing true and observed decision outcomes. This table facilitates examination of either marginal or

conditional probabilities of passing the configural rule. Standard indices for classification agreement can be applied to the resulting table, such as percent correct classification or Kappa index, and false negative and false positive rates can be estimated.

- *Conditional.* Through application of the appropriate measurement model, for any given value of θ, $g^t(\theta)$ can be calculated and the distribution $p(g(X)|\ \theta)$ can be produced either analytically or by simulation. In simulation, X is generated conditional on particular θ, and g(X) computed for each. The resulting distribution can be examined for qualities such as proportion of correct classifications, and types of misclassifications (i.e., false negatives versus false positives).
- *Marginal.* Given any particular distribution p(θ) of θ, a contingency table can be created comparing $g^t(\theta)$s versus $g(X)$s, again either analytically or by simulation. In simulation, first θ is drawn from p(θ) then X is generated conditional on drawn θ. Next $g(X)$ is computed for each, and a table of true and observed *g's* is constructed.

Application of Model Using Classical Test Theory

In classical test theory (CTT), each observed score is modeled as a combination of true score and error (Lord & Novick, 1968):

$$X = \theta + \varepsilon \tag{3.2}$$

For simplicity, normality is assumed for both θ and the conditional observed score distributions. However, other distributions could be applied.

$$p(X_i|\theta) = N(\theta,\sigma_e^2) \quad (3.3)$$

$$p(\theta) = N(\mu,\Sigma) \quad (3.4)$$

If X and θ are unidimensional, Σ is true-score variance, and (1) and (2) together imply that X is also normally distributed in the population:

$$p(X) = N(\mu,\Sigma + \sigma_e^2) \quad (3.5)$$

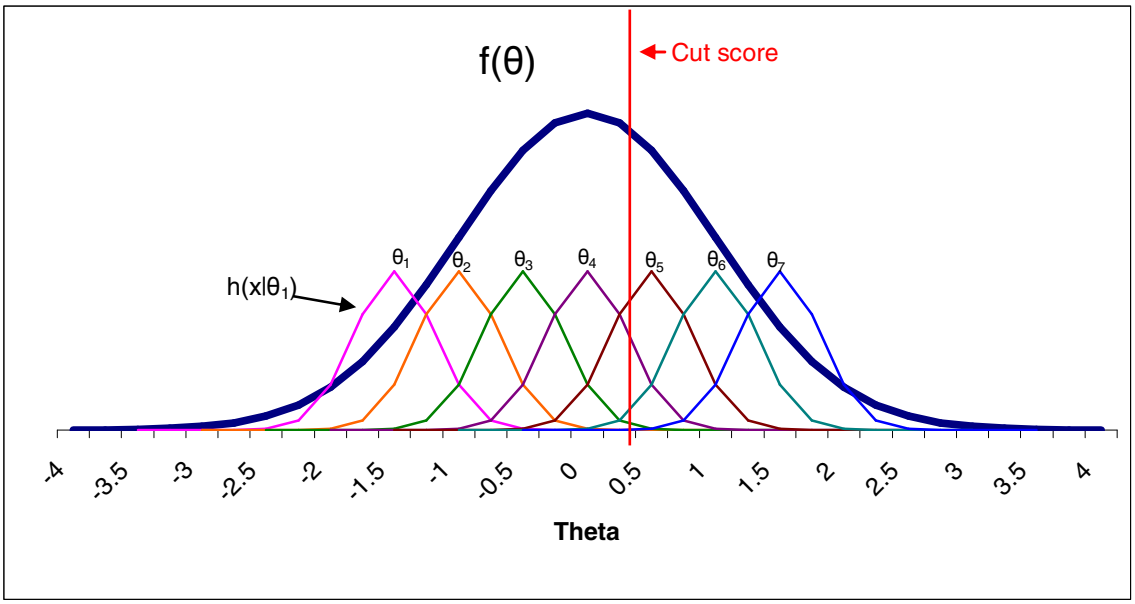
If θ is multidimensional, then $p(\theta)$ is multivariate normal, with mean vector μ and covariance matrix Σ. In this case,

$$p(X|\theta_k) = N(\theta_k,\sigma_e^2) \quad (3.6)$$

$$p(X_1, \ldots, X_{iK})=N((\mu_1, \ldots, \mu_K), \Sigma + \text{diag}(\sigma_{e1}^2, \ldots, \sigma_{eK}^2) \quad (3.7)$$

Figure 1 illustrates the CTT model approach for a unidimensional test with normally distributed true score and error . The heavy line represents the overall distribution of true scores for examinees (denoted $f(\theta)$). The expected distribution of observed scores for seven values of θ are drawn below the true score distribution, $h(X|\theta)$. An arbitrary cut score is shown to illustrate the importance of the proximity between true score and criterion in the potential for misclassification. For example, given the cut point in this diagram, an examinee with θ_1 has a probability of virtually zero of earning an observed score above the criterion. In contrast, an examinee with θ_4 has a moderate probability of obtaining an observed score above the criterion even though θ_4 is below the cut score.

Figure 1: Illustration of Individual Observed Score and True Score Distributions



A presentation of equations for calculating the probability of consistent and inconsistent classifications follows. The following notation is used to indicate true and observed scores, as well as outcome decisions.

- x_{ik} = observed score for student i for test k
- θ_{ik} = true score for student i for test k
- C_k = cut-score set for passing test k
- O = Observed. In the case of a single test, O indicates the observed score. In the case of a decision rule, O indicates the actual decision rendered.
- T = True. In the case of a single test, T indicates the student's true score. In the case of a decision rule, T indicates the decision that would be made based on true scores.
- m = mastery on a single test.
- nm = non-mastery on a single test
- p = passage of a decision rule
- f = failure of a decision rule

There are four possible outcomes in estimating the accuracy of a decision rule. These outcomes are illustrated in Table 3.1. A similar table can be constructed to

estimate the consistency of a decision rule by comparing outcomes on two observed scores.

Table 3.1: Comparison of the Accuracy of Decision Outcomes

True Status	Observed Status	
	Fail	Pass
Fail	True Non-Master	False Master *(false positive)*
Pass	False Non-Master *(false negative)*	True Master

Classification Reliability of a Single Test

Equations to estimate each of the cells in the contingency table for a single test in the CTT model appear below. These equations integrate over two distributions (true and observed) and dissect the joint distribution according to the decision criterion.

True Master:

$$\Pr\left(O_{m1}\left|T_{m1}\right.\right)=\int_{C1}^{\infty}\int_{C1}^{\infty}h\left(x_{i1}\left|\theta_{i1}\right.\right)\partial x_{i1}f\left(\theta_{i1};\mu,\Sigma\right)\partial\theta_{i1} \tag{3.8}$$

False Master:

$$\Pr\left(O_{m1}\left|T_{nm1}\right.\right)=\int_{-\infty}^{C1}\int_{C1}^{\infty}h\left(x_{i1}\left|\theta_{i1}\right.\right)\partial x_{i1}f\left(\theta_{i1};\mu,\Sigma\right)\partial\theta_{i1} \tag{3.9}$$

False Non-Master:

$$\Pr\left(O_{nm1}\left|T_{m1}\right.\right)=\int_{C1}^{\infty}\int_{-\infty}^{C1}h\left(x_{i1}\left|\theta_{i1}\right.\right)\partial x_{i1}f\left(\theta_{i1};\mu,\Sigma\right)\partial\theta_{i1} \tag{3.10}$$

True Non-Master:

$$\Pr\left(O_{nm1}\left|T_{nm1}\right.\right)=\int_{-\infty}^{C1}\int_{-\infty}^{C1}h\left(x_{i1}\left|\theta_{i1}\right.\right)\partial x_{i1}f\left(\theta_{i1};\mu,\Sigma\right)\partial\theta_{i1} \tag{3.11}$$

Classification Reliability of Two Tests Using a Conjunctive Rule

In a conjunctive rule, the examinee must pass all tests. Equations 3.8 – 3.11 can be extended to incorporate two tests for each examinee by integrating over the joint distributions of two true score and two observed scores, and obtaining the density of the area that corresponds to the classification of interest. For example, the following equation calculates the probability of correct classification as a Master (i.e., true score above the criterion on both tests) on the basis of observed scores. Therefore, the equation calculates the probability that all four scores are above the cut score.

True Master:

$$\begin{aligned}\Pr\left(O_{ip}\left|T_{ip}\right.\right)&=\Pr\left(O_{ip}\left|\theta_{i1}\geq C_1\cap\theta_{i2}\geq C_2\right.\right)\\&=\Pr\left(X_{i1}\geq C_1\ \cap\ X_{i2}\geq C_2\left|\theta_{i1}\geq C_1\cap\theta_{i2}\geq C_2\right.\right).\\&=\int_{\theta i2=C2}^{\infty}\int_{\theta i1=C1}^{\infty}\int_{xi2=C2}^{\infty}\int_{xi1=C1}^{\infty}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})\,f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}\end{aligned} \tag{3.12}$$

False Master:

$$\Pr\left(O_{ip}\left|T_{if}\right.\right)=\Pr\left(O_{ip}\left|\theta_{i1}<C_1\cup\theta_{i2}<C_2\right.\right)$$
$$=\Pr\left(X_{i1}\geq C_1\ \cap\ X_{i2}\geq C_2\left|\theta_{i1}<C_1\cup\theta_{i2}<C_2\right.\right). \tag{3.13}$$

$$=\int_{\theta_{i2}=-\infty}^{C_2}\int_{\theta_{i1}=-\infty}^{C_1}\int_{x_{i2}=C_2}^{\infty}\int_{x_{i1}=C_1}^{\infty}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}$$

$$+\int_{\theta_{i2}=C_2}^{\infty}\int_{\theta_{i1}=-\infty}^{C_1}\int_{x_{i2}=C_2}^{\infty}\int_{x_{i1}=C_1}^{\infty}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}$$

$$+\int_{\theta_{i2}=-\infty}^{C_2}\int_{\theta_{i1}=C_1}^{\infty}\int_{x_{i2}=C_2}^{\infty}\int_{x_{i1}=C_1}^{\infty}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}.$$

False Non-Master:

$$\Pr\left(O_{if}\left|T_{ip}\right.\right)=\Pr\left(O_{if}\left|\theta_{i1}\geq C_1\cap\theta_{i2}\geq C_2\right.\right)$$
$$=\Pr\left(X_{i1}<C_1\ \cup\ X_{i2}<C_2\left|\theta_{i1}\geq C_1\cap\theta_{i2}\geq C_2\right.\right). \tag{3.14}$$

$$=\int_{\theta_{i2}=C_2}^{\infty}\int_{\theta_{i1}=C_1}^{\infty}\int_{x_{i2}=-\infty}^{C_2}\int_{x_{i1}=-\infty}^{C_1}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}.$$

$$+\int_{\theta_{i2}=C_2}^{\infty}\int_{\theta_{i1}=C_1}^{\infty}\int_{x_{i2}=C_2}^{\infty}\int_{x_{i1}=-\infty}^{C_1}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}$$

$$+\int_{\theta_{i2}=C_2}^{\infty}\int_{\theta_{i1}=C_1}^{\infty}\int_{x_{i2}=-\infty}^{C_2}\int_{x_{i1}=C_1}^{\infty}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}.$$

True Non-Master:

$$\Pr\left(O_{if}\left|T_{if}\right.\right)=\Pr\left(O_{if}\left|\theta_{i1}<C_1\cup\theta_{i2}<C_2\right.\right)$$
$$=\Pr\left(X_{i1}<C_1\ \cup\ X_{i2}<C_2\left|\theta_{i1}<C_1\cup\theta_{i2}<C_2\right.\right). \tag{3.15}$$

$$=\int_{\theta_{i2}=-\infty}^{C_2}\int_{\theta_{i1}=-\infty}^{C_1}\int_{x_{i2}=-\infty}^{C_2}\int_{x_{i1}=-\infty}^{C_1}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}$$
$$+\int_{\theta_{i2}=-\infty}^{C_2}\int_{\theta_{i1}=-\infty}^{C_1}\int_{x_{i2}=C_2}^{\infty}\int_{x_{i1}=-\infty}^{C_1}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}$$
$$+\int_{\theta_{i2}=-\infty}^{C_2}\int_{\theta_{i1}=-\infty}^{C_1}\int_{x_{i2}=-\infty}^{C_2}\int_{x_{i1}=C_1}^{\infty}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}$$
$$+\int_{\theta_{i2}=C_2}^{\infty}\int_{\theta_{i1}=-\infty}^{C_1}\int_{x_{i2}=-\infty}^{C_2}\int_{x_{i1}=-\infty}^{C_1}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}$$
$$+\int_{\theta_{i2}=C_2}^{\infty}\int_{\theta_{i1}=-\infty}^{C_1}\int_{x_{i2}=C_2}^{\infty}\int_{x_{i1}=-\infty}^{C_1}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}$$
$$+\int_{\theta_{i2}=C_2}^{\infty}\int_{\theta_{i1}=-\infty}^{C_1}\int_{x_{i2}=-\infty}^{C_2}\int_{x_{i1}=C_1}^{\infty}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}$$
$$+\int_{\theta_{i2}=-\infty}^{C_2}\int_{\theta_{i1}=C_1}^{\infty}\int_{x_{i2}=-\infty}^{C_2}\int_{x_{i1}=-\infty}^{C_1}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}$$
$$+\int_{\theta_{i2}=-\infty}^{C_2}\int_{\theta_{i1}=C_1}^{\infty}\int_{x_{i2}=C_2}^{\infty}\int_{x_{i1}=-\infty}^{C_1}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}$$
$$+\int_{\theta_{i2}=-\infty}^{C_2}\int_{\theta_{i1}=C_1}^{\infty}\int_{x_{i2}=-\infty}^{C_2}\int_{x_{i1}=C_1}^{\infty}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}.$$

Classification Reliability of Two Tests Using a Complementary Rule

The complementary rule stipulates that the examinee must pass at least one of a series of tests. As in the conjunctive rule, each equation integrates over the joint distribution of four scores (two true and two observed).

True Master:

$$\begin{aligned}\Pr\left(O_{ip}\left|T_{ip}\right.\right) &= \Pr\left(O_{ip}\left|\theta_{i1}\geq C_1 \cup \theta_{i2}\geq C_2\right.\right)\\ &= \Pr\left(X_{i1}\geq C_1 \ \cup\ X_{i2}\geq C_2\left|\theta_{i1}\geq C_1 \cup \theta_{i2}\geq C_2\right.\right).\end{aligned} \tag{3.16}$$

$$\begin{aligned}
&= \int_{\theta_{i2}=C_2}^{\infty}\int_{\theta_{i1}=-\infty}^{C_1}\int_{x_{i2}=C_2}^{\infty}\int_{x_{i1}=-\infty}^{C_1} h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}\\
&+ \int_{\theta_{i2}=C_2}^{\infty}\int_{\theta_{i1}=-\infty}^{C_1}\int_{x_{i2}=-\infty}^{C_2}\int_{x_{i1}=C_1}^{\infty} h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}\\
&+ \int_{\theta_{i2}=C_2}^{\infty}\int_{\theta_{i1}=-\infty}^{C_1}\int_{x_{i2}=C_2}^{\infty}\int_{x_{i1}=C_1}^{\infty} h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}\\
&+ \int_{\theta_{i2}=-\infty}^{C_2}\int_{\theta_{i1}=C_1}^{\infty}\int_{x_{i2}=C_2}^{\infty}\int_{x_{i1}=-\infty}^{C_1} h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}\\
&+ \int_{\theta_{i2}=-\infty}^{C_2}\int_{\theta_{i1}=C_1}^{\infty}\int_{x_{i2}=-\infty}^{C_2}\int_{x_{i1}=C_1}^{\infty} h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}\\
&+ \int_{\theta_{i2}=-\infty}^{C_2}\int_{\theta_{i1}=C_1}^{\infty}\int_{x_{i2}=C_2}^{\infty}\int_{x_{i1}=C_1}^{\infty} h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}\\
&+ \int_{\theta_{i2}=C_2}^{\infty}\int_{\theta_{i1}=C_1}^{\infty}\int_{x_{i2}=-\infty}^{C_2}\int_{x_{i1}=C_1}^{\infty} h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}\\
&+ \int_{\theta_{i2}=C_2}^{\infty}\int_{\theta_{i1}=C_1}^{\infty}\int_{x_{i2}=C_2}^{\infty}\int_{x_{i1}=-\infty}^{C_1} h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}\\
&+ \int_{\theta_{i2}=C_2}^{\infty}\int_{\theta_{i1}=C_1}^{\infty}\int_{x_{i2}=C_2}^{\infty}\int_{x_{i1}=C_1}^{\infty} h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}.
\end{aligned}$$

False Master:

$$\Pr\left(O_{ip}\left|T_{if}\right.\right)=\Pr\left(O_{ip}\left|\theta_{i1}<C_1\cap\theta_{i2}<C_2\right.\right)$$
$$=\Pr\left(X_{i1}<C_1\ \cup\ X_{i2}<C_2\left|\theta_{i1}<C_1\cap\theta_{i2}<C_2\right.\right). \tag{3.17}$$

$$=\int_{\theta_{i2}=-\infty}^{C_2}\int_{\theta_{i1}=-\infty}^{C_1}\int_{x_{i2}=C_2}^{\infty}\int_{x_{i1}=-\infty}^{C_1}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}$$
$$+\int_{\theta_{i2}=-\infty}^{C_2}\int_{\theta_{i1}=-\infty}^{C_1}\int_{x_{i2}=-\infty}^{C_2}\int_{x_{i1}=C_1}^{\infty}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}$$
$$+\int_{\theta_{i2}=-\infty}^{C_2}\int_{\theta_{i1}=-\infty}^{C_1}\int_{x_{i2}=C_2}^{\infty}\int_{x_{i1}=C_1}^{\infty}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}.$$

False Non-Master:

$$\Pr\left(O_{if}\left|T_{ip}\right.\right)=\Pr\left(O_{if}\left|\theta_{i1}\geq C_1\cup\theta_{i2}\geq C_2\right.\right)$$
$$=\Pr\left(X_{i1}<C_1\ \cap\ X_{i2}<C_2\left|\theta_{i1}\geq C_1\cup\theta_{i2}\geq C_2\right.\right). \tag{3.18}$$

$$=\int_{\theta_{i2}=C_1}^{\infty}\int_{\theta_{i1}=-\infty}^{C_1}\int_{x_{i2}=-\infty}^{C_2}\int_{x_{i1}=-\infty}^{C_1}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}$$
$$+\int_{\theta_{i2}=-\infty}^{C_2}\int_{\theta_{i1}=C_1}^{\infty}\int_{x_{i2}=-\infty}^{C_2}\int_{x_{i1}=-\infty}^{C_1}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}$$
$$+\int_{\theta_{i2}=C_2}^{\infty}\int_{\theta_{i1}=C_1}^{\infty}\int_{x_{i2}=-\infty}^{C_2}\int_{x_{i1}=-\infty}^{C_1}h(x_{i1}\mid\theta_{i1})h(x_{i2}\mid\theta_{i2})f(\theta_{i1},\theta_{i2})\partial x_{i1}\partial x_{i2}\partial\theta_{i1}\partial\theta_{i2}.$$

True Non-Master:

$$\Pr\left(O_{if} \left| T_{if} \right.\right) = \Pr\left(O_{if} \left| \theta_{i1} < C_1 \bigcap \theta_{i2} < C_2 \right.\right)$$
$$= \Pr\left(X_{i1} < C_1 \ \bigcap \ X_{i2} < C_2 \left| \theta_{i1} < C_1 \bigcap \theta_{i2} < C_2 \right.\right). \quad (3.19)$$

$$= \int_{\theta_{i2}=-\infty}^{C_2} \int_{\theta_{i1}=-\infty}^{C_1} \int_{x_{i2}=-\infty}^{C_2} \int_{x_{i1}=-\infty}^{C_1} h(x_{i1} \mid \theta_{i1}) h(x_{i2} \mid \theta_{i2}) f(\theta_{i1}, \theta_{i2}) \partial x_{i1} \partial x_{i2} \partial \theta_{i1} \partial \theta_{i2}.$$

The complexity of solving these equations highlights the advantage of a simulation approach as the configural rule incorporates more tests. Calculating the density of partitioned areas under a multivariate distribution is problematic. Numeric solutions to the equations presented above are extremely difficult and are also subject to estimation assumptions (such as the number of quadratures utilized). A simulation approach, as used in the Wainer et al (2005) study, presents a reasonable method given that the multivariate distribution can be modeled and an adequate number of draws is made. The use of simulation approaches for single tests was demonstrated in a number of studies reviewed in Chapter 2 (Bradlow & Wainer, 1998; Brennan and Wan, 2004; Klein & Orlando, 2000; Rudner, 2001; Wainer & Thissen, 1996; Wainer, Wang, Skorupski & Bradlow, 2005).

Chapter 4: Method

Design

The purpose of this study is to demonstrate a general method for estimating the reliability of classifications based on multiple tests from a classical test theory perspective. Based on a review of the literature and characteristics of classical test theory, a set of simulation conditions is structured to investigate the influence of three important factors on classification accuracy and reliability: number of tests, covariance among tests, and number of attempts permitted. To investigate the possibility that these factors have differential effects depending on test difficulty, all questions are addressed for both a 50% and a 70% passing rate.

Research Question 1: Does increasing the number of tests used to make a decision increase decision accuracy and/or decision consistency?

Research Question 2: Does increasing the covariance among tests increase decision accuracy and/or decision consistency?

Research Question 3: Does allowing multiple attempts to pass increase decision accuracy and/or decision consistency? For simplicity, the assumption is made that true score is the same at each attempt.

Simulation Study

The five decision rules investigated in this study are referred to as conjunctive, complementary, compensatory, conjunctive-complementary, and conjunctive-compensatory. The first three rules, conjunctive (AND), complementary

(OR), and compensatory (+), exhaust the simple ways of combining multiple test scores.

A conjunctive rule requires the examinee to pass **all** of the tests in the battery. It is the rule used by most states for high school exit exams, if the stipulation that examinees can only take each test once is added. In practice, most states allow multiple attempts to pass each of the high school exit exams.

In the complementary rule the examinee must pass **at least one** of the tests in the battery. Allowance of repeated attempts to take a test is an example of this rule. Another example of the application of the complementary rule is the situation in which one test is allowed to substitute for another, such as in the case when AP tests can be substituted for high school exit exams. In Philadelphia, students are promoted to the ninth grade based on Stanford Achievement Test scores or by virtue of their scores on a citywide test in reading and math.

The third simple rule, compensatory, applies a criterion to the **sum (**or average) of a set of test scores. An example of the application of this rule is the case of a college admission criterion that requires a minimum total score for the math and verbal portions of the SAT Tests. It differs from the conjunctive and complementary rules in that the criterion is not applied directly to individual test scores, but rather to the sum of test scores. From a reliability standpoint it offers the potential advantage of allowing errors in test scores on different tests to counterbalance each other. However, if the goal of the rule is to certify a minimum performance on each test this rule is not sufficient.

There are a great number of ways in which these three simple rules can be combined to create more complex rules. Two combination rules are illustrated in this study to show the application of the approach to more complex decisions rules: conjunctive-complementary and conjunctive-compensatory. These rules were selected because they are similar in spirit to rules in use in testing programs today, as is illustrated below.

In the conjunctive-complementary rule used in this study, the examinee must pass several of the tests in a battery, but must only pass one of the remaining tests. A real world example of this rule would be if students were required to pass reading and math tests for promotion to a higher grade, but were only required to pass either a social studies or science test.

The conjunctive-compensatory rule used in the study requires the examinee to meet a criterion for two tests in a battery, and also a prescribed overall score for all tests. This is similar to the rule used by the GED Testing Service in which the examinee must meet a minimum criterion on each of five tests, and also attain a prescribed average over all five tests. Another variation of this rule is used by Maryland for high school exit tests. Students who do not attain the desired score on each of the exit exams can pass if they attain a minimum score on each test and an overall total score. The overall total score is set at a level that allows for superior performance on one or more tests to counterbalance lower performance on others.

The following steps were used to structure the simulations (using version 2.3.1 of *R)* for these five decision rules:

1. Generated 3 multivariate normal datasets consisting of true scores on five tests (T1, T2, T3, T4, T5) for 500,000 examinees. The covariance among the tests was set to either 0.0 ('COVAR.0'), 0.6 ('COVAR6'), or 0.9 ('COVAR9'). The COVAR.0 dataset was used only to show the credibility of the method, and therefore was not included in the full set of simulation conditions.
2. Three replicate scores were generated for each true score for each test for COVAR.0 and COVAR9. Research Question 3 required six replicate scores for COVAR6. Distributions from which the replicate scores were drawn were constructed using the true score, θ, as the mean, with a standard deviation equal to 0.31623 to simulate a test with a reliability coefficient equal to 0.9. For notational purposes, replicate scores are labeled with the test number first, followed by the replication number. For example, the second set of replicate scores for Test 1 is labeled R1.2; the third set of replicate scores for Test 2 is labeled R2.3.
3. Passage of each test was determined by applying a criterion to each score, both true and replicate. To simulate a 50% passing rate, a cut-score of 0.0 was applied to each of the five tests. The criterion for passage of the compensatory rule was also 0.0, and for the conjunctive-compensatory rule an average score of 0.5 (i.e., half of a standard deviation) was required for passage. In the 70% passing rate conditions, a cut-score of -0.525 was applied to each test score. The compensatory rule required an average score of -0.525 or better, and the conjunctive-composite rule required an average score of -.025.

4. Each of the five decision rules was then applied to the sets of true and replicate scores to determine whether the examinee passed each overall decision rule.
5. A contingency table was constructed for each decision rule showing the congruence between classification based on true scores and replicate scores to estimate classification accuracy, and the congruence between two replicate scores to estimate classification consistency.

Table 4.1: Contingency Table for Classification Accuracy and Consistency

	Failed	Passed	Total
Failed	a	b	g
Passed	c	d	h
Total	e	f	N

Measures of agreement were calculated for each contingency table. There are a surprising number of measures for characterizing the relationship between the agreement in a simple contingency table. Some measures, such as chi-square and phi, are most appropriate when the goal is to describe the dependence of one variable on the other. The purpose of this study is most akin to rater agreement, in which interest is in determining whether scores by two raters yield the same decision about the competence of the examinee. Rater agreement is typically assessed by examining the hit rate; i.e., the number of examinees who receive the same classification based on the two ratings. Conditional measures that compute agreement separately for Masters and Non-Masters are also of interest. In addition, measures of the type of disagreement may be important in some decisions. For example, in the case of medical diagnosis it may be more important to avoid false negatives (missing the presence of disease) rather than false positives (diagnosing the patient with a disease

when in fact it is not present). In a legal framework, the United States system of law is based on the premise that the accused is innocent until proven guilty. Therefore, judges may be more concerned with false positives (sentencing an innocent person) than with false negatives (releasing a guilty person). In educational decisions, students who are falsely identified as passing the decision rule may be unable to perform at the next grade level, whereas students who are falsely held back lose valuable instructional time.

Different estimates of agreement were calculated for contingency tables comparing accuracy and consistency. For all tables, however, the following two measures of agreement were calculated.

- Percentage agreement. The percentage of examinees who received the same decision based on both sets of scores. P represents the proportion of agreement. The proportion is multiplied by 100 to obtain percentage agreement.

$$P = (a + d) / N \quad (4.1)$$

$$PCT\ (Agree) = P * 100 \quad (4.2)$$

- Cohen's Kappa. The marginal proportions in a contingency table have a strong impact on percentage agreement. When almost all examinees pass (or fail) the decision, the probability of making the correct classification may be high strictly due to chance. Cohen suggested a correction to percentage agreement (PC) that adjusts for the likelihood of making the correct classification strictly by chance. PC represents

the proportion of agreement that would be obtained if the two classifications were completely independent of one another.

$$PC = (e/N * g/N) + (f/N * h/N) \quad (4.3)$$

$$K = (P - PC) / (1\text{-}PC) * 100 \quad (4.4)$$

Kappa is normally reported on a scale from 0 to 1. It was multiplied by 100 in this study to facilitate comparison with exact agreement.

It must be emphasized that Kappa represents a different, but not necessarily superior, alternative to exact agreement. First, kappa represents the gain in prediction over chance, and therefore its interpretation is not straight forward. Second, Kappa is also influenced by the marginal distribution as shown in Feinstein and Cicchetti (1990) who illustrate several paradoxical relationships between exact agreement and Kappa. Subkoviak (1988) demonstrated that the level of reliability has an inverse effect on Kappa and exact agreement. Both measures are included in this study in an effort to illustrate the utility of various measures of agreement.

For classification accuracy, four additional measures were calculated to further characterize agreement. All four measures take advantage of the fact that the correct decision outcome is known, and can be used to further investigate specific types of disagreement.

- Conditional percentage of agreement for those who passed the decision rule based on true score.

$$PCT (Agree|Master) = d / h * 100 \quad (4.5)$$

- Conditional percentage of agreement for those who failed the decision rule based on true score.

$$PCT\ (Agree|Non\text{-}Master) = a / g * 100 \quad (4.6)$$

- Percentage of false positives. The percentage of examinees that passed the decision, but should have failed based on true score.

$$PCT\ (FP) = b / N * 100 \quad (4.7)$$

- Percentage of false negatives: The percentage of examinees that failed the decision, but should have passed based on their true score.

$$PCT\ (FN) = c / N * 100 \quad (4.8)$$

Given that these measures of agreement are all based on the same simple contingency table, it is informative to consider how they relate to one another. For example, the sum of PCT(Agree), PCT(FP) and PCT(FN) is equal to 100. Therefore, as PCT(Agree) increases, the sum of PCT(FP) and PCT(FN) must decrease. This decrease be reflected equally in PCT(FP) and PCT(FN), or asymmetrically (i.e., an increase in one and a decrease in the other). Conditional measures operate independently of PCT(Agree), PCT(FP), and PCT(FN).

For tables comparing replicate scores (i.e., classification consistency), it is not appropriate to consider false negatives and positives; and conditional agreement is not of great interest. Uebersax (2003) suggests a measure called the "percentage of specific agreement" which is appropriate for such situations. Cicchetti and Feinstein (1990) and Fitzmaurice (2002) also present this measure in more general discussions of measures of agreement. This measure provides an estimate of conditional agreement that takes into account the number classified as passers (and failers) on both sets of replicate scores.

- Percentage of specific agreement for Passers: The number of consistent positive classifications divided by the average number of positive classifications across both tests.

$$PCT\ (SA|Passers) = d / [(d + b + d + c)/2] * 100 \quad (4.9)$$

- Percentage of specific agreement for Failers: The number of consistent negative classifications divided by the average number of negative classifications across both tests.

$$PCT\ (SA|Failers) = a / [(a+ b + a + c)]/2 * 100 \quad (4.10)$$

In all simulation conditions the following parameters remained fixed:

1. Single cut-score. Although in many situations decisions may be based on multiple proficiency categories, overall decisions typically apply only one criterion (e.g., must be in basic proficiency group or higher).
2. Individual test reliability equal to 0.9 for all tests. This is reflective of the lower level of reliability typically found in standardized, multiple-choice tests. It would not be typical, however, of many performance-based tests.
3. Standard, normal distribution, *N(0,1)*, for true score on all tests. Although many score distributions are not normally distributed, a normal distribution is used for simplicity in the simulations.

Because the five tests were identically distributed, and the same criterion was applied to each test, results are descriptive of a set of parallel tests. This stipulation makes sense in the case of repeated attempts on the same test, but is less likely to be true for sets of different tests. For example, it is likely that for many groups of examinees tests of math, reading, social studies, and science would have different

levels of difficulty. The case of differential difficulty and variability will be illustrated later when the general method is applied to actual data from GED test takers.

Details for the simulation conditions appear in Table 4.2. The rule for combining scores is illustrated to clarify the decision rule used in each condition. For Research Questions 1 and 2, all conditions determined classification accuracy by comparing true scores to scores from the first replicate group. For classification consistency, in which decisions are based on multiple sets of replicate scores, the rule was applied to the first and second replicate scores.

For Research Question 3, in which examinees were allowed multiple attempts to pass the test, classification accuracy was estimated by comparing the decision outcomes based on true scores to that based on Replications 1 and 2 for the condition allowing two attempts, and Replications 1, 2, and 3 for the condition allowing 3 attempts. Classification consistency for two attempts compared the decision outcomes based on Replications 1 and 2 with those for Replications 4 and 5. Classification consistency for three attempts compared Replications 1, 2 and 3 to Replications 4, 5, and 6. For the composite rules, the maximum of replication scores for each test was used in creating the overall score.

Table 4.2 shows a symbolic representation for each decision rule. For Research Questions 1 and 2, the same rule was applied to true score and replicate score in each condition. Research Question 3 varied the number of attempts, and used multiple replicate scores in each decision. Table 4.2 shows the rule for replicate scores; the rules for true score are the same as in Research Question 1.

Table 4.2: Summary of Simulation Conditions for Research Questions 1, 2, and 3

Type of Decision Rules	Combination Rule	Number of Tests	Covar-iance	Number of Attempts
RESEARCH QUESTION 1				
Conjunctive (AND)	1 ∩ 2 1 ∩ 2 ∩ 3 1 ∩2 ∩ 3 ∩4 1 ∩ 2 ∩ 3 ∩ 4 ∩ 5	2 3 4 5	.6	1
Complementary (OR)	1 ∪ 2 1 ∪ 2 ∪ 3 1 ∪ 2 ∪ 3 ∪ 4 1 ∪ 2 ∪ 3 ∪ 4 ∪ 5	2 3 4 5	.6	1
Compensatory (+)	(1 + 2) (1 + 2 + 3) (1 + 2 + 3 + 4) (1 + 2 + 3 + 4 + 5)	2 3 4 5	.6	1
Conjunctive-Complementary (AND/OR)	1 ∩ 2 ∩ (3 ∪ 4) 1 ∩ 2 ∩ (3 ∪ 4 ∪ 5)	4 5	.6	1
Conjunctive-Compensatory (AND/+)	1 ∩ 2 ∩ 3 ∩ (1+ 2+ 3) 1 ∩ 2 ∩ 3 ∩ 4 ∩ 5 ∩ (1+ 2 +3 + 4 + 5)	3 5	.6	1
RESEARCH QUESTION 2				
Conjunctive (AND)	1 ∩ 2 ∩ 3 ∩ 4 ∩ 5	5	.6 .9	1
Complementary (OR)	1 ∪ 2 ∪ 3 ∪ 4 ∪ 5	5	.6 .9	1
Compensatory (+)	(1 + 2 +3 + 4+ 5)	5	.6 .9	1
Conjunctive-Complementary (AND/OR)	1 ∩ 2 ∩ (3 ∪ 4 ∪ 5)	5	.6 .9	1
Conjunctive-Compensatory (AND/+)	1 ∩ 2 ∩ 3 ∩ 4 ∩ 5 ∩ (1+ 2 +3 + 4 + 5)	5	.6 .9	1

Type of Decision Rules	Rule	Number of Tests	Covar-iance	Number of Attempts
RESEARCH QUESTION 3				
Conjunctive (AND)	(R1.1 U R 1.2) ∩ (R2.1 U R 2.2) ∩ (R3.1 U R 3.2) ∩ (R4.1 U R 4.2) ∩ (R5.1 U R 5.2)			2
	(R1.1 U R 1.2 U R1.3) ∩ (R2.1 U R 2.2 U R2.3) ∩ (R3.1 U R 3.2 U R3.3) ∩ (R4.1 U R 4.2 U R4.3) ∩ (R5.1 U R 5.2 U R5.3)	5	.6	3
Complementary (OR)	(R1.1 U R 1.2) U (R2.1 U R 2.2) U (R3.1 U R 3.2) U (R4.1 U R4.2) U (R5.1 U R 5.2)			2
	(R1.1 U R 1.2 U R1.3) U (R2.1 U R 2.2 U R2.3) U (R3.1 U R 3.2 U R3.3) U (R4.1 U R4.2 U R4.3) U (R5.1 U R5.2 U R5.3)	5	.6	3
Compensatory (+)	[R1.1 U R1.2] + [R2.1 U R 2.2] + [R3.1 U R 3.2] + [R4.1 U R 4.2] + [R5.1 U R 5.2]			2
	[R1.1 U R1.2 U R1.3] + [R2.1 U R2.2 U R2.3] + [R3.1 U R3.2 U R3.3] + [R4.1 U R4.2 U R4.3] + [R5.1 U R5.2 U R5.3]	5	.6	3
Conjunctive-Complementary (AND/OR)	(R1.1 U R 1.2) ∩ (R2.1 U R 2.2) ∩ ([R3.1 U R 3.2] U [R4.1 U R 4.2] U [R5.1 U R 5.2])			2
	(R1.1 U R 1.2 U R1.3) ∩ (R2.1 U R 2.2 U R2.3) ∩ ([R3.1 U R 3.2 U R3.3] U [R4.1 U R 4.2 U R4.3] U [R5.1 U R 5.2 U R5.3])	5	.6	3
Conjunctive-Compensatory (AND/+)	([R1.1 U R1.2] ∩ [R2.1 U R 2.2] ∩ ([R1.1 U R1.2] + [R2.1 U R 2.2] + [R3.1 U R 3.2] + [R4.1 U R 4.2] + [R5.1 U R 5.2])			2
	[R1.1 U R1.2 U R1.3] ∩ [R2.1 U R 2.2 U R2.3] ∩ ([R1.1 U R1.2 U R1.3] + [R2.1 U R 2.2 U R 2.3] + [R3.1 U R 3.2 U R3.3] + [R4.1 U R 4.2 U R4.3] + [R5.1 U R 5.2 U R5.3])	5	.6	3

Chapter 5: Simulation Results

Presentation of simulation results begins with a description of the characteristics of the simulated datasets, followed by results for Research Questions 1, 2, and 3. For each research question, results are presented first for the conditions in which 50% of examinees pass the individual tests, followed by results for the 70% passing rate. Contingency tables used to calculate passing rates and all measures of classification reliability appear in Appendix I. The properties of the simulated datasets are presented first, followed by the classification accuracy and reliability for the individual tests. Throughout the chapter, tables showing classification consistency are shaded to differentiate them from those for classification accuracy.

Description of Simulated Datasets

Means and standard deviations for the true and replicate score distributions for COVAR6 and COVAR9 appear in Tables 5.1, 5.2, and 5.3. The means for all distributions are quite close to the desired value of zero. The standard deviations for the true scores are all near to 1.0; and, as expected, the replicate scores have higher standard deviations due to the additional error variance. For COVAR6, the covariance among the five sets of true scores ranged from .599 to .601; covariance among the tests for each replicate group ranged from .598 to .603. For COVAR9, covariance among the set of true scores range from .899 to .901, and the covariance among replicate scores ranges from .899 to .902. The reliability of the resulting distributions is obtained by correlating the replicate scores for each test. For both COVAR6 and COVAR9, the reliability of the replicate scores is .909.

Table 5.1: Descriptive Statistics and Covariance for True Scores

COVAR6						
Test	Mean	1	2	3	4	5
1	0.00069	0.99882	0.60013	0.59847	0.60129	0.60068
2	0.00162	0.60013	0.99941	0.59890	0.60013	0.60009
3	0.00156	0.59847	0.59890	0.99850	0.60002	0.59949
4	0.00218	0.60129	0.60013	0.60002	1.00313	0.60238
5	0.00201	0.60068	0.60009	0.59949	0.60238	1.00208
COVAR9						
Test	Mean	1	2	3	4	5
1	-0.00199	1.00107	0.90079	0.90132	0.89968	0.90020
2	-0.00189	0.90079	1.00074	0.90108	0.89976	0.89999
3	-0.00201	0.90132	0.90108	1.00153	0.90017	0.90031
4	-0.00138	0.89968	0.89976	0.90017	0.99883	0.89937
5	-0.00211	0.90020	0.89999	0.90031	0.89937	0.99945

Table 5.2: Descriptive Statistics and Covariance for Six Sets of Replicate Scores, COVAR6

Replicate 1						
Test	Mean	1	2	3	4	5
1	0.00105	1.10029	0.60000	0.59878	0.60234	0.60131
2	0.00178	0.60000	1.09912	0.59826	0.60044	0.59995
3	0.00109	0.59878	0.59826	1.09892	0.60023	0.60074
4	0.00236	0.60234	0.60044	0.60023	1.10480	0.60285
5	0.00185	0.60131	0.59995	0.60074	0.60285	1.10280
Replicate 2						
Test	Mean	1	2	3	4	5
1	0.00075	1.09813	0.59888	0.59828	0.60031	0.60116
2	0.00075	0.59888	1.09946	0.59996	0.59977	0.60022
3	0.00144	0.59828	0.59996	1.09927	0.60068	0.59953
4	0.00223	0.60031	0.59977	0.60068	1.10284	0.60290
5	0.00181	0.60116	0.60022	0.59953	0.60290	1.10311
Replicate 3						
Test	Mean	1	2	3	4	5
1	0.00080	1.09885	0.59999	0.59890	0.60101	0.60074
2	0.00247	0.59999	1.09887	0.59856	0.59962	0.60017
3	0.00138	0.59890	0.59856	1.09895	0.59977	0.59979
4	0.00169	0.60101	0.59962	0.59977	1.10251	0.60253
5	0.00166	0.60074	0.60017	0.59979	0.60253	1.10320
Replicate 4						
Test	Mean	1	2	3	4	5
1	0.00047	1.09895	0.60134	0.59908	0.60136	0.60086
2	0.00132	0.60134	1.09957	0.59897	0.59865	0.60008
3	0.00143	0.59908	0.59897	1.09887	0.59958	0.59998
4	0.00231	0.60136	0.59865	0.59958	1.10311	0.60213
5	0.00120	0.60086	0.60008	0.59998	0.60213	1.10264
Replicate 5						
Test	Mean	1	2	3	4	5
1	0.00057	1.09951	0.60053	0.59849	0.60184	0.60079
2	0.00191	0.60053	1.09937	0.59861	0.60069	0.60060
3	0.00157	0.59849	0.59861	1.09854	0.60115	0.59944
4	0.00192	0.60184	0.60069	0.60115	1.10498	0.60316
5	0.00227	0.60079	0.60060	0.59944	0.60316	1.10226
Replicate 6						
Test	Mean	1	2	3	4	5
1	0.00016	1.09847	0.59928	0.59893	0.60099	0.60077
2	0.00192	0.59928	1.09777	0.59870	0.59996	0.59902
3	0.00206	0.59893	0.59870	1.09908	0.60074	0.59966
4	0.00259	0.60099	0.59996	0.60074	1.10280	0.60252
5	0.00238	0.60077	0.59902	0.59966	0.60252	1.10195

Table 5.3: Descriptive Statistics and Covariance for Three Sets of Replicate Scores, COVAR9

Replicate 1						
Test	Mean	1	2	3	4	5
1	-0.00199	1.10010	0.90067	0.90047	0.89850	0.89949
2	-0.00189	0.90067	1.10116	0.90089	0.90019	0.90014
3	-0.00201	0.90047	0.90089	1.10129	0.89952	0.89997
4	-0.00138	0.89850	0.90019	0.89952	1.09817	0.89863
5	-0.00211	0.89949	0.90014	0.89997	0.89863	1.09896
Replicate 2						
Test	Mean	1	2	3	4	5
1	-0.00164	1.10285	0.90176	0.90198	0.90080	0.90053
2	-0.00174	0.90176	1.10135	0.90119	0.89989	0.89985
3	-0.00248	0.90198	0.90119	1.10019	0.89992	0.89898
4	-0.00120	0.90080	0.89989	0.89992	1.09962	0.89886
5	-0.00227	0.90053	0.89985	0.89898	0.89886	1.09893
Replicate 3						
Test	Mean	1	2	3	4	5
1	-0.00193	1.10048	0.90112	0.90126	0.89970	0.89904
2	-0.00157	0.90112	1.10095	0.90170	0.90072	0.89960
3	-0.00214	0.90126	0.90170	1.10257	0.90095	0.90008
4	-0.00133	0.89970	0.90072	0.90095	1.09994	0.89939
5	-0.00231	0.89904	0.89960	0.90008	0.89939	1.09802

Individual Tests

The classification accuracy and consistency for individual tests serves as the baseline for considering the impact of combining additional tests. Because the five tests were all generated to have similar means and standard deviations, a comparison of the various measures of agreement for true and replicate scores also informs the precision of such measures. Table 5.4 presents the measures of accuracy for all five tests for COVAR6 at the 50% passing rate. The standard deviation of each statistic serves as an estimate of its standard error.

The mean exact agreement for single tests is 90.3%, and a similar level of conditional agreement is demonstrated. The mean percentage of false positives and false negatives is 4.88% and 4.87% respectively.

Table 5.4: Accuracy for Individual Tests, COVAR6, 50%

Scores	Exact Agreement	Kappa	Condition. Agree True Master	Conditional Agree, True Non-Master	False Positives	False Negatives
Test 1: True Score With -						
Rep 1	90.30	80.60	90.31	90.29	4.86	4.84
Rep 2	90.26	80.51	90.27	90.24	4.88	4.86
Rep 3	90.27	80.53	90.23	90.31	4.85	4.89
Rep 4	90.30	80.61	90.32	90.28	4.86	4.84
Rep 5	90.32	80.64	90.29	90.34	4.83	4.85
Rep 6	90.22	80.43	90.23	90.20	4.90	4.88
Test 2: True Score With -						
Rep 1	90.22	80.45	90.23	90.22	4.89	4.89
Rep 2	90.17	80.34	90.13	90.21	4.89	4.94
Rep 3	90.22	80.43	90.21	90.22	4.89	4.90
Rep 4	90.27	80.53	90.28	90.25	4.87	4.86
Rep 5	90.27	80.54	90.28	90.26	4.87	4.86
Rep 6	90.24	80.48	90.26	90.23	4.88	4.87
Test 3: True Score With -						
Rep 1	90.33	80.66	90.31	90.35	4.82	4.85
Rep 2	90.27	80.53	90.27	90.27	4.86	4.87
Rep 3	90.27	80.53	90.26	90.28	4.86	4.88
Rep 4	90.34	80.67	90.26	90.41	4.79	4.87
Rep 5	90.24	80.49	90.21	90.28	4.86	4.90
Rep 6	90.27	80.54	90.31	90.24	4.88	4.85
Test 4: True Score With -						
Rep 1	90.23	80.46	90.24	90.23	4.88	4.89
Rep 2	90.24	80.47	90.28	90.20	4.89	4.87
Rep 3	90.13	80.25	90.17	90.08	4.95	4.92
Rep 4	90.19	80.37	90.25	90.12	4.93	4.88
Rep 5	90.26	80.52	90.30	90.22	4.88	4.86
Rep 6	90.24	80.48	90.29	90.19	4.90	4.86
Test 5: True Score With -						
Rep 1	90.25	80.50	90.27	90.23	4.89	4.87
Rep 2	90.25	80.51	90.29	90.22	4.89	4.86
Rep 3	90.36	80.72	90.42	90.31	4.84	4.79
Rep 4	90.22	80.43	90.20	90.24	4.88	4.90
Rep 5	90.22	80.44	90.33	90.11	4.94	4.84
Rep 6	90.22	80.43	90.28	90.15	4.92	4.86
Mean	90.25	80.50	90.27	90.24	4.88	4.87
StdDev	0.049	0.099	0.054	0.070	0.034	0.028

Measures of consistency for the six replicate scores for Test 1 appear in Table 5.5. As expected, the measures of consistency are lower than similar measures of

accuracy since both scores incorporate error in the case of consistency. The mean exact agreement in consistency is 86.4% (compared to 90.3% for accuracy), and the measures of conditional agreement are also lower than the corresponding accuracy measures. The increased error incorporated into replicate scores is also reflected in a higher percentage of examinees who pass one test but fail the other (a corollary to false negatives and false positives in tables showing accuracy) -- 6.8% for consistency compared to 4.9% for accuracy.

Table 5.5: Measures of Consistency for Test 1, COVAR6, 50%

Replicates	Exact Agreement	Kappa	Pass Rep1 and Fail Rep 2	Specific Agreement, Masters	Specific Agreement, Nonmasters
Rep 1 & Rep 2	86.41	72.82	6.79	86.41	86.41
Rep 1 & Rep 3	86.41	72.82	6.82	86.40	86.41
Rep 1 & Rep 4	86.41	72.83	6.79	86.42	86.41
Rep 1 & Rep 5	86.39	72.79	6.82	86.39	86.40
Rep 1 & Rep 6	86.36	72.71	6.82	86.36	86.35
Rep 2 & Rep 3	86.33	72.66	6.86	86.33	86.34
Rep 2 & Rep 4	86.38	72.77	6.81	86.39	86.38
Rep 2 & Rep 5	86.36	72.72	6.84	86.36	86.36
Rep 2 & Rep 6	86.31	72.62	6.85	86.31	86.30
Rep 3 & Rep 4	86.32	72.65	6.81	86.32	86.33
Rep 3 & Rep 5	86.37	72.75	6.80	86.37	86.38
Rep 3 & Rep 6	86.34	72.67	6.80	86.33	86.34
Rep 4 & Rep 5	86.44	72.88	6.80	86.44	86.44
Rep 4 & Rep 6	86.37	72.75	6.81	86.38	86.37
Rep 5 & Rep 6	86.32	72.64	6.82	86.32	86.32
Mean	86.37	72.74	6.82	86.37	86.37
StdDev	0.0384	0.0768	0.0193	0.0391	0.0381

Descriptive statistics presented for the individual tests indicate that the simulations were successful in generating data with the desired means, standard deviations, and covariances for a set of highly reliable tests. The sample size of 500,000 in each condition contributes to a high level of precision for estimates. The mean exact agreement in regard to accuracy and consistency is similar to that

obtained in previous simulation studies cited in Chapter 2 (Bradlow & Wainer, 1998; Klein & Orlando, 2000; Rudner, 2001).

Another test of the proposed method is shown by calculating the percentage of examinees passing the conjunctive rule using a dataset generated in a similar fashion but with covariance equal to zero among the tests. Probability theory states that the probability of passing a series of uncorrelated tests is equal to the product of the probabilities of passing each individual test. It follows that for a set of five tests, each with a probability of passing of .50, the probability of passing two uncorrelated tests is .25; for three tests, the probability reduces to .125, four tests to .0625, and five tests to .03125. The conjunctive rule was applied to the COVAR0 dataset combining 2, 3, 4, and 5 tests. The resulting proportion of examinees who passed the decision rule was .249, .125, .062, and .031 respectively. This example supports the utility of the proposed method for other types of rules and for use with related tests.

Research Question 1: Increasing the Number of Tests

This research question investigates the effect of increasing the number of tests on the classification accuracy and consistency of simple and complex decisions. Results are first presented for conditions in which 50% of examinees pass each test; then for conditions in which 70% of examinees pass; and then an overall summary of the results for the research question follows.

Fifty Percent Passing Rate

Table 5.6 shows the percentage of examinees that passed each type of decision rule based on true score.

Table 5.6: Percentage Passing Based on True Score, COVAR6, 50%

	Number of Tests				
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	50.00	35.28	27.91	23.36	20.27
Complementary	50.00	64.74	72.14	76.69	79.74
Compensatory	50.00	50.02	50.02	50.06	50.11
Conj-Complem	50.00			32.43	33.88
Conj-Compens	50.00		23.65		18.94

The results for a single test are presented first in each row to provide a point of comparison[4]. The passing rate decreases with each additional test in the conjunctive rule, from 50.0% to 20.3%, whereas the passing rate increases in the complementary rule to 79.8%. The passing rate remains stable at around 50% with the addition of more tests in the compensatory rule. The direction of these changes in the percentage passing the simple rules follows expectations. The conjunctive rule becomes more stringent with each additional test, whereas the complementary rule allows more opportunities to pass with additional tests. In the compensatory rule each additional test contributes equally to passage of the overall rule, and therefore the passing rate does not change.

The addition of more tests in the complex rules combines the features of these simple rules toward less predictable outcomes. The conjunctive-complementary rule becomes both more stringent (the examinee must pass more tests) and less stringent (there are more ways to pass) with additional tests. The net effect is that fewer examines pass when comparing the four- and five-test conditions to a single test. Similarly, the conjunctive-compensatory rule combines the requirements of the

[4] For individual tests, the mean value over all five tests is presented. This value is the same for all decision rules in each condition.

conjunctive rule with the stable effect of adding tests in a compensatory manner. In this case, the conjunctive-compensatory rule was structured to require a minimum score and a more stringent overall average score. The result is that fewer examinees pass this rule than either of the simple conjunctive or compensatory rule, and the percentage passing the conjunctive-compensatory rule declines with the addition of more tests. In fact, the conjunctive-compensatory rule yields the lowest passing rate in Table 5.6, with only 18.9% passing the five test rule.

Table 5.7 shows the percentage who pass based on replicate scores as more tests are included in the decision rule. Passing rates show a similar pattern as that seen for accuracy, but at reduced levels. Only 17.3% pass the conjunctive-compensatory rule when comparing two replicate scores.

Table 5.7: Percentage Passing Based on Two Replicate Scores, COVAR6, 50%

	Number of Tests				
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	50.00	34.25	26.32	21.50	18.26
Complementary	50.00	65.79	73.69	78.52	81.72
Compensatory	50.00	50.07	50.07	50.12	50.12
Conj-Complem	50.00			31.14	32.73
Conj-Compens	50.00		22.81		17.33

As described in the Chapter 4, there are a number of measures for characterizing classification reliability in contingency tables and an optimal measure for all purposes has not been clearly identified. It is possible that the choice of measure depends on the primary purpose of the decision. Furthermore, all measures are percentages and comparisons are complicated by differences in the percentage passing the decision rules. As shown in Tables 5.6 and 5.7, there are substantial differences among the decision rules in the percentage passing as more tests are added to the rule. Conditional measures that estimate the reliability for Masters and

Non-Masters separately are less affected by the differences in marginal distributions, and may provide a more consistent pattern of results. The following measures of classification accuracy compare decision outcomes based on true and replicate scores for the five decision rules.

- *Exact Agreement* is the 'hit rate'; that is, the percentage of examinees who received the correct classification as Masters or Non-Masters. There is no established value of acceptable levels of exact agreement across all decision purposes. Subkoviak (1988) suggests 85% as the lower bound of acceptable exact agreement for decisions with important outcomes. As the percentage of Masters increases, which is more typical in mastery tests, Subkoviak advises expecting higher levels of exact agreement.
- *Kappa* is a measure of agreement that has been adjusted for the likelihood of obtaining the same classification purely by chance. It can be interpreted as the *additional* contribution of true score in predicting replicate score over and above the prediction that would be made based on chance (i.e., the overall proportion of passers). Similar to exact agreement, acceptable levels for Kappa vary according to the decision context. Subkoviak (1988) illustrates that Kappa *decreases* as the percentage of Masters increases. He suggests that Kappa's between 60% and 70% are acceptable for decisions in which 50% of examinees are Masters, and adjusts the acceptable level of Kappa to 65% when 90% are Masters.

- *Conditional Agreement for Masters* is the percentage of Masters who are **correctly** classified. Since the denominator for this percentage is the marginal total, it is less affected by the overall percentage passing the rule.
- *Conditional Agreement for Non-Masters* is the percentage of Non-Masters who are **correctly** classified. This percentage is also less affected by differences in the percentage passing the rule.
- *False Negatives* estimate the percentage of all examinees who are **incorrectly** classified as Non-Masters.
- *False Positives* estimate the percentage of all examinees who are **incorrectly** classified as Masters.

The following measures of consistency compare decision outcomes for two sets of replicate scores. As in the case of accuracy, measures that use the total number of examinees in the denominator (exact agreement and percentage of examinees who receive different decisions based on the two sets of replicate scores) are most affected by differences in the percentage passing the overall decision rules.

- *Exact Agreement* is the percentage of examinees who receive the same classification on both sets of replicate scores.
- *Kappa* estimates the agreement between classifications on two replicate scores adjusting for the marginal distribution.
- *Pass 1, Fail 2* is the percentage of examinees who passed on the basis of one set of replicate scores, but not on the other. It is similar to the false negative and false positive measure used to describe classification accuracy.

- *Specific agreement for Passers* is the number of examinees who pass the decision rule on the basis of both sets of replicate scores divided by the number who pass on either set of scores.
- *Specific agreement for Failers* is the number of examinees who fail the decision rule on the basis of both sets of scores divided by the number who fail on either set of scores.

Criteria for Comparing Measures of Agreement

Given the number of comparisons made possible by the large number of conditions in this simulation study, and the computational intensity required to calculate standard errors for each measure in each condition, a more general approach was taken in characterizing differences between measures of agreement. Standard errors estimated for the measures of agreement for individual tests were used to estimate confidence intervals for each measure. Differences among the measures are highlighted when they differ by more than four standard errors (analogous to non-overlapping 95% confidence intervals). To facilitate the reading of results tables, the decision rule label is shown in **bold typeface** if the decision shows an **increase** in reliability across the conditions in the table. The decision rule label is shown in *italicized typeface* if the decision shows a **decrease** in reliability. Note that measures of false negatives, false positives, and Pass 1/Fail 2 show *increased* classification reliability when the estimates *decrease*.

Results for the accuracy of classification as more tests are added to the decision rule appear in Table 5.8.

Table 5.8: Accuracy for One to Five Tests, COVAR6, 50%

Exact Agreement					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	90.25	90.80	91.70	92.46	93.10
Complementary	90.25	90.84	91.78	92.57	93.17
Compensatory	90.25	92.21	93.38	94.09	94.56
Conj-Complem	90.25			91.17	91.02
Conj-Compens	90.25		93.61		94.06
Kappa					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	80.50	79.71	79.02	78.34	77.84
Complementary	80.50	79.81	79.20	78.62	78.07
Compensatory	80.50	84.42	86.75	88.17	89.12
Conj-Complem	80.50			79.63	79.80
Conj-Compens	80.50		82.09		80.01
False Negatives					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	4.87	5.12	4.95	4.70	4.45
Complementary	4.87	4.06	3.34	2.80	2.42
Compensatory	4.87	3.87	3.29	2.93	2.73
Conj-Complem	4.87			5.07	5.06
Conj-Compens	4.87		3.61		3.78
False Positives					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	4.88	4.09	3.35	2.84	2.44
Complementary	4.88	5.10	4.88	4.63	4.41
Compensatory	4.88	3.92	3.34	2.99	2.72
Conj-Complem	4.88			3.77	3.91
Conj-Compens	4.88		2.78		2.16
Conditional Agreement for True Masters					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	90.27	85.50	82.28	79.87	78.03
Complementary	90.27	93.73	95.37	96.35	96.96
Compensatory	90.27	92.27	93.43	94.15	94.56
Conj-Complem	90.27			84.38	85.06
Conj-Compens	90.27		84.73		80.06
Conditional Agreement for True Non-Masters					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	90.24	93.69	95.35	96.30	96.93
Complementary	90.24	85.53	82.47	80.13	78.26
Compensatory	90.24	92.15	93.33	94.02	94.56
Conj-Complem	90.24			94.42	94.08
Conj-Compens	90.24		96.36		97.33

Table 5.8 shows that, with the exception of the compensatory rule, different patterns are found for various measures of agreement as the number of tests incorporated into the decision rule increases. Exact agreement increases with the addition of more tests for all decision rules, with the largest increase seen for the compensatory rule (94.6%). Kappa, however, decreases consistently with the addition of more tests for all but the two rules that incorporate a compensatory rule. The addition of more tests decreases the percentage of false positives for all rules, but variable effects are found for false negatives.

A clear picture emerges when considering measures of conditional agreement. Once again the compensatory rule increases reliability for all decision rules. For Masters, the conjunctive rule shows *decreased* classification reliability with the addition of more tests, and the complementary rule shows *increased* reliability. The reverse is seen for Non-Masters – the conjunctive rule shows *increased* reliability and the complementary rule shows *decreased* reliability. The complex rules show similar results to those for the simple conjunctive rule.

Measures of consistency show similar patterns to those for accuracy, but at somewhat lower levels.

Table 5.9: Consistency for One to Five Tests, COVAR6, 50%

Exact Agreement					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	86.37	87.45	88.99	90.25	91.28
Complementary	86.37	87.42	89.03	90.30	91.31
Compensatory	86.37	89.06	90.53	91.61	92.36
Conj-Complem	86.37			88.17	87.91
Conj-Compens	86.37		91.24		92.26
Kappa					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	72.74	72.13	71.60	71.10	70.77
Complementary	72.74	72.05	71.70	71.23	70.86
Compensatory	72.74	78.13	81.06	83.22	84.71
Conj-Complem	72.74			72.41	72.52
Conj-Compens	72.74		75.14		72.95
Pass Rep 1, Fail Rep 2					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	6.82	6.30	5.51	4.89	4.38
Complementary	6.82	6.28	5.45	4.81	4.31
Compensatory	6.82	5.51	4.76	4.22	3.83
Conj-Complem	6.82			5.93	6.09
Conj-Compens	6.82		4.38		3.89
Specific Agreement for Passers					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	86.37	81.66	79.07	77.31	76.10
Complementary	86.37	90.44	92.56	93.83	94.68
Compensatory	86.37	89.07	90.54	91.62	92.37
Conj-Complem	86.37			80.99	81.50
Conj-Compens	86.37		80.81		77.63
Specific Agreement for Failers					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	86.37	90.46	92.53	93.79	94.67
Complementary	86.37	81.61	79.13	77.40	76.18
Compensatory	86.37	89.06	90.52	91.59	92.34
Conj-Complem	86.37			91.41	91.02
Conj-Compens	86.37		94.33		95.32

Seventy Percent Passing Rate

The following tables present results for a passing criterion set to yield a 70% passing rate for each individual test. This higher rate of passage is reflected in Tables 5.10 and 5.11. The conjunctive-compensatory rule yields the lowest passing rate (39.3%) based on true score, compared to 18.9% in the corresponding 50% passing rate condition.

Table 5.10: Percentage Passing Based on True Score, COVAR6, 70%

	Number of Tests				
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	70.06	57.37	49.81	44.63	40.79
Complementary	70.06	82.76	87.95	90.73	92.45
Compensatory	70.06	72.17	73.03	73.52	73.79
Conj-Complem	70.06			54.99	56.33
Conj-Compens	70.06		45.23		39.32

Table 5.11: Percentage Passing Based on Two Replicate Scores, COVAR6, 70%

	Number of Tests				
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	69.21	55.43	47.14	41.52	37.42
Complementary	69.21	82.98	88.55	91.48	93.28
Compensatory	69.21	71.59	72.59	73.15	73.50
Conj-Complem	69.21			52.80	55.43
Conj-Compens	69.21		43.42		36.41

Table 5.12 presents the estimated accuracy of decision rules in the 70% passing condition. Whereas in the 50% condition exact agreement increased with the addition of more tests in all five rules, now there are differences in the rules in regard to the impact of adding more tests. The simple conjunctive rule, and the two complex rules that incorporate a conjunctive element, show stable or slight decreases in exact agreement as they incorporate more tests. The complementary and compensatory rules show increased exact agreement. Conditional measures show consistent results with those found in the 50% condition.

Table 5.12: Accuracy for One to Five Tests, COVAR6, 70%

Exact Agreement					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	91.41	90.00	89.79	89.84	90.05
Complementary	91.41	93.72	95.15	96.03	96.65
Compensatory	91.41	93.37	94.46	95.08	95.55
Conj-Complem	91.41			90.02	89.57
Conj-Compens	91.41		91.47		91.05
Kappa					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	79.69	79.68	79.57	79.31	79.14
Complementary	79.69	77.89	76.61	75.54	74.76
Compensatory	79.69	83.61	86.02	87.43	88.53
Conj-Complem	79.69			79.93	78.86
Conj-Compens	79.69		82.71		81.00
False Negatives					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	4.73	5.97	6.44	6.63	6.66
Complementary	4.73	3.03	2.13	1.61	1.26
Compensatory	4.73	3.60	2.99	2.64	2.37
Conj-Complem	4.73			6.08	5.66
Conj-Compens	4.73		5.17		5.93
False Positives					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	3.86	4.03	3.77	3.53	3.29
Complementary	3.86	3.25	2.73	2.36	2.08
Compensatory	3.86	3.03	2.55	2.27	2.08
Conj-Complem	3.86			3.90	4.76
Conj-Compens	3.86		3.36		3.02
Conditional Agreement for True Masters					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	93.24	89.59	87.07	85.14	83.67
Complementary	93.24	96.34	97.58	98.23	98.63
Compensatory	93.24	95.01	95.91	96.40	96.79
Conj-Complem	93.24			88.94	89.95
Conj-Compens	93.24		88.56		84.92
Conditional Agreement for True Non-Masters					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	87.11	90.55	92.48	93.63	94.45
Complementary	87.11	81.16	77.38	74.56	72.39
Compensatory	87.11	89.13	90.56	91.42	92.05
Conj-Complem	87.11			91.35	89.09
Conj-Compens	87.11		93.86		95.02

For the 70% passing rate condition, measures of consistency show similar results to those for accuracy.

Table 5.13: Consistency for One to Five Tests, COVAR6, 70%

Exact Agreement					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	87.89	86.36	86.37	86.78	87.28
Complementary	87.89	91.41	93.50	94.88	95.83
Compensatory	87.89	90.69	92.16	93.04	93.78
Conj-Complem	87.89			86.55	86.36
Conj-Compens	87.89		88.28		88.26
Kappa					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	71.60	72.39	72.66	72.78	72.85
Complementary	71.60	69.57	67.93	67.11	66.72
Compensatory	71.60	77.11	80.30	82.29	84.03
Conj-Complem	71.60			73.01	72.39
Conj-Compens	71.60		76.15		74.65
Pass Rep 1, Fail Rep 2					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	6.04	6.83	6.79	6.60	6.36
Complementary	6.04	4.28	3.24	2.55	2.07
Compensatory	6.04	4.66	3.92	3.47	3.10
Conj-Complem	6.04			6.73	6.83
Conj-Compens	6.04		5.85		5.89
Specific Agreement for Passers					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	91.25	87.69	85.55	84.09	83.01
Complementary	91.25	94.83	96.33	97.20	97.77
Compensatory	91.25	93.50	94.60	95.25	95.77
Conj-Complem	91.25			87.26	87.69
Conj-Compens	91.25		86.50		83.87
Specific Agreement for Failers					
	ONE	TWO	THREE	FOUR	FIVE
Conjunctive	80.35	84.70	87.10	88.70	89.84
Complementary	80.35	74.75	71.60	69.91	68.95
Compensatory	80.35	83.61	85.70	87.04	88.26
Conj-Complem	80.35			85.75	84.70
Conj-Compens	80.35		89.64		90.78

Summary of Results for Research Question 1

A summary of the overall trends in measures of classification reliability for the 50% and 70% conditions appears in Tables 5.14 and 5.15. These tables highlight the similarity of results for conditional measures of classification reliability across the five decision rules in contrast to those based on other measures.

The addition of more tests to the decision rule has a large impact on the percentage of examinees who pass all decision rules, except the compensatory rule. For the conditions in which only 50% of examinees pass each individual test, only 20% pass the five-test conjunctive rule. In contrast, the five test complementary rule yields an 80% passing rate. Only the compensatory rule maintains the same passing rate as more tests are added to the decision rule. As expected, the addition of a complementary component to the conjunctive rule raises the five-test passing rate (34%), but the addition of the compensatory rule to the conjunctive rule slightly decreases the passing rate (19%).

Given the variability in results for different measures of accuracy and consistency across different types of decision rules, what can be said about the effect of adding more tests to decision rules?

- Measures of accuracy and consistency provide similar results for most, but not all, conditions.
- The compensatory rule shows increased classification reliability for all measures and all decision rules with the addition of more tests. In the five-test rule, the percentage of agreement is 95% (compared to 90% for one test).

- For Masters, adding more tests *decreases* classification reliability when combined in a conjunctive manner, and the same is true when complementary and compensatory rules are used in combination with the conjunctive rule. The opposite is true for the complementary rule, for which classification reliability *increases* with the addition of more tests. The complementary rule yields the highest agreement for Masters (97%). The conjunctive rule yields the lowest agreement for Masters (78%).
- For Non-Masters, classification reliability shows an opposite pattern of results to that for Masters. The addition of more tests *increases* reliability in the conjunctive rule and complex rules that include the conjunctive rule, but *decreases* reliability in the complementary rule. The conjunctive rule, and the complex conjunctive-compensatory rule, yield the highest agreement for Non-Masters at 97%. The complementary rule yields the lowest agreement for Non-Masters at (78%).
- The impact of adding more tests shows a similar pattern of results for both levels of test difficulty, but the level of accuracy differs by test difficulty and number of tests. For a single test, the easier test (70% passing) has higher accuracy for overall agreement and conditional agreement for Masters than the more difficult test (50% passing), but lower accuracy for Non-Masters. For the five-test rules, similar results are found for the conditional comparisons, but exact agreement is higher for rules with a conjunctive element on the more difficult test, and lower for the other rules.

Table 5.14: Effect of Adding More Tests on Classification Reliability, COVAR6, 50%

	Exact Agreement		Kappa		**Conditional Agreement|Master***		**Conditional Agreement|Non-Master***		False Negative**	False Positive**	Pass 1/ Fail 2**
	ACC	CON	ACC	CON	**ACC**	**CON**	**ACC**	**CON**	ACC	ACC	CON
Conjunctive	H	H	L	L	**L**	**L**	**H**	**H**	~	H	H
Complementary	H	H	L	L	**H**	**H**	**L**	**L**	H	H	H
Compensatory	H	H	H	H	**H**	**H**	**H**	**H**	H	H	H
Conj-Complem	H	H	L	~	**L**	**L**	**H**	**H**	L	H	H
Conj-Compens	H	H	~	~	**L**	**L**	**H**	**H**	H	H	H
H indicates increased reliability; L indicates decreased reliability; ~ indicates no or mixed effect											

*For CON, this column indicates Specific Agreement for Passers and Failers respectively.
**Higher reliability is associated with lower percentage of false negatives, false positives, and Pass 1/Fail 2.

Table 5.15: Effect of Adding More Tests on Classification Reliability, COVAR6, 70%

	Exact Agreement		Kappa		**Conditional Agreement|Master***		**Conditional Agreement|Non-Master***		False Negative**	False Positive**	Pass 1/ Fail 2**
	ACC	CON	ACC	CON	**ACC**	**CON**	**ACC**	**CON**	ACC	ACC	CON
Conjunctive	L	L	L	H	**L**	**L**	**H**	**H**	L	H	~
Complementary	H	H	L	L	**H**	**H**	**L**	**L**	H	H	H
Compensatory	H	H	H	H	**H**	**H**	**H**	**H**	H	H	H
Conj-Complem	L	L	~	H	**L**	**L**	**H**	**H**	~	L	L
Conj-Compens	L	H	H	H	**L**	**L**	**H**	**H**	L	H	H
H indicates increased reliability; L indicates decreased reliability; ~ indicates no or mixed effect											

*For CON, this column indicates Specific Agreement for Passers and Failers respectively.
**Higher reliability is associated with lower percentage of false negatives, false positives, and Pass 1/Fail 2.

Research Question 2: Varying the Covariance Among Tests

For this question, results for the five-test condition for each decision rule are compared for COVAR6 and COVAR9 to investigate the effect of the strength of relationship among tests on classification reliability.

Fifty Percent Passing Rate

The percentage of examinees who passed each rule is shown in Table 5.16. The compensatory rule shows similar passing rates for each level of covariance, as was also the case when the number of tests was increased in Research Question 1. For the conjunctive rule the passing rate based on true score increases with higher covariance from 20.3% to 35.3%, whereas the passing rate decreases in the similar comparison for the complementary rule (from 79.7% to 64.7%).

Table 5.16: Passing Rate for COVAR6 and COVAR9, 50%, Five Tests

	COVAR6, 50%		COVAR9, 50%	
	TRUE SCORE	REPLICATE	TRUE SCORE	REPLICATE
Conjunctive	20.27	18.26	35.25	30.05
Complementary	79.74	81.72	64.67	69.91
Compensatory	50.12	50.11	49.88	49.88
Conj-Complem	33.88	32.73	42.13	39.33
Conj-Compens	18.94	17.33	29.36	27.13

Table 5.17 shows that the effect of increasing covariance on exact agreement, false negatives, and false positives depends on the decision rule. Kappa, and the conditional measures of agreement increase for all decision rules as covariance increases.

Table 5.17 : Accuracy for COVAR6 and COVAR9, Five Tests, 50%

	COVAR6	COVAR9
Exact Agreement		
Conjunctive	93.10	91.65
Complementary	93.17	91.60
Compensatory	94.56	95.33
Conj-Complem	91.02	91.43
Conj-Compens	94.06	94.55
Kappa		
Conjunctive	77.84	81.08
Complementary	78.07	80.98
Compensatory	89.12	90.65
Conj-Complem	79.80	82.26
Conj-Compens	80.01	86.57
False Negatives		
Conjunctive	4.45	6.77
Complementary	2.42	1.58
Compensatory	2.73	2.33
Conj-Complem	5.06	5.68
Conj-Compens	3.78	3.84
False Positives		
Conjunctive	2.44	1.57
Complementary	4.41	6.82
Compensatory	2.72	2.34
Conj-Complem	3.91	2.89
Conj-Compens	2.16	1.61
Conditional Agreement|Masters		
Conjunctive	78.03	80.79
Complementary	96.96	97.56
Compensatory	94.56	95.32
Conj-Complem	85.06	86.51
Conj-Compens	80.06	86.93
Conditional Agreement|Non-Masters		
Conjunctive	96.93	97.57
Complementary	78.26	80.70
Compensatory	94.56	95.33
Conj-Complem	94.08	95.01
Conj-Compens	97.33	97.72

Although exact agreement and kappa show similar results for measures of consistency to those found for accuracy, conditional measures of agreement show different effects. The conjunctive and complementary rules exhibit opposite patterns.

Specific agreement for Passers increases in the conjunctive rule, but decreases in the complementary rule. The opposite pattern is found for Failers.

Table 5.18: Consistency for COVAR6 and COVAR9, Five Tests, 50%

	COVAR6	COVAR9
Exact Agreement		
Conjunctive	91.28	90.46
Complementary	91.31	90.46
Compensatory	92.36	93.42
Conj-Complem	87.91	88.67
Conj-Compens	92.26	92.72
Kappa		
Conjunctive	70.77	77.30
Complementary	70.86	77.33
Compens	84.71	86.84
Conj-Complem	72.52	76.26
Conj-Compens	72.95	81.59
Pass Rep 1, Fail Rep 2		
Conjunctive	4.34	4.79
Complementary	4.31	4.80
Compensatory	3.82	3.27
Conj-Complem	6.09	5.67
Conj-Compens	3.89	3.64
Specific Agreement for Passers		
Conjunctive	76.10	84.11
Complementary	94.68	93.17
Compensatory	92.37	93.41
Conj-Complem	81.50	85.59
Conj-Compens	77.63	86.58
Specific Agreement for Failers		
Conjunctive	94.67	93.18
Complementary	76.18	84.15
Compensatory	92.34	93.44
Conj-Complem	91.02	90.66
Conj-Compens	95.32	95.01

Seventy Percent Passing Rate

Results comparing two levels of covariance with a 70% passing rate are similar to those for the 50% passing rate, but with more examinees passing each of the decision rules.

Table 5.19 : Passing Rates for COVAR6 and COVAR9, Five Tests, 70%

	COVAR6, 70%		COVAR9, 70%	
	TRUE	OBSERVED	TRUE	OBSERVED
Conjunctive	40.79	37.42	56.33	50.13
Complementary	92.45	93.28	82.05	85.32
Compensatory	73.79	73.50	70.72	70.54
Conj-Complem	56.33	55.43	63.14	59.78
Conj-Compens	39.32	36.41	50.06	46.99

Results for classification reliability are shown in Tables 5.20 and 5.21. The pattern of results is very similar to those for the 50% passing rate, but at generally higher levels of agreement.

Table 5.20: Accuracy for COVAR6 and COVAR9, Five Tests, 70%

	COVAR6	COVAR9
Exact Agreement		
Conjunctive	90.05	90.48
Complementary	96.65	94.43
Compensatory	95.55	95.95
Conj-Complem	89.57	91.31
Conj-Compens	91.05	93.41
Kappa		
Conjunctive	79.14	80.95
Complementary	74.76	79.64
Compensatory	88.53	90.24
Conj-Complem	78.86	81.67
Conj-Compens	81.00	86.82
False Negatives		
Conjunctive	6.66	7.86
Complementary	1.26	1.15
Compensatory	2.37	2.11
Conj-Complem	5.66	6.03
Conj-Compens	5.93	4.83
False Positives		
Conjunctive	3.29	1.66
Complementary	2.08	4.42
Compensatory	2.08	1.93
Conj-Complem	4.76	2.67
Conj-Compens	3.02	1.76
Conditional Agreement|Masters		
Conjunctive	83.67	86.05
Complementary	98.63	98.60
Compensatory	96.79	97.01
Conj-Complem	89.95	90.45
Conj-Compens	84.92	90.34
Conditional Agreement|Non-Masters		
Conjunctive	94.45	96.20
Complementary	72.39	75.37
Compensatory	92.05	93.40
Conj-Complem	89.09	92.77
Conj-Compens	95.02	96.48

Table 5.21 : Consistency for COVAR6 and COVAR9, Five Tests, 70%

	COVAR6	COVAR9
Exact Agreement		
Conjunctive	87.28	88.97
Complementary	95.83	93.84
Compensatory	93.78	94.28
Conj-Complem	86.36	88.43
Conj-Compens	88.26	91.21
Kappa		
Conjunctive	72.85	77.94
Complementary	66.72	75.42
Compens	84.03	86.23
Conj-Complem	72.39	75.95
Conj-Compens	74.65	82.35
Pass Rep 1, Fail Rep 2		
Conjunctive	6.36	5.51
Complementary	2.07	3.08
Compensatory	3.10	2.87
Conj-Complem	6.83	5.80
Conj-Compens	5.89	4.37
Specific Agreement for Passers		
Conjunctive	83.01	89.00
Complementary	97.77	96.39
Compensatory	95.77	95.94
Conj-Complem	87.69	90.32
Conj-Compens	83.87	90.65
Specific Agreement for Failers		
Conjunctive	89.84	88.94
Complementary	68.95	79.02
Compensatory	88.26	90.29
Conj-Complem	84.70	85.63
Conj-Compens	90.78	91.70

Summary of Results for Research Question 2

Findings for this research question yield less consistent patterns than those for Research Question 1 (Tables 5.22 and 5.23). One consistent result, however, is that the compensatory rule shows increased classification reliability and consistency for all rules.

The relationship between covariance and the percentage passing varies according to the decision rule. Whereas for the conjunctive rules, the percentage who pass is higher in the higher covariance condition, the reverse is found for the complementary rule.

- For comparisons of accuracy, measures of conditional agreement for both Masters and Non-Masters increase for all but the complementary rule (70% passing).
- For comparisons of consistency, measures of conditional agreement increase for Passers in the conjunctive rule, but decrease in the complementary rule. The opposite is true for Failers, in which measures of agreement decrease in the conjunctive rule but increase in the complementary rule.
- The pattern of results for the 70% passing rate are generally similar to those for the 50% passing rate in regard to accuracy. An exception is found for measures of consistency for Failers, which shows opposite results for the complex conjunctive rules based on test difficulty.
- As was the case when considering rules based on additional tests, test difficulty has differential impact on exact agreement for the five decision

rules, but similar results for conditional measures of agreement. For COVAR9, conditional agreement for Masters is lower for the more difficult test than the easier test, and conditional agreement for Non-Masters is higher on the more difficult test.

Table 5.22: Effect of Increasing Covariance, Five Tests, 50%

	Exact Agreement		Kappa		**Conditional Agreement|Master***		**Conditional Agreement|Non-Master***		False Negative**	False Positive**	Pass 1/ Fail 2**
	ACC	CON	ACC	CON	**ACC**	**CON**	**ACC**	**CON**	ACC	ACC	CON
Conjunctive	L	L	H	H	**H**	**H**	**H**	**L**	L	H	L
Complementary	L	L	H	H	**H**	**L**	**H**	**H**	H	L	L
Compensatory	H	H	H	H	**H**	**H**	**H**	**H**	H	H	H
Conj-Complem	H	H	H	H	**H**	**H**	**H**	**L**	L	H	H
Conj-Compens	H	H	H	H	**H**	**H**	**H**	**L**	~	H	H
H indicates increased reliability; L indicates decreased reliability; ~ indicates no or mixed effect											

*For CON, this column indicates Specific Agreement for Passers and Failers respectively.
**Higher reliability is associated with lower percentage of false negatives, false positives, and Pass 1/Fail 2.

Table 5.23: Effect of Increasing Covariance, Five Tests, 70%

	Exact Agreement		Kappa		**Conditional Agreement|Master***		**Conditional Agreement|Non-Master***		False Negative**	False Positive**	Pass 1/ Fail 2**
	ACC	CON	ACC	CON	**ACC**	**CON**	**ACC**	**CON**	ACC	ACC	CON
Conjunctive	H	H	H	H	**H**	**H**	**H**	**L**	L	H	H
Complementary	L	L	H	H	**L**	**L**	**H**	**H**	~	L	L
Compensatory	H	H	H	H	**H**	**H**	**H**	**H**	H	H	H
Conj-Complem	H	H	H	H	**H**	**H**	**H**	**H**	L	H	H
Conj-Compens	H	H	H	H	**H**	**H**	**H**	**H**	H	H	H
H indicates increased reliability; L indicates decreased reliability; ~ indicates no or mixed effect											

*For CON, this column indicates Specific Agreement for Passers and Failers respectively.
**Higher reliability is associated with lower percentage of false negatives, false positives, and Pass 1/Fail 2.

Research Question 3: Allowing Multiple Opportunities to Pass

The third research question investigates the effect of allowing examinees multiple attempts to pass each of five tests on the classification reliability of the five decision rules. The assumption is made that true score remains the same across all attempts. Results are presented first for the conditions in which 50% of examinees pass each individual test, and then for conditions in which 70% pass each individual test.

Fifty Percent Passing Rate

The percentage of examinees who passed on the basis of true score, as well as on the basis of replicate scores for one, two, or three attempts is shown in Table 5.24. As expected, for each of the five rules the percentage passing increases with each additional attempt. This is the case even for the compensatory rule which did not show increases in the percentage passing when the number of tests and the covariance was increased.

Table 5.24: Percentage Passing Multiple Attempts, COVAR6, Five Tests, 50%

		REPLICATE SCORE		
	TRUE SCORE	ONE ATTEMPT	TWO ATTEMPTS	THREE ATTEMPTS
Conjunctive	20.27	18.26	24.70	28.28
Complementary	79.74	81.72	86.10	88.06
Compensatory	50.12	50.11	58.56	62.67
Conj-Complem	33.88	32.73	40.21	44.14
Conj-Compens	18.94	17.33	23.60	27.11

The accuracy of classification reliability for each number of attempts is shown in Table 5.25. The results for each type of measure are quite consistent for all decision rules. Conditional agreement for Masters, as well as the percentage of false

positives, increases with more attempts; almost all other measures of reliability decrease for all the decision rules.

Table 5.25 : Accuracy For Multiple Attempts, COVAR6, Five Tests, 50%

	Number of Attempts		
Exact Agreement			
	ONE	TWO	THREE
Conjunctive	93.10	92.85	91.01
Complementary	93.20	92.53	91.35
Compensatory	94.56	91.26	87.44
Conj-Complem	91.02	90.68	88.67
Conj-Compens	94.06	93.27	91.13
Kappa			
	ONE	TWO	THREE
Conjunctive	77.84	79.54	75.76
Complementary	78.07	73.82	68.40
Compensatory	89.12	82.52	74.87
Conj-Complem	79.80	80.09	76.44
Conj-Compens	80.01	79.97	75.22
False Negatives			
	ONE	TWO	THREE
Conjunctive	4.45	1.36	0.49
Complementary	2.42	0.55	0.16
Compensatory	2.73	0.15	0.01
Conj-Complem	5.06	1.50	0.54
Conj-Compens	3.78	1.03	0.35
False Positives			
	ONE	TWO	THREE
Conjunctive	2.44	5.79	8.50
Complementary	4.41	6.92	8.48
Compensatory	2.72	8.59	12.55
Conj-Complem	3.91	7.83	10.80
Conj-Compens	2.16	5.70	8.52
Conditional Agreement|Masters			
	ONE	TWO	THREE
Conjunctive	78.03	93.28	97.58
Complementary	96.96	99.31	99.80
Compensatory	94.56	99.70	99.99
Conj-Complem	85.06	95.57	98.41
Conj-Compens	80.06	94.54	98.15
Conditional Agreement|Non-Masters			
	ONE	TWO	THREE
Conjunctive	96.93	92.74	89.34
Complementary	78.26	65.86	58.13
Compensatory	94.56	82.79	74.84
Conj-Complem	94.08	88.16	83.67
Conj-Compens	97.33	92.97	89.49

The consistency of classification shows a different pattern of results to those for accuracy. Kappa and specific agreement for Passers increases for all decision rules with additional attempts, but specific agreement for Failers decreases for the conjunctive rule and increases for complementary and compensatory rules.

Table 5.26 : Consistency for Multiple Attempts, COVAR6, Five Tests, 50%

	Number of Attempts		
Exact Agreement			
	ONE	TWO	THREE
Conjunctive	91.28	91.15	91.32
Complementary	91.31	93.77	94.84
Compensatory	92.36	93.86	94.62
Conj-Complem	87.91	89.03	89.78
Conj-Compens	92.26	92.13	92.29
Kappa			
	ONE	TWO	THREE
Conjunctive	70.77	76.23	78.61
Complementary	70.86	73.99	75.42
Compensatory	84.71	87.34	88.51
Conj-Complem	72.52	77.18	79.27
Conj-Compens	72.95	78.19	80.52
Pass Rep 1, Fail Rep 2			
	ONE	TWO	THREE
Conjunctive	4.34	4.40	4.29
Complementary	4.31	3.12	2.57
Compensatory	3.82	3.08	2.69
Conj-Complem	6.09	5.51	5.12
Conj-Compens	3.89	3.91	3.82
Specific Agreement for Passers			
	ONE	TWO	THREE
Conjunctive	76.10	82.11	84.67
Complementary	94.68	96.38	97.07
Compensatory	92.37	94.75	95.71
Conj-Complem	81.50	86.35	88.42
Conj-Compens	77.63	83.34	85.80
Specific Agreement for Failers			
	ONE	TWO	THREE
Conjunctive	94.67	94.12	93.94
Complementary	76.18	77.61	78.35
Compensatory	92.34	92.59	92.80
Conj-Complem	91.02	90.83	90.85
Conj-Compens	95.32	94.85	94.71

Seventy Percent Passing Rate

When the passing criterion is lowered to produce a 70% passing rate, the percentage passing with multiple attempts is even higher than that found for the 50% passing rate. In the complementary rule, 96% pass when allowed three attempts, compared with 88% at one attempt.

Table 5.27: Percentage Passing Multiple Attempts, COVAR6, Five Tests, 70%

		OBSERVED SCORE		
	TRUE SCORE	ONE ATTEMPT	TWO ATTEMPTS	THREE ATTEMPTS
Conjunctive	40.79	37.42	46.26	50.71
Complementary	92.45	93.28	95.37	96.22
Compensatory	73.79	73.50	80.11	83.01
Conj-Complem	56.33	55.43	62.28	66.11
Conj-Compens	39.32	36.41	45.15	49.62

Results in Table 5.28 show a similar pattern for the accuracy of classifications to that seen for the 50% passing rate condition. The reliability of classification is particularly high for conditional agreement for Masters, for which the compensatory rule yields an agreement of 99.99%.

Table 5.28: Accuracy For Multiple Attempts, COVAR6, Five Tests, 70%

	Number of Attempts		
Exact Agreement			
	ONE	TWO	THREE
Conjunctive	90.05	90.66	88.72
Complementary	96.65	96.55	96.09
Compensatory	95.55	93.42	90.78
Conj-Complem	89.57	90.66	89.06
Conj-Compens	91.05	90.99	88.61
Kappa			
	ONE	TWO	THREE
Conjunctive	79.14	81.07	77.50
Complementary	74.76	69.92	63.62
Compensatory	88.53	81.55	73.10
Conj-Complem	78.86	80.72	77.19
Conj-Compens	81.00	81.60	77.18
False Negatives			
	ONE	TWO	THREE
Conjunctive	6.66	1.93	0.68
Complementary	1.26	0.27	0.08
Compensatory	2.37	0.13	0.00
Conj-Complem	5.66	1.69	0.58
Conj-Compens	5.93	1.59	0.54
False Positives			
	ONE	TWO	THREE
Conjunctive	3.29	7.40	10.60
Complementary	2.08	3.19	3.84
Compensatory	2.08	6.45	9.22
Conj-Complem	4.76	7.65	10.36
Conj-Compens	3.02	7.42	10.85
Conditional Agreement|Masters			
	ONE	TWO	THREE
Conjunctive	83.67	95.26	98.34
Complementary	98.63	99.71	99.92
Compensatory	96.79	99.82	99.99
Conj-Complem	89.95	96.99	98.97
Conj-Compens	84.92	95.96	98.62
Conditional Agreement|Non-Masters			
	ONE	TWO	THREE
Conjunctive	94.45	87.50	82.09
Complementary	72.39	57.79	49.13
Compensatory	92.05	75.39	64.82
Conj-Complem	89.09	82.49	76.28
Conj-Compens	95.02	87.77	82.12

In contrast, the pattern for consistency in classification reliability is quite different than that for accuracy. With the exception of specific agreement for Failers, all measures showed increased classification reliability.

Table 5.29: Consistency for Multiple Attempts, COVAR6, Five Tests, 70%

	Number of Attempts		
Exact Agreement			
	ONE	TWO	THREE
Conjunctive	87.28	88.69	89.66
Complementary	95.83	97.29	97.88
Compensatory	93.78	95.58	96.37
Conj-Complem	86.36	89.17	90.49
Conj-Compens	88.26	89.57	90.48
Kappa			
	ONE	TWO	THREE
Conjunctive	72.85	77.25	79.31
Complementary	66.72	69.29	70.81
Compensatory	84.03	86.13	87.13
Conj-Complem	72.39	76.95	78.80
Conj-Compens	74.65	78.93	80.97
Pass Rep 1, Fail Rep 2			
	ONE	TWO	THREE
Conjunctive	6.36	5.67	5.20
Complementary	2.07	1.37	1.06
Compensatory	3.10	2.21	1.81
Conj-Complem	6.83	5.43	4.79
Conj-Compens	5.89	5.21	4.79
Specific Agreement for Passers			
	ONE	TWO	THREE
Conjunctive	83.01	87.77	89.79
Complementary	97.77	98.58	98.90
Compensatory	95.77	97.24	97.81
Conj-Complem	87.69	91.30	92.81
Conj-Compens	83.87	88.45	90.40
Specific Agreement for Failers			
	ONE	TWO	THREE
Conjunctive	89.84	89.48	89.51
Complementary	68.95	70.71	71.92
Compensatory	88.26	88.88	89.32
Conj-Complem	84.70	85.65	85.99
Conj-Compens	90.78	90.49	90.56

Summary of Results for Research Question 3

As expected, as more opportunities are provided for retaking tests, more examinees pass the decision rule. The passing rate is even higher for complex rules as more students pass each individual test. However, the passing rate does have a differential effect on the accuracy and consistency of classifications.

- For accuracy, adding more attempts to pass increases measures of conditional agreement for Masters for all decision rules for both passing rates. The opposite is found for Non-Masters, for which conditional agreement and false positives decrease.
- For consistency, a different pattern is seen for the reliability of classifications in the 70% passing rate conditions to that in the 50% passing rate conditions. In the 70% passing condition all measures (except one) show increases with multiple attempts.
- Test difficulty shows similar effects as those found in Research Questions 1 and 2, with the more difficult test showing lower accuracy for Masters, and higher accuracy for Non-Masters when candidates are given multiple opportunities to pass.
- Multiple attempts results in the highest classification reliability found in any of the simulation conditions – 99.99% for conditional agreement for Masters. It also results in the highest rate of false positives and the lowest rate of false negatives. In the compensatory rule with a 50% passing rate, 12.5% of examinees are classified as false positives and 0.01% are classified as false negatives.

Table 5.30: Effect of Multiple Attempts to Pass, COVAR6, Five Tests, 50%

	Exact Agreement		Kappa		**Conditional Agreement|Master***		**Conditional Agreement|Non-Master***		False Negative**	False Positive**	Pass 1/ Fail 2**
	ACC	CON	ACC	CON	**ACC**	**CON**	**ACC**	**CON**	ACC	ACC	CON
Conjunctive	L	~	~	H	**H**	**H**	**L**	**L**	H	L	~
Complementary	L	H	L	H	**H**	**H**	**L**	**H**	H	L	H
Compensatory	L	H	L	H	**H**	**H**	**L**	**H**	H	L	H
Conj-Complem	L	H	~	H	**H**	**H**	**L**	**L**	H	L	H
Conj-Compens	L	~	L	H	**H**	**H**	**L**	**L**	H	L	~
H indicates increased reliability; L indicates decreased reliability; ~ indicates no or mixed effect											

*For CON, this column indicates Specific Agreement for Passers and Failers respectively.
**Higher reliability is associated with lower percentage of false negatives, false positives, and Pass 1/Fail 2.

Table 5.31: Effect of Multiple Attempts to Pass, COVAR6, Five Tests, 70%

	Exact Agreement		Kappa		**Conditional Agreement|Master***		**Conditional Agreement|Non-Master***		False Negative**	False Positive**	Pass 1/ Fail 2**
	ACC	CON	ACC	CON	**ACC**	**CON**	**ACC**	**CON**	ACC	ACC	CON
Conjunctive	L	H	~	H	**H**	**H**	**L**	**L**	H	L	H
Complementary	L	H	L	H	**H**	**H**	**L**	**H**	H	L	H
Compensatory	L	H	L	H	**H**	**H**	**L**	**H**	H	L	H
Conj-Complem	~	H	~	H	**H**	**H**	**L**	**H**	H	L	H
Conj-Compens	L	H	~	H	**H**	**H**	**L**	**L**	H	L	H
H indicates increased reliability; L indicates decreased reliability; ~ indicates no or mixed effect											

*For CON, this column indicates Specific Agreement for Passers and Failers respectively.
**Higher reliability is associated with lower percentage of false negatives, false positives, and Pass 1/Fail 2.

Overview of Simulation Results

Simulation results for Research Questions 1, 2, and 3 reveal both expected and unexpected findings. Results for individual tests, and for passing rates, perform as expected and provide evidence that the simulation method is a valid approach to the estimation of classification reliability. The consistency of findings with other studies of individual tests lends support to the viability of the simulation method for investigating questions about multiple measures and complex decision rules. The simulation method is particularly valuable in its ability to illustrate the different effects that characteristics of tests and decision rules have on accuracy and consistency. Overall, results for classification reliability are not always intuitive, which offers support to the value of structuring simulations to investigate various decision rules, test characteristics, and examinee populations.

One particular contribution of this study is the presentation of a wide variety of measures of agreement. Previous studies have generally used measures of exact agreement, and sometimes Kappa, to characterize classification reliability. Although the use of fewer measures would simplify the interpretation of results, it also obscures an important finding of this study –factors such as number of tests, covariance, and multiple attempts may have differential effects on different measures of agreement. Some may suggest that it is desirable to summarize agreement using a single measure, and the literature cited previously illustrates the debate surrounding use of exact agreement versus Kappa to describe such agreement. Cicchetti and Feinstein (1990) question the desirability of searching for a single measure of agreement. Findings from this study provide support for their viewpoint – the use of a single index would

result in quite different interpretations of the results. Some decision rules provide higher accuracy and consistency for students who should pass, and others are more accurate for students who should fail. Rather than viewing this phenomenon as a problem to be overcome, researchers and decision makers need to appreciate this distinction and consider its implications when structuring decision rules. There may be situations in which it is advantageous to use a rule that provides greater accuracy for particular subgroups.

Simulation results for individual tests are similar to those found in previous analytic and simulation studies (Bradlow & Wainer, 1998; Klein & Orlando, 2000; Rudner, 2001; Subkoviak, 1988). For a moderately difficult test (50% passing rate) with high reliability (i.e, r = .90), the same decision would be made based on two parallel tests for 86% of examinees. This percentage increases to 88% for an easier test in which 70% of examinees pass. If the decision could be based on the examinee's true score, the corresponding percentages would increase to 90% and 91% respectively. These levels of reliability serve as the benchmark for the comparisons of the effects of increasing number of tests, covariance, and opportunities to retest. In an important decision, misclassification of 12% of the examinees may not be acceptable. In cases of lower test reliability and therefore lower classification reliability, the need to improve classification reliability becomes that much more important.

In the midst of conflicting findings among the many conditions in the simulation, one clear pattern emerges. The compensatory rule provides, in almost every condition, the most accurate and consistent classification. This occurs because

test errors counterbalance each other in the compensatory rule due to the additive method of combining scores, in contrast to the conjunctive and disjunctive rules in which errors compound. The effect of this compounding varies according to whether we are interested in classification reliability for Masters and Non-Masters.

In the testing world there are strong recommendations that important decisions should be based on multiple measures. With the exception of the compensatory rule, findings from conditions that increase the number of tests in the decision rule show nuanced results. Measures of exact agreement increase for all types of decision rules. However, when we look at the effect for Masters and Non-Masters, we see that increasing the number of tests increases agreement for Masters when the rule is complementary or compensatory, but not with any rule that incorporates a conjunctive element. The opposite is true for Non-Masters – increasing the number of tests increases agreement for the conjunctive rule, but not the complementary rule. Increasing the number of tests can result in very high levels of classification accuracy for some rules – over 97% for some measures of agreement.

It is reasonable to believe that increasing the covariance between tests should increase classification reliability. Results from the simulation studies are quite mixed except for the compensatory rule, and differ according to the difficulty of the test. Classification accuracy is higher for both Masters and Non-Masters, but consistency shows opposite results for Masters versus Non-Masters. One curious finding is that the passing rate is actually lower with higher covariance between tests in the complementary rule than in the lower covariance condition. This suggests that variability among tests plays an important role in complementary decisions, and

lower covariance relates to higher variability among the scores for an examinee across the five tests.

Allowing students to retest is a commonly used strategy in practical testing situations to increase classification reliability. The simulation results clearly illustrate the benefits and liabilities of allowing examinees to retest. With each retest, the classification reliability for the overall group and for Non-Masters declines, but increases for Masters. The pattern of false negatives and positives supports this finding – false negatives increase and false positives decrease for every rule. The percentages of classification reliability become quite high, with over 97% agreement for Masters for all decision rules given three attempts. This increased accuracy for Masters comes with a price – false positive rates also increase, with over 12% of examinees identified as Masters who in fact are Non-Masters.

Test difficulty shows consistent effects across all three research questions. The easier test in which 70% passed each test shows higher accuracy for Non-Masters, and the harder test (50% passing rate for each test) shows higher accuracy for Masters. The impact of test difficulty for the overall measure of exact agreement is different based on the decision rule utilized: the difficult test has higher accuracy for all rules utilizing a conjunctive element, but the opposite is true for the complementary and compensatory rule.

Simulation results in this study also provide a comparison between measures of accuracy and consistency. Although in most conditions the pattern of results is similar for accuracy and consistency, when multiple attempts are permitted results for exact agreement and kappa are quite different according to whether accuracy or

consistency is assessed. This is an example of the unpredictable effect of highly complex rules, since in the multiple attempt conditions a complementary rule is layered on top of all the rules. If only the consistency measures were available, the conclusion would be reached that increased opportunities to retest improves all types of classification consistency. Measures of accuracy, however, show lower agreement for the overall group and for Non-Masters.

In summary, results from the simulation study confirm the value of the approach for exploring classification accuracy and reliability for complex decision rules under a variety of testing conditions since important determinants of classification reliability may combine to produce unexpected outcomes. Major findings have been highlighted, but there are many more comparisons that can be made based on the tables presented in this chapter.

Given the recommendation to use simulation methods to explore the classification reliability of complex decisions, it is valuable to explore the application of the method to an actual testing scenario. Chapter 6 provides just such an application to data obtained from the GED Testing Service.

Chapter 6: Illustration using GED Test Data

The fourth research question investigates the application of the general method explored through the simulation approach in Chapter 5 to a real dataset. The utility of the approach is evaluated by comparing classification reliability estimated by the simulation method with that obtained using actual data for approximately 100,000 examinees who took Form G of the General Educational Development (GED) Tests in 2004.

Research Question 4: How well does the suggested simulation method for estimating classification reliability of complex decisions estimate the classification reliability that would be obtained if actual data for two test administrations could be compared?

The question of classification consistency is best answered through the administration of two parallel test forms to each examinee. For practical reasons, obtaining parallel form data for an operational test is usually not feasible. An alternative approach is to create two, half-tests from a single administration as a proxy for parallel form data. Such an approach was used in the Livingston and Lewis (1995) study to evaluate their method for calculating the classification reliability for a test based on one administration, and by Subkoviak in a 1988 article exploring practical guidelines for considering the reliability of mastery tests.

Use of split-half scores requires item-level data, which is frequently not readily available to researchers. The simulation method presents a practical alternative to the split-half approach because it can be implemented based on information easily obtained for a set of tests: descriptive statistics, test reliability, and

the covariance matrix. It also represents an advantage over the split-half method because it creates the opportunity to investigate decision accuracy as well as consistency. In this chapter, a comparison will be provided of the classification consistency estimated by creating two, split-half scores to those obtained by simulating scores through the method illustrated in Chapter 5. The empirical baseline will be compared directly with the model-based approach as it is applied to a half-length GED. If substantial agreement is found, we can apply the model-based approach to the full-length GED with confidence. The accuracy and consistency of the GED passing rule will be explored as well.

Description of GED Test Battery

Over 700,000 examinees take the GED Test Battery in the U.S. states and affiliated territories, Canadian provinces, military installations, and prisons as a means of earning a high school equivalency credential. The battery includes five tests: Language Arts, Reading; Language Arts, Writing; Mathematics, Social Studies, and Science. All tests are comprised of multiple-choice items, with the exception of Language Arts, Writing which includes both multiple-choice items and an essay scored using a 4-point rating scale. The multiple-choice and essay scores are combined to yield a single score for the Writing test. GED candidates must earn a minimum scale score of 410 on each test, and an overall average scale score of 450, to be awarded a high school equivalency credential. The GED Tests are normed on graduating high-school seniors, and the passing criterion for each test is set so that 40% of graduating high-school seniors would not pass each GED Test.

The entire battery of tests requires about 7 ½ hours to complete, and not all candidates take all of the tests at one sitting. Candidates are also given multiple opportunities to pass each test, and the number of attempts permitted varies depending on the locale in which the candidate tests. For the current study, only data from one attempt for candidates who took all five tests are included.

Method

A dataset containing item-level responses for 110,991 candidates who took all five GED tests was obtained from the GED Testing Service.[5] To provide a proxy for parallel form data for the examinees, the item-level responses were used to construct half-test scores. These half-test scores for each examinee were then used to provide a measure of classification consistency for each test and for the overall GED decision rule.

A number of different techniques have been suggested for constructing half-test scores. Crocker and Algina (1986) outline the most common methods:

1. Construct each half-test, to the extent possible, to match the table of specifications for the overall test.
2. Rank the items in order of difficulty, and then include items with odd-numbered ranks in one half-test, and even-numbered ranks in the other.
3. Randomly assign half the items to each half-test.
4. Include odd-numbered items in one half-test, and even-numbered scores in the other.

[5]Since a classical test theory approach was used, only the multiple-choice portion for the Writing Test is used in this study.

Of these methods, the first one is most appealing from a validity standpoint because it maximizes the similarity of content between the two half-tests. It is also the most difficult to implement because it requires more information about the test items than may be available to researchers who are not test specialists. The second approach creates two half-tests that are most likely quite similar in difficulty, but may in fact be measuring different constructs. The same is true for the third method. The last method, the odd/even approach, is simple to carry out based on item-level responses, and controls for some important factors known to affect test performance, such as the effects of familiarity and fatigue.

Feldt and Brennan (1993) discuss the pros and cons of different strategies of allocating items to half-tests given the current practice of constructing items that relate to a common passage (as is the case with most of the tests in the GED battery). Feldt and Brennan reason that splitting items from the same passage among different half-scores creates positive bias in the correlation coefficient, whereas assigning all items related to a specific passage to the same half score creates negative bias. Although Feldt and Brennan do not offer any empirical evidence to support this reasoning, based on their hypothesis the source of bias in using the odd/even method with the GED Tests would be to positively bias the correlation between the two half-tests.

The following steps were used to create and score split-half scores for the GED Tests:

1. Create two scores for each examinee on each test by summing the number of correct responses separately for odd and even numbered items.

2. A cut-score for each half-test was determined using the following steps:
    a. Found the percentage who passed each full test by earning a scale score of 410 or higher (the GED criterion).
    b. Selected the score for each half-test that yielded a similar pass rate to that on the full test.
    c. Applied this score for each half-test to determine whether the examinee passed each half-test.
3. The overall average criterion for the half-scores (equivalent to an average scale score of 450) was calculated using the following steps:
    a. Converted the half-test scores to z-scores (using mean and standard deviation for the same half-test). This was done prior to averaging scores for the five tests since the Reading Test has fewer items than the other tests.
    b. Found the average z-score for each examinee for odd and even half-tests.
    c. Fouind the z-score for each half-test that corresponded to an average scale score of 450.
    d. Apply this average z-score criterion to each examinee's average z-score for odd and even half-tests.
4. Apply GED rule (must pass each test with a scale score equivalent of 410 and overall average equivalent to 450) to each examinee's scores on odd and even half-tests.

Overview of GED Test Data

Descriptive statistics for the raw and split-half scores for all five GED Tests appear in Table 6.1. The Writing, Social Studies, Science, and Math tests each include 50 multiple-choice items; the Reading Test contains 40 multiple-choice items.

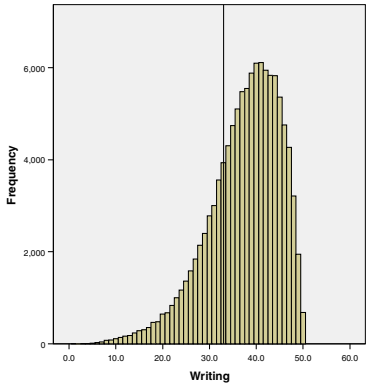
Table 6.1: Descriptive Statistics for Half- and Full-Length GED Tests

	Mean	StdDev	Skewness	Kurtosis
Writing	**37.40**	**7.66**	**-.859**	**.700**
Odd	19.49	3.85	-1.129	1.414
Even	17.91	4.28	-.577	-.003
Social Studies	**37.26**	**8.55**	**-.727**	**.044**
Odd	18.14	4.55	-.587	-.222
Even	19.11	4.40	-.869	.379
Science	**38.26**	**7.45**	**-.952**	**.811**
Odd	19.84	3.85	-1.145	1.375
Even	18.42	4.04	-.705	.191
Reading	**31.95**	**5.88**	**-1.220**	**1.649**
Odd	16.03	3.16	-1.191	1.461
Even	15.92	3.11	-1.097	1.347
Math	**29.99**	**9.58**	**-.209**	**-.626**
Odd	14.63	5.18	-.168	-.647
Even	15.36	4.87	-.261	-.582

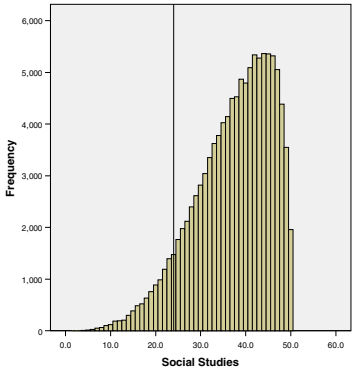
The score distributions for the GED tests are shown in Figures 6.1. The cut-score set by the GED Testing Service is also indicated in each figure. Several characteristics of test performance are evident in the histograms. First, none of these tests appears to be normally distributed. With the exception of Math, the tests are negatively skewed and kurtotic. The degree of skewness and kurtosis is documented in Table 6.1. Such distributions are not uncommon on criterion-referenced tests and licensing tests such as the GED Tests. Second, the passing rate is high on all tests, with the lowest rates seen on the Writing and Math tests.

Figure 6.1: Histograms Showing Raw Score Distributions for GED Tests

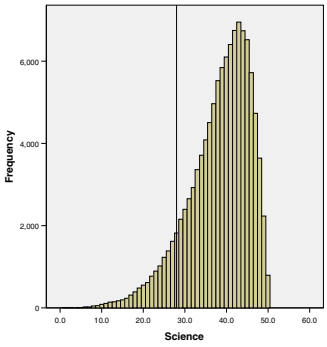
Writing



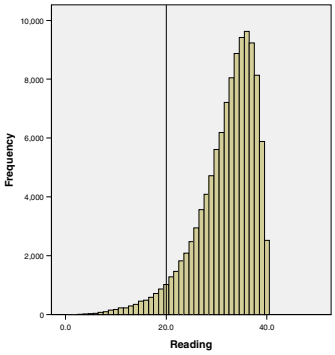
Social Studies



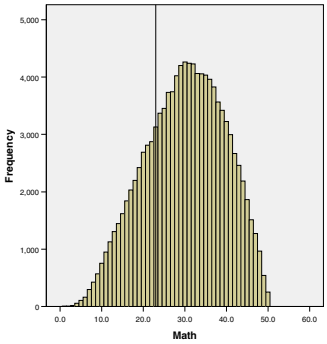
Science



Reading



Math



The GED Tests are moderately correlated with each other (Table 6.2), with the highest relationship between Social Studies and Science (r=.80) and the lowest relationship for Reading and Math (r=.51). Table 6.2 also shows the correlations among the odd half-tests and the even half-tests, which are fairly similar to each other and, as expected due to their shorter length, lower than the correlations among the full-length tests.

Table 6.2: Correlations Among Half- and Full-Length Tests

	Writing	Social Studies	Science	Reading	Math	
Writing	1	.695 .636 .612	.701 .650 .606	.694 .637 .589	.651 .554 .612	Full Test Odd Even
Social Studies		1	.803 .729 .717	.787 .704 .708	.634 .575 .581	Full Test Odd Even
Science			1	.740 .674 .640	.685 .597 .632	Full Test Odd Even
Reading				1	.559 .488 .513	Full Test Odd Even

The reliability of the GED tests was calculated using Coefficient Alpha for the full-length tests and each of the half-length tests. The correlation between the split-half scores also serves as an estimate of reliability. It is presented in Table 6.3 along with the adjusted estimate of the reliability of the full-length test through use of the Spearman Brown prophecy formula. Inspection of Table 6.3 shows that internal consistency (as measured by Coefficient Alpha) for the two half-tests for each GED Test is similar, as is the adjusted split-half correlation and the Coefficient Alpha for the full-length test. All half-test correlations are similar for each test, and are similar to the full-test correlations.

Table 6.3: Reliability Estimates for GED Test Data

	Coefficient Alpha			Correlation	
	Full Test	Odd Half-Test	Even Half-Test	Half-Tests	Adjusted*
Writing	.877	.781	.784	.773	.872
Social Studies	.900	.815	.821	.828	.906
Science	.872	.783	.762	.782	.878
Reading	.853	.751	.732	.757	.862
Math	.898	.819	.810	.819	.901

*Spearman Brown prophecy formula was used to estimate the reliability of the full-length tests based on the correlation between the half-tests.

Passage of the full-length tests was determined using the raw score passing criterion for each test obtained from the GED Testing Service. The criterion for each half-test was determined by selecting the score yielding the closest passing rate to that obtained on the full-length test. Passing scores for full- and half-length tests, as well as the resulting passing rates, are presented in Table 6.4.

Table 6.4: Passing Criteria and Rates for Half- and Full-Length Tests.

Test	Passing Score: Full Test	Passing Score: Odd	Passing Score: Even	Passing Rate: Odd	Passing Rate: Even	Passing Rate: Full Test
1. Writing*	33	18	16	74.99%	72.73%	76.62%
2. Social Studies	24	12	13	90.74%	91.05%	92.34%
3. Science	28	15	13	89.98%	90.91%	90.59%
4. Reading	20	10	10	95.39%	95.75%	95.72%
5. Math	23	11	12	77.12%	77.16%	76.82%
Passed GED rule				61.59%	60.54%	63.20%

*The passing criterion for the Writing test combines the multiple-choice and essay score. For purposes of this example, the passing score was set to yield a similar pass rate to that obtained on the combined score.

For most of the tests, the passing rates are quite similar for the full- and half-length tests. However, for the Writing and Social Studies Tests the closest approximation to the passing rate is somewhat lower for the half-length tests than the full-length tests. This contributes to a slightly lower overall passing rate for the half-

tests (61.6% and 60.5% for odd and even scores respectively) as compared that for the full-length test (63.0%). Although the GED decision rule incorporates both a conjunctive and compensatory component, very few (1.9%) of examinees fail the overall decision rule because they do not earn an overall scale score of 450. This is a reflection of the very high passing rates for several of the tests.

The similarity between passing rates for the half-length tests is encouraging for their use for estimating classification consistency, which is presented later in the chapter in comparison to that obtained based on simulated data.

Simulated Data

Using the method illustrated in Chapter 5, a dataset was generated with similar true score variance, covariances, and test reliability to those of the GED tests. Next, two sets of replicate scores were generated for each examinee on the five tests, and the classification consistency between the two sets was assessed.

The simulation procedure illustrated in Chapter 5 assumes that each test is normally distributed (and therefore the set of tests has a multivariate normal distribution), an assumption that is questionable for the GED Tests given the histograms in Figure 6.1 and descriptive statistics in Table 6.1. This assumption of normality is important because it provides the framework for estimating the true score distribution from that of the observed scores. It is therefore desirable to normalize the observed score distributions before constructing the covariance matrix for the simulated data. If normalization cannot be achieved, an alternate approach is to simulate data for a non-normal multivariate distribution. However, such an approach introduces new challenges into the simulation approach by requiring the estimation of

the true score distribution from a non-normal observed score distribution. For the interested reader, methods for simulating multivariate non-normal distributions are presented in Vale and Maurelli (1983), Headrick and Sawilowsky (1999), and Nevitt and Hancock (1999). Each of these studies estimates multivariate non-normal distributions by allowing for the estimation of skew and kurtosis for each individual variable. An alternate approach is suggested by Mislevy (1984) through use of mixture modeling for estimating non-normal latent distributions.

Peng and Subkoviak (1980) studied the impact of non-normality on classification agreement for individual tests, and found that estimates remained fairly stable as long as the distribution was unimodal. Their findings for individual tests offer support for using the multivariate normal distribution to approximate classification reliability for non-normal distributions. However, since their study did not explicitly address the case of multivariate distributions, the GED Test distributions were normalized.

Several methods were applied to the raw scores for the full-length test in an attempt to transform the scores to a normal distribution. Both log and exponential transformations were unsuccessful in producing normally distributed scores. However, the normalized ranking transformation in SPSS, using Blom's formula (6.1), was moderately successful in transforming the scores for each test to a standard, normal distribution, as shown in Fig. 6.2.

$$ns = \Phi(p)\frac{(r-3/8)}{(N+1/4)} \qquad (6.1)$$

where: ns = normalized score

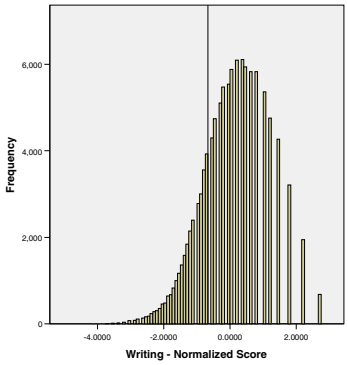
Φ = pth quantile from the standard normal distribution

r = rank

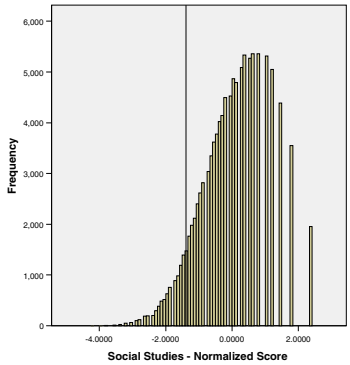
N = number of observations

Figure 6.2: Histograms Showing Normalized Scores for GED Tests

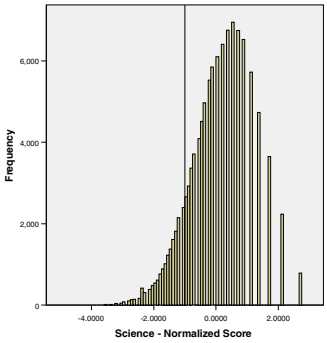
Writing



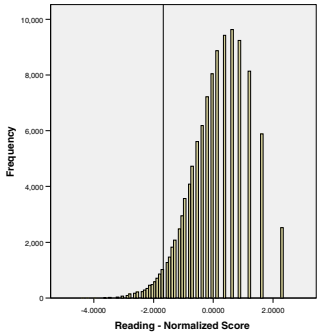
Social Studies



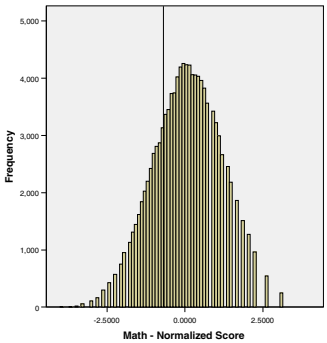
Science



Reading



Math



Given the goal of replicating the relationships seen in the raw scores for the GED Tests, comparison of the correlation matrices for the raw scores and normalized scores is informative. As shown in Table 6.5, the correlational structure observed in the raw scores for the GED tests is well duplicated by the normalized scores.

Table 6.5: Correlations Among Tests for Raw Score Versus Normalized Scores

	Writing	Social Studies	Science	Reading	Math	
Writing	1	.695 .688	.701 .689.	.694 .681	.651 .660	Raw Score Normalized
Social Studies		1	.803 .795	.787 .775.	.634 .642	Raw Score Normalized
Science			1	.740 .719	.685 .699	Raw Score Normalized
Reading				1	.559 .568	Raw Score Normalized

The following steps were used in constructing the simulated dataset in which true and replicate scores were generated to study classification accuracy and consistency of the GED Tests.

1. The covariance matrix for the normalized scores among the five tests was calculated using the half-test scores from the odd-numbered items. (Table 6.6).

Table 6.6: Covariance Matrix for Normalized Scores for Odd Half-Length Test

	Writing	Social Studies	Science	Reading	Math
Writing	.963872 (.765159)	.603844	.596741	.576515	.545608
Social Studies	.603844	.965223 (815213)	.692102	.655146	.564010
Science	.596741	.692102	.956095 (.773073)	.603774	.591530
Reading	.576515	.655146	.603774	.935907 (.738919)	.470365
Math	.545608	.564010	.591530	.470365	.984991 (.815174)

2. The true score variance for each test was calculated by multiplying the observed variance for each test by the test reliability. Table 6.6 shows the true score variance in parentheses for each test. The observed score covariances remain the same.
3. A simulated dataset containing 100,000 true scores, and two replicate scores for each true score were generated (using *R,* as illustrated in Chapter 5).
4. Passing status for true and replicate scores was obtained by applying the normalized score equivalent to the passing score for the odd half-test.
5. Contingency tables were constructed showing classification accuracy (by comparing the outcome for true score to that of one replicate score) and consistency (comparing the outcomes based on two replicate scores) for each test, as well as for the overall GED passing rule.
6. Measures of agreement were calculated to characterize classification reliability as in Chapter 5.

The covariances among tests and reliability[6] of test scores in the simulated data are shown in Table 6.7. In general, the simulated data provided a good approximation of test reliability and covariances among tests for the odd, half-length test.

[6] Estimated by correlating the two replicate scores for each test.

Table 6.7: Covariance and Reliability for Half-Length Test Simulated Data

Test	Reliability	Writing	Social Studies	Science	Reading	Math
Writing	.774	.965565	.608858	.601254	.579545	.549217
Social Studies	.830	.608858	.969608	.696622	.657921	.568894
Science	.782	.601254	.696622	.961652	.606085	.597079
Reading	.756	.579545	.657921	.606085	.936459	.473104
Math	.821	.549217	.568894	.597079	.473104	.992670

A comparison of the passing rates for both sets of data is shown in Table 6.8. The highly similar passing rates between the simulated and odd-half scores support the success of the simulation in approximating the score distribution.

Table 6.8: Passing Rates for Raw Score and Simulated Data, Half-Length Tests

Test	Raw Score: Odd Half-Test	Simulated, Half-Length Test
Writing	74.99%	75.14%
Social Studies	90.74%	92.00%
Science	89.98%	90.20%
Reading	95.39%	95.67%
Math	77.12%	75.46%
Passed GED rule	61.59%	59.40%

The consistency of classification based on the comparison of split-half scores and simulated half-length scores is shown in the first two columns in Table 6.9. Counts from which the measures of agreement were calculated appear in Appendix I.

Table 6.9: Consistency Among Split-Half, Half-Length Simulated Scores, and Full Length Simulated Scores

	Split-Half	Simulated Half Length	**Simulated Full Length**
Exact Agreement			
Writing	83.08	82.99	**86.95**
Social Studies	92.88	93.35	**94.86**
Science	92.45	91.06	**93.05**
Reading	96.29	95.08	**95.98**
Math	86.38	84.91	**88.97**
Overall Rule	83.62	82.82	**87.80**
Kappa			
Writing	56.22	54.44	**65.42**
Social Studies	56.99	54.64	**65.88**
Science	56.32	49.56	**61.97**
Reading	56.15	40.95	**53.81**
Math	61.38	59.28	**70.20**
Overall Rule	65.56	64.38	**74.32**
Pass 1, Fail 2			
Writing	9.59	8.49	**6.52**
Social Studies	3.41	3.28	**2.58**
Science	3.31	4.49	**3.47**
Reading	1.68	2.48	**2.01**
Math	6.79	7.56	**5.47**
Overall Rule	8.71	8.58	**5.98**
Specific Agreement, Passers			
Writing	88.55	88.68	**91.27**
Social Studies	96.08	96.39	**97.20**
Science	95.82	95.04	**96.13**
Reading	98.06	97.43	**97.90**
Math	91.17	90.00	**92.69**
Overall Rule	86.59	85.54	**90.02**
Specific Agreement, Failers			
Writing	67.64	65.76	**74.15**
Social Studies	60.90	58.25	**68.68**
Science	60.48	54.51	**65.83**
Reading	58.08	43.53	**55.91**
Math	70.21	69.28	**77.51**
Overall Rule	78.97	78.85	**84.30**

In general, the simulation method provided somewhat lower estimates of classification reliability than the split-half method for the individual tests, but quite

similar estimates for the overall decision rule. Results are consistent with Feldt and Brennan's supposition that using an odd/even approach with passage-based tests may result in overestimation of classification consistency.

Both the simulation and split-half methods provide estimates of classification consistency for tests half the length of the GED tests. It is expected that the half-test provides an underestimate of test reliability given the strong relationship between test length and reliability. A more accurate estimate of reliability for the GED Tests may be found by simulating data for the full length tests. The same method (Steps 1-6) was utilized, but the covariances for the full-length tests (Table 6.10) and the stepped-up reliability for each test were used to generate the simulated data.

Table 6.10: Covariance Matrix for Normalized Scores for Full-Length Test

	Writing	Social Studies	Science	Reading	Math
Writing	.989856 (.863154)	.678962	.681302	.669293	.654642
Social Studies	.678962	.984558 (.892009)	.784387	.759690	.635370
Science	.681302	.784387	.988584 (.867977)	.706269	.693413
Reading	.669293	.759690	.706269	.976115 (.8414111)	.559507 (.896791)
Math	.654642	.635370	.693413	.559507	.995329

Table 6.11 shows the covariances among tests and reliability[7] of test scores in the simulated data. Comparisons with Table 6.10 indicate that the simulated data provided a good approximation to the normalized raw scores for the full-length test.

[7] Estimated by correlating the two replicate scores for each test.

Table 6.11: Covariance and Test Reliability for Full-Length Test Simulated Data

Test	Reliability	Writing	Social Studies	Science	Reading	Math
Writing	.871	.987146	.679478	.683724	.666164	.654553
Social Studies	.907	.679478	.987192	.788571	.758539	.640294
Science	.879	.683724	.788571	.989186	.705644	.696441
Reading	.862	.666164	.758539	.705644	.973531	.559052
Math	.901	.654553	.640294	.696441	.559052	.997333

The passing rates for the raw and simulated data for full length tests are shown in Table 6.12. As was the case for the half-length tests, the simulated full-length test had very similar passing rates to that observed in the raw data.

Table 6.12: Passing Rates for Raw Score and Simulated Data for Full-Length Tests

Test	Raw Score	Simulated*
Writing	76.62%	74.76%
Social Studies	92.34%	91.81%
Science	90.59%	89.83%
Reading	95.72%	95.45%
Math	76.82%	75.43%
Passed GED rule	62.81%	61.03%

*For Replicate 1

Results of the full-length test simulation are presented in Table 6.9 in the third column. The effect of a longer test was to increase classification reliability, a similar result to the well documented effect of increasing test length on test reliability. The simulation results for the full-length test provide a better estimate of classification consistency for the GED Tests. Exact agreement for the overall decision rule was 87.8%, and 6% of examinees would receive a different decision based on which replicate score was used.

Although all the GED Tests exhibit fairly high levels of reliability, unlike in the simulation study in Chapter 5 there is some variability among the tests. In

addition, there is variability in passing rates among the tests. This provides the opportunity to examine classification consistency in relation to test reliability and passing rate. The influence of test difficulty is illustrated by results for the Reading test. Reading has the *highest* classification consistency for almost all measures, with the exception of Kappa and Specific Agreement for Failers. Reading has the *lowest* test reliability (r=.757). This inverse relationship is most likely due to the effect of test difficulty – Reading is the easiest test, and therefore the scores for most of the examinees are not near to the cut-score. A comparison of the test and classification reliability for the other tests is consistent with this interpretation. There is no monotonic relationship between test reliability and classification reliability. However, classification reliability increases consistently with higher passing rates.

One advantage of the simulation method over the split-half approach is the opportunity to estimate classification accuracy. Table 6.13 shows measures of classification accuracy for the simulated full-length tests.

Table 6.13: Classification Accuracy for Simulated, Full-Length Tests

	Exact Agreement	Kappa	False Negatives	False Positives	CA|Masters	CA|Non-Masters
Writing	90.69	74.83	5.42	3.89	92.90	83.59
Social Studies	96.46	75.08	2.27	1.27	97.55	82.40
Science	95.10	71.30	3.20	1.70	96.49	80.42
Reading	97.24	64.18	1.92	0.84	98.01	75.88
Math	92.19	78.59	4.49	3.32	94.14	85.80
Overall Rule	90.85	80.42	6.44	2.71	90.06	92.30

The classification accuracy of the overall GED rule is 90.9%, with the individual test accuracy varying from 90.7% for Writing to 97.2% Reading. The overall rule provides slightly better classification accuracy for Non-Masters than

Masters, as evidenced by conditional agreement for Non-Masters and the percentage of false negatives.

Summary of GED Test Illustration

This illustration investigated the utility of the proposed simulation method for providing information about the classification reliability of actual test data. A comparison of results for the simulation method and the traditional, split-half approach also sheds light on the validity of the proposed method. The split-half approach, however, is not without pitfalls and differences in results could reflect shortcomings in either method. The split-half approach can only estimate the classification reliability of a test half the length of the original test Although the Spearman Brown prophesy formula provides an adjustment for test length to the reliability coefficient, there is no such adjustment for classification reliability. In addition, the lower number of items on the half-tests contributes to difficulty in replicating the passing rate for the full-length test, and the similarity between the two scores may be influenced by how the items were divided between the two tests.

The simulation method brings its own challenges, primarily the necessity of modeling test scores using a normal distribution. The GED Tests are not normally distributed, and this is not an uncommon finding for high stakes tests. Although it may be possible to adjust the covariance matrix through normalization of scores, this does not address the underlying problem of non-normality in the data. Nevertheless, the effect of non-normality did not have much of an effect on findings in the simulation study, perhaps because the passing rates are relatively high on all the GED

Tests. This is an area that will need more study if the simulation method is to prove useful in applied settings.

Results for the GED Test data show, however, that simulated data provided reasonably close approximation to the raw score data in terms of passing rates, and very similar results for measures of classification consistency with the split-half approach. These results estimate that 88% of examinees would receive the same overall decision if they took two parallel forms of the GED Test Battery; that 91% are accurately classified in regard to mastery; and that the overall decision is somewhat better at detecting Non-Masters than Masters. Similar results were demonstrated for conjunctive decisions in Chapter 5. The simulation differed from actual GED test data in two important ways. First, only the multiple-choice portion of the writing test was modeled. It is likely that inclusion of the essay portion would result in lower reliability, and therefore lower accuracy, for the writing test. Second, only one attempt was modeled for each test. Since GED test takers are given multiple attempts to pass, the estimated accuracy is most likely higher than that of the actual GED Tests.

The impact of choice of decision rule on percentage of candidates passing the GED is illustrated in Table 6.14. Some decision rules are logically nested in all testing applications given that similar passing criteria are applied. All students who pass each of the five tests (conjunctive rule) will also pass at least one test (complementary rule). Similarly, all students who pass the conjunctive-compensatory rule will pass the simple conjunctive rule. Students who pass the conjunctive rule will also pass the compensatory rule if the same average passing criteria is used. The other

rules – compensatory and complex combinations of the simple rules – are not necessarily nested. The GED passing criterion for each test was applied to each individual test. For the simple compensatory rule, an average score of 410 was required across all five tests. The conjunctive-compensatory rule was the same as the actual GED rule, and required an average score of 450.

Table 6.14: Percentage of Candidates Who Pass Various Decision Rules

98% Pass Complementary Rule				
	88% Pass Compensatory Rule			
		65% Pass Conjunctive Rule		
			63% Pass Conjunctive-Compensatory Rule	
All Candidates				

Almost all candidates (98%) pass at least one of the GED tests, and the next highest percentage (88%) earns an average scale score of 410 and therefore passes the compensatory rule. The percentage passing the GED (conjunctive-compensatory) rule is only slightly less than the percentage passing the simple conjunctive rule. In this case, use of a compensatory rule would result in the passage of many candidates who did not meet minimum mastery of all five individual tests. This is most likely due to the high passing rates on the tests, which provides great opportunity for a high score on one test to compensate for a low score on another. Table 6.14 also illustrates that, given the choice, candidates would most likely choose to be evaluated using the complementary rule.

The only rule that does not exhibit a nested relationship with the other rules for the GED is the conjunctive-complementary rule. For purposes of this example, the conjunctive-complementary rule requires the student to pass reading, writing, and math tests and either the social studies or science test. Thirty-two percent of the candidates pass the complementary but fail the conjunctive-complementary rule; a

very small percentage (1%) pass the conjunctive-complementary but fail the conjunctive rule; and 23% fail the conjunctive-complementary but pass the compensatory rule.

Extension to More Complex Configural Rules

In the previous illustration, promising results were presented regarding the utility of the simulation method for estimating the classification reliability of complex decision rules for actual test data. One advantage of the simulation method is the ease with which highly complex configural rules can be accommodated. Once simulated data has been created, it is straight forward to apply even the most complex configural rule. The following hypothetical example uses the simulated data for the GED Tests to illustrate application of the simulation method to a configural rule that is more complex than those modeled in Chapter 5.

This example is structured to show how the classification reliability of a simple conjunctive rule for five tests is impacted by increasing the complexity of passing conditions. It is similar in spirit to the Louisiana rule in which students may compensate for a lower score on one test by earning a higher score on another test. Interest in changing the decision rule might stem from a concern about the potential error in applying a criterion to test scores. Therefore a proposal is made to alter the decision rule to allow multiple criteria for passing each test.

For purposes of this example, the simple rule requires a student to earn a standard score of zero or higher on each of five tests (Writing, Social Studies, Reading, Science, and Math).

For the complex rule, each test is viewed as measuring primarily verbal or quantitative skills. A student can demonstrate proficiency on each type of skill in a number of ways. This complex rule is defined below.

Verbal: There are three configural rules for passing.

1. Writing >= 0 AND Reading >= -.5 AND Social Studies >= -.5
   OR
2. Reading >= 0 AND Writing >= -.5 AND Social Studies >=-.5
   OR
3. Social Studies >= 0 AND Reading >= -.5 AND Writing >=-.5

Quantitative: There are two configural rules for passing.

1. Math >= 0 AND Science >= -.5
   OR
2. Science >= 0 AND Math >= -.5

Overall Decision: Pass one Verbal and one Quantitative rule.

Table 6.15 shows the resulting contingency tables for accuracy and consistency of decision outcomes for both the simple and complex rules. As expected, the additional opportunities to retest increases the number of Masters who pass: 28,167 pass the simple rule versus 45,655 for the complex rule. A similar pattern, but at a lower level, is seen for consistency which compares the percentage who pass based on two replicate scores.

Table 6.15: Contingency Table for Simple and Complex Rule

Simple Rule				**Complex Rule**			
Accuracy				Accuracy			
	Replicate 1				Replicate 1		
TRUE	Fail	Pass	Total	TRUE	Fail	Pass	Total
Fail	69615	2218	71833	Fail	51714	2631	54345
Pass	6444	21723	28167	Pass	7129	38526	45655
Total	76059	23941	100000	Total	588443	41157	100000
Consistency				Consistency			
	Replicate 2				Replicate 2		
Replicate 1	Fail	Pass	Total	Replicate 1	Fail	Pass	Total
Fail	70970	5089	76059	Fail	52735	6108	58843
Pass	5103	18838	23941	Pass	6030	35127	41157
Total	76073	23927	100000	Total	58765	41235	100000

Classification accuracy is shown in Table 6.16. Although more students pass the complex rule than the simple rule, most measures of classification reliability remain fairly stable. Exact agreement is 91.3% in the simple rule, and 90.2% in the complex rule. The largest difference is found for the accuracy of classifying Masters, 77.1% in the simple rule versus 84.4% in the complex rule. Exceptions are found for conditional agreement for Masters, which is higher in the complex rule than in the simple rule, and the percentage of false negatives which is also higher in the complex rule.

Table 6.16: Classification Accuracy for Simple and Complex Rule

	Exact Agreement	Kappa	False Negatives	False Positives	CA|Masters	CA|Non-Masters
Simple	91.34	77.57	6.44	2.22	77.12	96.91
Complex	90.24	80.18	7.13	2.63	84.39	95.16

Results for classification consistency (Table 6.17) are similar to those found for accuracy with the largest difference found for the consistency with which Masters are classified.

Table 6.17: Classification Consistency for Simple and Complex Rule

	Exact Agreement	Kappa	Pass 1/Fail 2	SA|Passers	SA|Failers
Simple	89.81	72.01	5.10	78.71	93.30
Complex	87.86	74.95	6.03	85.27	89.68

Therefore, if the goal is to increase the overall accuracy of the decision rule, the suggested change to the rule would not be successful. However, if the goal is to increase the accurate identification of Masters, such a change would be beneficial but at the expense of accuracy for Non-Masters.

Summary of Hypothetical Example

This example illustrates the value of the simulation method for investigating a seemingly endless variety of configural rules. Using the simulation method, various decision rules can be evaluated for the level of classification reliability provided and their ability to meet the particular needs of a given decision. As shown in Chapter 5, different decision rules maximize accuracy and consistency for Masters versus Non-Masters, and also create different percentages of false negatives versus false positives.

Other Strategies for Improving Accuracy of GED Decision Rule

The previous illustration is just one of many that could be conducted to investigate ways to increase the accuracy of decisions based on the GED tests. Previous studies, and the simulations in this study, suggest the following factors increase overall accuracy. The effect of these factors varies based on test difficulty and type of decision rule, which supports the value of simulations in establishing the impact on accuracy for a specific testing situation and population. All suggestions must be evaluated in regard to validity; i.e., whether the change makes sense in terms of the purpose of the decision.

1. Increasing the reliability of individual tests. The writing test has the lowest reliability, and lowest accuracy. Simulations could be constructed to identify the necessary increase in reliability to obtain the desired level of overall accuracy for the GED decision rule.

2. Increasing the number of tests. Simulations show increased overall accuracy for all decision rules with the addition of more tests. Validity questions dictate whether it makes sense to add another test to the battery.

3. Increasing the covariance among tests. This factor is more difficult to conceptualize and manipulate in a practical testing situation. One could argue that to the extent that the tests become more similar the validity of the decision is reduced. There may be ways, however, to increase the covariance among the tests without sacrificing the integrity of the individual constructs that are measured.

4. Use of compensatory decision rule. This factor may also be difficult to put into practice due to validity concerns. The GED credential represents mastery of each

separate content area, and allowing high performance on one task to compensate for another may be inappropriate. Simulations could identify how many examinees would pass the overall rule without passing each individual test.

5. Acceptable number of attempts. Although this illustration did not model multiple attempts to pass, the actual GED tests allow for several attempts to pass. Simulations could be structured to estimate the loss of accuracy with each additional attempt.

If it were deemed desirable to increase accuracy for Non-Masters, increasing the difficulty of the tests could be effective in raising accuracy.

Chapter 7: Discussion

The purpose of this study is to explore how decision making strategies contribute to the reliability of decision outcomes given that the scoring of tests necessitates imposing arbitrary cut-scores on errorful test scores. More specifically, the subject of this study is the extent to which methods of combining information in complex decisions contribute to errors in classification. The central question posed is whether we would make the same decision based on a replicate source of information about the examinee. Previous studies explored this topic for a single test, but cannot shed light on this question for decisions based on multiple measures.

The answer to questions of classification reliability requires comparison of two scores for each examinee. Parallel test data provide such data, but are difficult and costly to obtain. A practical alternative is to artificially create two scores for each examinee by dividing each test into two, half-tests that are equivalent in content, difficulty, and variability. However, the split-half approach requires item-level data, and constructing equivalent half-scores can be challenging. In addition, the resulting classification reliability is estimated for a test half the length of the original test, and therefore provides an underestimate of reliability. The Spearman-Brown formula provides a way to adjust the reliability coefficient of a single test for length, but not the reliability of multiple tests combined by an arbitrary decision rule. The general method outlined in this study estimates classification reliability based on a single test administration for a variety of testing purposes and situations. It is illustrated for the particular case of continuous test scores and pass/fail decisions but the method

extends to other testing situations as well, such as multiple proficiency categories applied to continuous scores.

The proposed general method includes the following basic steps:

1. Using the appropriate measurement model, generate a dataset containing true and replicate scores with a similar distribution to the target data.
2. Determine passage of both true and replicate scores for each individual test.
3. Determine passage of the overall decision rule for all sets of scores.
4. Construct a contingency table comparing the decision outcome for true and replicate scores for decision accuracy, and two sets of replicate score for decision consistency.
5. Calculate the appropriate measures of agreement based on the contingency table.

These same steps can be utilized for any complex decision rule for a variety of measurement models, including the classical test theory situations explored in this study, and latent variable approaches such as item response theory and latent class models.

The general method is illustrated using conditions that are plausible given actual testing situations in many high stakes decisions, incorporating moderately correlated tests ($r = .6$) with high reliability ($r = .9$). The difficulty of the test is examined at two levels: 50%, the point at which classification reliability for each individual test is lowest; and also at a somewhat higher rate (70%). However, in many

practical applications passing rates are higher than 70% as demonstrated by the GED Tests example.

What is the utility of this simulation method for answering questions of classification reliability? This question is best answered by examining the adequacy of the simulations in matching desired distributions and providing results that are congruent with previous studies. As shown in Chapter 5, the simulated datasets provide a good approximation to the desired distributions; estimates of classification for individual tests are similar to those in previous studies; and the number of examinees who pass complex decision rules makes sense. When applied to real test data, results are also encouraging.

Application to real data, however, forces the modeling of data that may not meet all the model assumptions. In the case of the GED Test data, credible results are obtained despite some non-normality in the distributions. The utility of the method for non-normal distributions can be interpreted in light of the following quote from Box and Draper (p. 424, 1987), "Essentially, all models are wrong, but some models are useful." The question of how much departure from normality influences estimates is left for future studies, but the model worked reasonably well for the GED Tests.

Simulation results in this study illustrate the importance of how measures are combined, which measure is used to characterize agreement, and whether accuracy or consistency is considered. The lack of straight-forward results emphasizes the value of constructing simulations as a means of exploring the consequences of decision rules. Findings support Chester's assertion (pg. 39, 2003) that the manner in which

scores are combined may be the most important factor when evaluating the validity and reliability of decisions.

The validity of complex decision rules is a subject for another study, but the motivation to increase validity lays the groundwork for practical scenarios related to reliability. For example, if interest lies in identifying examinees who have acquired necessary skills in a variety of subjects (such as the case for high school exit exams), it is logical to propose a conjunctive rule in which the examinee must pass each individual test. If the purpose of the decision is to increase the validity of measurement of a particular skill, such as writing, it is logical to allow examinees to demonstrate skills using tests with different types of items. Allowing examinees multiple opportunities to pass a test may be undertaken in the quest to reduce the effect of error from a particular test.

What do the simulation results tell us about the reliability of decisions based on these strategies? Despite the seemingly mixed findings among the many conditions in the study a few general principles are apparent.

Choice of Agreement Measure is Important. Different measures of agreement may provide different answers. Overall measures of agreement can be affected by the overall passing rate, and changes in decision rules can result in large differences in passing rates. In the most extreme case, decisions that result in all (or no) examinees passing will tend to show higher agreement than those with a more moderate passing rate. Conditional measures of agreement describe reliability separately for Masters and Non-Masters and may therefore be more useful. Conditional measures may provide a different answer than overall measures so decision makers need to consider

what type of accuracy is most important for the given purpose. Some decision rules have different effects for Masters and Non-Masters. Based on the consequences for incorrect decisions, certain types of error may be preferable over others. For example, if a decision is designed to certify newly graduated surgeons, most of us would prefer that it minimize the percentage of examinees that are not qualified but in fact pass the decision rule. On the other hand, a decision rule that determines high school graduation may be more concerned with minimizing the number of examinees who should pass but in fact fail.

Compensatory Rules Increase Classification Reliability. First and foremost, adding up scores provides the most reliable decision for almost every testing scenario examined. From a validity standpoint, a compensatory decision seems less desirable for the high school exit exam situation because it would not guarantee that the examinee had mastered all the desired skills. The more common decision rule in this situation is to use a simple conjunctive rule. Simulation results show that the conjunctive rule approach does a better job of correctly identifying examinees who have not acquired the necessary skills at the expense of misclassifying some examinees who in fact have acquired such mastery. Results consistently show that, compared to the simple conjunctive rule, adding a compensatory component to the conjunctive rule increases classification accuracy and consistency.

The second scenario, in which examinees are given multiple ways of demonstrating skills, uses a complementary rule to combine measures. The complementary rule frequently shows opposite results to those obtained in the conjunctive rule because the two rules apply inverse logic. In the conjunctive rule,

only one of the four cells in the contingency table *passes* the rule; in the complementary rule, the inverse cell in the table *fails* the rule. The complementary rule does a better job of identifying examinees who actually meet the desired criteria, but at the same time incorrectly classifies examinees who should fail the criteria. Perhaps due to the inverse relationship between conjunctive and complementary rules, the combination of conjunctive and complementary rules does not generally increase classification reliability. The choice of a conjunctive versus a complementary rule forces us to consider which type of error is preferable – false identification of examinees as Masters or as Non-Masters.

Cronbach et al's (1997) admonition concerning the potential error introduced by adding more tests to a decision in conjunctive decisions was not fulfilled in this study. Overall, the conjunctive decision becomes more reliable as more tests are added. Cronbach et al, however, assumed that tests were unrelated in making calculations. This study shows that, for the overall group, the percentage of examinees correctly classified increases with the addition of related tests. Cronbach et al also estimated reliability for one individual at the greatest risk for misclassification, whereas this study looked at the impact on the overall group.

*Providing Multiple Opportunities to Pass Produces Mixed Results. A*llowing examinees multiple attempts presents a good example of unanticipated findings in the simulation method. When considering classification accuracy, allowing retests is clearly related to lower reliability for all decision rules. However, the opposite is true when considering classification consistency -- measures of agreement are higher for all types of decisions with the addition of more opportunities to pass the test. This

presents a conundrum in terms of recommendations. On the one hand, we don't have true scores in actual testing situations so consistency is the more appropriate consideration. However, the fact that accuracy decreases with the addition of more tests is concerning. The reason for the conflicting findings may be found in the strong complementary component that overlays all decision rules in the case of multiple attempts to pass. Comparison of the tables for accuracy and consistency (Table 5.25 and 5.26) shows the most marked difference between rates of false positives and false negatives for accuracy, but a more moderate percentage for the analogous measure for consistency (pass one test but fail the other). Clarification of this finding might be helped by more in-depth analysis of the score configurations for misclassified examinees based on replicate scores.

It is also important to recognize that this simulation study cannot address the impact of multiple attempts when examinees receive remediation and further instruction in preparation for subsequent testing. In practical situations the purpose of allowing repeated attempts is to allow students to demonstrate increased knowledge as well to account for error in test scores on one particular attempt.

Highly-Related Tests Yield Higher Classification Reliability. The purpose of adding more tests to a decision rule may be to increase the reliability of the decision. In such a case, the question arises as to what types of tests and test characteristics would most benefit classification reliability. Although the content of the tests will be of central importance, the similarity of the tests is another consideration. Is it better to add *similar* tests to reinforce the information already obtained, or *dissimilar* tests that bring new information to the decision? Simulation results suggest that increasing the

covariance among tests results in higher agreement unless the tests are combined in a complementary fashion. An unexpected finding, however, is that the percentage of examinees who pass the decision rule actually *decreases* as covariance increases.

Some Complex Rules Provide Poorer Classification Reliability Than Individual Tests. In terms of the absolute levels of agreement demonstrated in the simulations, measures of accuracy vary from a low of 49% in the complementary condition with three attempts to 99% in the same condition. Consistency estimates vary from 71% to 99%. Therefore, there are rules that decrease, as well as increase, classification reliability beyond that obtained for a single test (90% for accuracy; 86% for consistency).

Accuracy Varies According to Test Difficulty. Two students at different levels of proficiency who take the same set of tests to which the same passing criteria are applied may have a different likelihood of correct classification. This finding highlights the potential unfairness inherent in applying both simple and complex decision rules, and therefore has important policy implications. Furthermore, simulation results suggest that difficult tests provide higher accuracy for Non-Masters, but lower accuracy for Masters. In the current climate in which increasing pressure is being applied to make tests more challenging, this finding provides useful information for consideration by policy makers.

Future Directions

Findings from this study are encouraging, and support the consideration of additional questions that can be answered using the proposed general method. Some questions extend the classical test theory approach to address the importance of

distributional assumptions and test characteristics, as well as provide more in-depth study of the nature of classification error. Other questions extend the method to encompass other measurement models that are found in practical testing situations.

- The examples in the study all used the same level of test reliability, and classification reliability is strongly related to test reliability. It would be informative to construct simulations to investigate the impact of lower test reliability. In particular, what is the effect of adding a less reliable test to a set of highly reliable tests? This corresponds to a real world example in which a performance based test may be added to a decision to increase validity, but it may have a negative impact on reliability.
- The multiple attempt conditions in this study are structured so that the examinee's true score remains the same throughout all retests. This is an unlikely assumption in real life testing situations. Efforts to model change in true score, perhaps using mixture models in which examinees vary in how much true score changes between testings, would be more authentic.
- Further analysis of the types of incorrect decisions that are made would be informative. For example, are there unusual test score configurations that are more likely to receive incorrect decisions? Such analysis could be useful in practical situations to identify examinees for which additional information would help to improve the reliability of the decision.
- The assumption of normality is central to classical test theory, but may not be tenable for actual test scores. Simulation studies can be structured to investigate the effect of departures from normality on classification

reliability. In addition, the utility of the simulation method for decisions with multiple cut-scores requires more accurate modeling of the full score distribution. The development of methods for modeling multivariate non-normal distributions, either through specification of additional moments or mixture modeling, is an important requirement in extending the simulation method.

- This study explored the modeling of data that can be appropriately addressed using classical test theory. However, the method is equally applicable to other measurement models, such as item response theory and latent class analysis. Such application would extend the utility of the method to a variety of practical applications that use performance-based assessments to assess examinee knowledge and skills. In both classical test theory and item response theory the underlying trait is modeled as a continuous scale. Classification of examinees using a latent class approach, in which the underlying scale is categorical, could also prove useful in categorizing examinees and could be accommodated using the general method.

The use of test scores to make educational decisions about examinees is likely to continue in the foreseeable future. Measurement specialists, teachers, parents, and examinees are rightfully concerned about the potential error inherent in all test scores on which important decisions are based. The motivation for this study is to provide a useful methodology for improving the reliability of such decisions. Given the

complexity of findings for the conditions investigated in this study, the practical recommendation is made to use simulation methods to investigate decision rules and maximize classification reliability through the choice of tests, configural strategies, and number of opportunities to retest before implementing or changing policies that dictate important outcomes for examinees.

Appendix I: Contingency Tables with Counts

INDIVIDUAL TESTS: COVAR.6, 50%							
Accuracy				**Consistency**			
TEST 1	REPLICATE 1			**TEST 1**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	225737	24283	250020	FAIL	215965	33986	249951
PASS	24214	225766	249980	PASS	33970	216079	250049
TOTAL	249951	250049	500000	TOTAL	249935	250065	500000
TEST 2	REPLICATE 1			**TEST 2**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	225414	24436	249850	FAIL	215576	34282	249858
PASS	24444	225706	250150	PASS	34502	215640	250142
TOTAL	249858	250142	500000	TOTAL	250078	249922	500000
TEST 3	REPLICATE 1			**TEST 3**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	225701	24117	249818	FAIL	215913	34023	249936
PASS	24235	225947	250182	PASS	33946	216118	250064
TOTAL	249936	250064	500000	TOTAL	249859	250141	500000
TEST 4	REPLICATE 1			**TEST 4**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	225171	24390	249561	FAIL	215416	34206	249622
PASS	24451	225988	250439	PASS	34032	216346	250378
TOTAL	249622	250378	500000	TOTAL	249448	250552	500000
TEST 5	REPLICATE 1			**TEST 5**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	225477	24427	249904	FAIL	215726	34078	249804
PASS	24327	225769	250096	PASS	34030	216166	250196
TOTAL	249804	250196	500000	TOTAL	249756	250244	500000

INDIVIDUAL TESTS: COVAR.6, 70%							
Accuracy				**Consistency**			
TEST 1	REPLICATE 1			**TEST 1**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	130752	19341	150093	FAIL	124243	30048	154291
PASS	23539	326368	349907	PASS	30035	315674	345709
TOTAL	154291	345709	500000	TOTAL	154278	345722	500000
TEST 2	REPLICATE 1			**TEST 2**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	129901	19393	149294	FAIL	123304	30387	153691
PASS	23790	326916	350706	PASS	30322	315987	346309
TOTAL	153691	346309	500000	TOTAL	153626	346374	500000
TEST 3	REPLICATE 1			**TEST 3**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	130598	19337	149935	FAIL	123718	30643	154361
PASS	23763	326302	350065	PASS	30356	315283	345639
TOTAL	154361	345639	500000	TOTAL	154074	345926	500000
TEST 4	REPLICATE 1			**TEST 4**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	130238	19313	149551	FAIL	123606	30342	153948
PASS	23710	326739	350449	PASS	30066	315986	346052
TOTAL	153948	346052	500000	TOTAL	153672	346328	500000
TEST 5	REPLICATE 1			**TEST 5**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	130515	19121	149636	FAIL	123785	30260	154045
PASS	23530	326834	350364	PASS	30222	315733	345955
TOTAL	154045	345955	500000	TOTAL	154007	345993	500000

RESEARCH QUESTION 1: COUNTS FOR ONE TO FIVE TESTS

CONJUNCTIVE: RQ1, COVAR.6, 50%							
Accuracy				**Consistency**			
2 TESTS	REPLICATE 1			**2 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	303162	20428	323590	FAIL	297526	31221	328747
PASS	25585	150825	176410	PASS	31524	139729	171253
TOTAL	328747	171253	500000	TOTAL	329050	170950	500000
3 TESTS	REPLICATE 1			**3 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	343695	16757	360452	FAIL	340931	27490	368421
PASS	24726	114822	139548	PASS	27565	104014	131579
TOTAL	368421	131579	500000	TOTAL	368496	131504	500000
4 TESTS	REPLICATE 1			**4 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	368991	14191	383182	FAIL	368190	24315	392505
PASS	23514	93304	116818	PASS	24440	83055	107495
TOTAL	392505	107495	500000	TOTAL	392630	107370	500000
5 TESTS	REPLICATE 1			**5 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	386420	12223	398643	FAIL	386987	21704	408691
PASS	22271	79086	101357	PASS	21890	69419	91309
TOTAL	408691	91309	500000	TOTAL	408877	91123	500000

COMPLEMENTARY: RQ1, COVAR.6, 50%							
Accuracy				**Consistency**			
2 TESTS	REPLICATE 1				REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	150778	25502	176280	FAIL	139561	31501	171062
PASS	20284	303436	323720	PASS	31402	297536	328938
TOTAL	171062	328938	500000	TOTAL	170963	329037	500000
3 TESTS	REPLICATE 1				REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	114875	24410	139285	FAIL	103971	27592	131563
PASS	16688	344027	360715	PASS	27237	341200	368437
TOTAL	131563	368437	500000	TOTAL	131208	368792	500000
4 TESTS	REPLICATE 1				REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	93404	23160	116564	FAIL	83007	24408	107415
PASS	14011	369425	383436	PASS	24073	368512	392585
TOTAL	107415	392585	500000	TOTAL	107080	392920	500000
5 TESTS	REPLICATE 1				REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	79282	22027	101309	FAIL	69484	21909	91393
PASS	12111	386580	398691	PASS	21554	387053	408607
TOTAL	91393	408607	500000	TOTAL	91038	408962	500000

COMPENSATORY: RQ1, COVAR.6, 50%							
Accuracy				**Consistency**			
2 TESTS	REPLICATE 1			**2 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	230296	19609	249905	FAIL	222517	27121	249638
PASS	19342	230753	250095	PASS	27555	222807	250362
TOTAL	249638	250362	500000	TOTAL	250072	249928	500000
3 TESTS	REPLICATE 1			**3 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	233207	16678	249885	FAIL	226071	23577	249648
PASS	16441	233674	250115	PASS	23779	226573	250352
TOTAL	249648	250352	500000	TOTAL	249850	250150	500000
4 TESTS	REPLICATE 1			**4 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	234763	14934	249697	FAIL	228541	20853	249394
PASS	14631	235672	250303	PASS	21108	229498	250606
TOTAL	249394	250606	500000	TOTAL	249649	250351	500000
5 TESTS	REPLICATE 1			**5 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	235805	13576	249381	FAIL	230344	19086	249430
PASS	13625	236994	250619	PASS	19138	231432	250570
TOTAL	249430	250570	500000	TOTAL	249482	250518	500000

CONJUNCTIVE-COMPLEMENTARY: RQ1, COVAR.6, 50%							
Accuracy				**Consistency**			
4 TESTS	REPLICATE 1			**4 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	318985	18841	337826	FAIL	314834	29479	344313
PASS	25328	136846	162174	PASS	29670	126017	155687
TOTAL	344313	155687	500000	TOTAL	344504	155496	500000
5 TESTS	REPLICATE 1			**5 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	311029	19561	330590	FAIL	306321	30024	336345
PASS	25316	144094	169410	PASS	30442	133213	163655
TOTAL	336345	163655	500000	TOTAL	336763	163237	500000

CONJUNCTIVE-COMPOSITE: RQ1, COVAR.6, 50%							
Accuracy				**Consistency**			
3 TESTS	REPLICATE 1			**3 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	367885	13881	381766	FAIL	364042	21896	385938
PASS	18053	100181	118234	PASS	21885	92177	114062
TOTAL	385938	114062	500000	TOTAL	385927	114073	500000
5 TESTS	REPLICATE 1			**5 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	394484	10812	405296	FAIL	394118	19248	413366
PASS	18882	75822	94704	PASS	19461	67173	86634
TOTAL	413366	86634	500000	TOTAL	413579	86421	500000

CONJUNCTIVE: RQ1, COVAR.6, 70%							
Accuracy				**Consistency**			
2 TESTS	REPLICATE 1			**2 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	193024	20150	213174	FAIL	188794	34075	222869
PASS	29845	256981	286826	PASS	34150	242981	277131
TOTAL	222869	277131	500000	TOTAL	222944	277056	500000
3 TESTS	REPLICATE 1			**3 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	232090	18861	250951	FAIL	230094	34188	264282
PASS	32192	216857	249049	PASS	33953	201765	235718
TOTAL	264282	235718	500000	TOTAL	264047	235953	500000
4 TESTS	REPLICATE 1			**4 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	259217	17630	276847	FAIL	259274	33104	292378
PASS	33161	189992	223153	PASS	32987	174635	207622
TOTAL	292378	207622	500000	TOTAL	292261	207739	500000
5 TESTS	REPLICATE 1			**5 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	279626	16442	296068	FAIL	281127	31791	312918
PASS	33292	170640	203932	PASS	31790	155292	187082
TOTAL	312918	187082	500000	TOTAL	312917	187083	500000

COMPLEMENTARY: RQ1, COVAR.6, 70%							
Accuracy				**Consistency**			
2 TESTS	REPLICATE 1				REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	69968	16245	86213	FAIL	63563	21550	85113
PASS	15145	398642	413787	PASS	21397	393490	414887
TOTAL	85113	414887	500000	TOTAL	84960	415040	500000
3 TESTS	REPLICATE 1				REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	46634	13635	60269	FAIL	40970	16291	57261
PASS	10627	429104	439731	PASS	16207	426532	442739
TOTAL	57261	442739	500000	TOTAL	57177	442823	500000
4 TESTS	REPLICATE 1				REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	34571	11797	46368	FAIL	29749	12853	42602
PASS	8031	445601	453632	PASS	12755	444643	457398
TOTAL	42602	457398	500000	TOTAL	42504	457496	500000
5 TESTS	REPLICATE 1				REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	27311	10418	37729	FAIL	23136	10487	33623
PASS	6312	455959	462271	PASS	10347	456030	466377
TOTAL	33623	466377	500000	TOTAL	33483	466517	500000

COMPENSATORY: RQ1, COVAR.6, 70%							
Accuracy				Consistency			
2 TESTS	REPLICATE 1			**2 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	124037	15131	139168	FAIL	118778	23254	142032
PASS	17995	342837	360832	PASS	23302	334666	357968
TOTAL	142032	357968	500000	TOTAL	142080	357920	500000
3 TESTS	REPLICATE 1			**3 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	122102	12733	134835	FAIL	117441	19604	137045
PASS	14943	350222	365165	PASS	19586	343369	362955
TOTAL	137045	362955	500000	TOTAL	137027	362973	500000
4 TESTS	REPLICATE 1			**4 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	121030	11360	132390	FAIL	116803	17445	134248
PASS	13218	354392	367610	PASS	17337	348415	365752
TOTAL	134248	365752	500000	TOTAL	134140	365860	500000
5 TESTS	REPLICATE 1			**5 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	120626	10416	131042	FAIL	116891	15588	132479
PASS	11853	357105	368958	PASS	15510	352011	367521
TOTAL	132479	367521	500000	TOTAL	132401	367599	500000

CONJUNCTIVE-COMPLEMENTARY: RQ1, COVAR.6, 70%							
Accuracy				**Consistency**			
4 TESTS	REPLICATE 1			**4 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	205581	19478	225059	FAIL	202358	33635	235993
PASS	30412	244529	274941	PASS	33639	230368	264007
TOTAL	235993	264007	500000	TOTAL	235997	264003	500000
5 TESTS	REPLICATE 1			**5 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	194552	23817	218369	FAIL	188794	34075	222869
PASS	28317	253314	281631	PASS	34150	242981	277131
TOTAL	222869	277131	500000	TOTAL	222944	277056	500000

CONJUNCTIVE-COMPENSATORY: RQ1, COVAR.6, 70%							
Accuracy				**Consistency**			
3 TESTS	REPLICATE 1			**3 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	257051	16802	273853	FAIL	253570	29354	282924
PASS	25873	200274	226147	PASS	29250	187826	217076
TOTAL	282924	217076	500000	TOTAL	282820	217180	500000
5 TESTS	REPLICATE 1			**5 TESTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	288314	15104	303418	FAIL	288724	29236	317960
PASS	29646	166936	196582	PASS	29443	152597	182040
TOTAL	317960	182040	500000	TOTAL	318167	181833	500000

RESEARCH QUESTION 2: COMPARING COVAR6 TO COVAR9 FOR FIVE TESTS

CONJUNCTIVE: RQ2, FIVE TESTS, 50%							
ACCURACY				CONSISTENCY			
	COVAR6				COVAR6		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	386420	12223	398643	FAIL	386987	21704	408691
PASS	22271	79086	101357	PASS	21890	69419	91309
TOTAL	408691	91309	500000	TOTAL	408877	91123	500000
	COVAR9				COVAR9		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	315881	7872	323753	FAIL	325976	23765	349741
PASS	33860	142387	176247	PASS	23946	126313	150259
TOTAL	349741	150259	500000	TOTAL	349922	150078	500000

COMPLEMENTARY: RQ2, FIVE TESTS, 50%							
ACCURACY				CONSISTENCY			
	COVAR6				COVAR6		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	79282	22027	101309	FAIL	69484	21909	91393
PASS	12111	386580	398691	PASS	21554	387053	408607
TOTAL	91393	408607	500000	TOTAL	91038	408962	500000
	COVAR9				COVAR9		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	142541	34096	176637	FAIL	126715	23725	150440
PASS	7899	315464	323363	PASS	23995	325565	349560
TOTAL	150440	349560	500000	TOTAL	150710	349290	500000

COMPENSATORY: RQ2, FIVE TESTS, 50%							
ACCURACY				CONSISTENCY			
	COVAR6				COVAR6		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	235805	13576	249381	FAIL	230344	19086	249430
PASS	13625	236994	250619	PASS	19138	231432	250570
TOTAL	249430	250570	500000	TOTAL	249482	250518	500000
	COVAR9				COVAR9		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	238926	11693	250619	FAIL	234060	16538	250598
PASS	11672	237709	249381	PASS	16350	233052	249402
TOTAL	250598	249402	500000	TOTAL	250410	249590	500000

CONJUNCTIVE-COMPLEMENTARY: RQ2, FIVE TESTS, 50%							
ACCURACY				CONSISTENCY			
	COVAR6				COVAR6		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	311029	19561	330590	FAIL	306321	30024	336345
PASS	25316	144094	169410	PASS	30442	133213	163655
TOTAL	336345	163655	500000	TOTAL	336763	163237	500000
	COVAR9				COVAR9		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	274925	14443	289368	FAIL	275054	28294	303348
PASS	28423	182209	210632	PASS	28358	168294	196652
TOTAL	303348	196652	500000	TOTAL	303412	196588	500000

CONJUNCTIVE-COMPENSATORY: RQ2, FIVE TESTS, 50%							
ACCURACY				CONSISTENCY			
	COVAR6				COVAR6		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	394484	10812	405296	FAIL	394118	19248	413366
PASS	18882	75822	94704	PASS	19461	67173	86634
TOTAL	413366	86634	500000	TOTAL	413579	86421	500000
	COVAR9				COVAR9		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	345162	8042	353204	FAIL	346188	18163	364351
PASS	19189	127607	146796	PASS	18224	117425	135649
TOTAL	364351	135649	500000	TOTAL	364412	135588	500000

CONJUNCTIVE: RQ2, FIVE TESTS, 70%							
ACCURACY				CONSISTENCY			
	COVAR6				COVAR6		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	279626	16442	296068	FAIL	281127	31791	312918
PASS	33292	170640	203932	PASS	31790	155292	187082
TOTAL	312918	187082	500000	TOTAL	312917	187083	500000
	COVAR9				COVAR9		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	210050	8306	218356	FAIL	221746	27604	249350
PASS	39300	242344	281644	PASS	27538	223112	250650
TOTAL	249350	250650	500000	TOTAL	249284	250716	500000

COMPLEMENTARY: RQ2, FIVE TESTS, 70%							
ACCURACY				CONSISTENCY			
	COVAR6				COVAR6		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	27311	10418	37729	FAIL	23136	10487	33623
PASS	6312	455959	462271	PASS	10347	456030	466377
TOTAL	33623	466377	500000	TOTAL	33483	466517	500000
	COVAR9				COVAR9		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	67643	22105	89748	FAIL	58012	15380	73392
PASS	5749	404503	410252	PASS	15416	411192	426608
TOTAL	73392	426608	500000	TOTAL	73428	426572	500000

COMPENSATORY: RQ2, FIVE TESTS, 70%							
ACCURACY				CONSISTENCY			
	COVAR6				COVAR6		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	120626	10416	131042	FAIL	116891	15588	132479
PASS	11853	357105	368958	PASS	15510	352011	367521
TOTAL	132479	367521	500000	TOTAL	132401	367599	500000
	COVAR9				COVAR9		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	136725	9669	146394	FAIL	133050	14248	147298
PASS	10573	343033	353606	PASS	14364	338338	352702
TOTAL	147298	352702	500000	TOTAL	147414	352586	500000

CONJUNCTIVE-COMPLEMENTARY: RQ2, FIVE TESTS, 70%							
ACCURACY				CONSISTENCY			
	COVAR6				COVAR6		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	194552	23817	218369	FAIL	188794	34075	222869
PASS	28317	253314	281631	PASS	34150	242981	277131
TOTAL	222869	277131	500000	TOTAL	222944	277056	500000
	COVAR9				COVAR9		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	170954	13328	184282	FAIL	172236	28855	201091
PASS	30137	285581	315718	PASS	28975	269934	298909
TOTAL	201091	298909	500000	TOTAL	201211	298789	500000

CONJUNCTIVE-COMPENSATORY: RQ2, FIVE TESTS, 70%							
ACCURACY				CONSISTENCY			
	COVAR6				COVAR6		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	288314	15104	303418	FAIL	288724	29236	317960
PASS	29646	166936	196582	PASS	29443	152597	182040
TOTAL	317960	182040	500000	TOTAL	318167	181833	500000
	COVAR9				COVAR9		
5 TESTS	REPLICATE 1			5 TESTS	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	240899	8791	249690	FAIL	242966	22104	265070
PASS	24171	226139	250310	PASS	21854	213076	234930
TOTAL	265070	234930	500000	TOTAL	264820	235180	500000

RESEARCH QUESTION 3: COMPARING TWO AND THREE ATTEMPTS

CONJUNCTIVE: RQ3, FIVE TESTS, COVAR6, 50%							
Accuracy				**Consistency**			
2 ATTEMPTS	REPLICATE 1			**2 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	369697	28946	398643	FAIL	354236	22268	376504
PASS	6807	94550	101357	PASS	21982	101514	123496
TOTAL	376504	123496	500000	TOTAL	376218	123782	500000
3 ATTEMPTS	REPLICATE 1			**3 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	356153	42490	398643	FAIL	336650	21960	358610
PASS	2457	98900	101357	PASS	21461	119929	141390
TOTAL	358610	141390	500000	TOTAL	358111	141889	500000

COMPLEMENTARY: RQ3, FIVE TESTS, COVAR6, 50%							
Accuracy				**Consistency**			
2 ATTEMPTS	REPLICATE 1			**2 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	66722	34587	101309	FAIL	53946	15538	69484
PASS	2762	395929	398691	PASS	15595	414921	430516
TOTAL	69484	430516	500000	TOTAL	69541	430459	500000
3 ATTEMPTS	REPLICATE 1			**3 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	58888	42421	101309	FAIL	46732	12967	59699
PASS	811	397880	398691	PASS	12853	427448	440301
TOTAL	59699	440301	500000	TOTAL	59585	440415	500000

COMPENSATORY: RQ3, FIVE TESTS, COVAR6, 50%							
Accuracy				**Consistency**			
2 ATTEMPTS	REPLICATE 1			**2 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	206455	42926	249381	FAIL	191915	15303	207218
PASS	763	249856	250619	PASS	15418	277364	292782
TOTAL	207218	292782	500000	TOTAL	207333	292667	500000
3 ATTEMPTS	REPLICATE 1			**3 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	186630	62751	249381	FAIL	173233	13431	186664
PASS	34	250585	250619	PASS	13458	299878	313336
TOTAL	186664	313336	500000	TOTAL	186691	313309	500000

CONJUNCTIVE-COMPLEMENTARY: RQ3, FIVE TESTS, COVAR6, 50%							
Accuracy				**Consistency**			
2 ATTEMPTS	REPLICATE 1			**2 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	291464	39126	330590	FAIL	271669	27293	298962
PASS	7498	161912	169410	PASS	27548	173490	201038
TOTAL	298962	201038	500000	TOTAL	299217	200783	500000
3 ATTEMPTS	REPLICATE 1			**3 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	276613	53977	330590	FAIL	253789	25512	279301
PASS	2688	166722	169410	PASS	25609	195090	220699
TOTAL	279301	220699	500000	TOTAL	279398	220602	500000

CONJUNCTIVE-COMPENSATORY: RQ3, FIVE TESTS, COVAR6, 50%							
Accuracy				**Consistency**			
2 ATTEMPTS	REPLICATE 1			**2 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	376808	28488	405296	FAIL	362205	19775	381980
PASS	5172	89532	94704	PASS	19573	98447	118020
TOTAL	381980	118020	500000	TOTAL	381778	118222	500000
3 ATTEMPTS	REPLICATE 1			**3 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	362710	42586	405296	FAIL	345013	19448	364461
PASS	1751	92953	94704	PASS	19084	116455	135539
TOTAL	364461	135539	500000	TOTAL	364097	135903	500000

CONJUNCTIVE: RQ3, FIVE TESTS, COVAR6, 70%							
Accuracy				**Consistency**			
2 ATTEMPTS	REPLICATE 1			**2 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	259049	37019	296068	FAIL	240513	28197	268710
PASS	9661	194271	203932	PASS	28350	202940	231290
TOTAL	268710	231290	500000	TOTAL	268863	231137	500000
3 ATTEMPTS	REPLICATE 1			**3 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	243057	53011	296068	FAIL	220740	25711	246451
PASS	3394	200538	203932	PASS	26012	227537	253549
TOTAL	246451	253549	500000	TOTAL	246752	253248	500000

COMPLEMENTARY: RQ3, FIVE TESTS, COVAR6, 70%							
Accuracy				**Consistency**			
2 ATTEMPTS	REPLICATE 1			**2 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	21804	15925	37729	FAIL	16388	6748	23136
PASS	1332	460939	462271	PASS	6826	470038	476864
TOTAL	23136	476864	500000	TOTAL	23214	476786	500000
3 ATTEMPTS	REPLICATE 1			**3 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	18536	19193	37729	FAIL	13604	5308	18912
PASS	376	461895	462271	PASS	5317	475771	481088
TOTAL	18912	481088	500000	TOTAL	18921	481079	500000

COMPENSATORY: RQ3, FIVE TESTS, COVAR6, 70%							
Accuracy				**Consistency**			
2 ATTEMPTS	REPLICATE 1			**2 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	98790	32252	131042	FAIL	88368	11075	99443
PASS	653	368305	368958	PASS	11027	389530	400557
TOTAL	99443	400557	500000	TOTAL	99395	400605	500000
3 ATTEMPTS	REPLICATE 1			**3 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	84941	46101	131042	FAIL	75882	9082	84964
PASS	23	368935	368958	PASS	9072	405964	415036
TOTAL	84964	415036	500000	TOTAL	84954	415046	500000

CONJUNCTIVE-COMPLEMENTARY: RQ3, FIVE TESTS, COVAR6, 70%							
Accuracy				**Consistency**			
2 ATTEMPTS	REPLICATE 1			**2 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	180138	38231	218369	FAIL	161598	27010	188608
PASS	8470	273161	281631	PASS	27157	284235	311392
TOTAL	188608	311392	500000	TOTAL	188755	311245	500000
3 ATTEMPTS	REPLICATE 1			**3 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	166564	51805	218369	FAIL	145890	23575	169465
PASS	2901	278730	281631	PASS	23956	306579	330535
TOTAL	169465	330535	500000	TOTAL	169465	330154	500000

CONJUNCTIVE-COMPENSATORY: RQ3, FIVE TESTS, COVAR6, 70%							
Accuracy				**Consistency**			
2 ATTEMPTS	REPLICATE 1			**2 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	266321	37097	303418	FAIL	248160	26101	274261
PASS	7940	188642	196582	PASS	26069	199670	225739
TOTAL	274261	225739	500000	TOTAL	274229	225771	500000
3 ATTEMPTS	REPLICATE 1			**3 ATTEMPTS**	REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	249177	54241	303418	FAIL	228271	23615	251886
PASS	2709	193873	196582	PASS	23968	224146	248114
TOTAL	251886	248114	500000	TOTAL	252239	247761	500000

RESEARCH QUESTION 4: GED TESTS

Observed Split-Half Scores				Simulated Half-Length Tests			
Writing							
	REPLICATE 2				REPLICATE 2		
REPLICATE 1	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	19623	8134	27757	FAIL	16333	8526	24859
PASS	10641	72593	83234	PASS	8485	66656	75141
Total	30264	80727	110991	Total	24818	75182	100000
Social Studies							
	REPLICATE 2				REPLICATE 2		
REPLICATE 1	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	6158	4125	10283	FAIL	4638	3364	8002
PASS	3781	96927	100708	PASS	3284	88714	91998
Total	9939	101052	110991	Total	7922	92078	100000
Science							
	REPLICATE 2				REPLICATE 2		
REPLICATE 1	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	6416	4709	11125	FAIL	5356	4444	9800
PASS	3675	96191	99866	PASS	4494	85706	90200
Total	10091	100900	110991	Total	9850	90150	100000
Reading							
	REPLICATE 2				REPLICATE 2		
REPLICATE 1	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	2854	2259	5113	FAIL	1896	2438	4334
PASS	1860	104018	105878	PASS	2482	93184	95666
Total	4714	106277	110991	Total	4378	95622	100000
Math							
	REPLICATE 2				REPLICATE 2		
REPLICATE 1	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	17815	7581	25396	FAIL	17013	7526	24539
PASS	7537	78058	85595	PASS	7560	67901	75461
Total	25352	85639	110991	Total	24573	75427	100000
Overall Rule							
	REPLICATE 2				REPLICATE 2		
REPLICATE 1	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	34124	8508	42632	FAIL	32019	8599	40618
PASS	9669	58690	68359	PASS	8582	50800	59382
Total	43793	67198	110991	Total	40601	59399	100000

SIMULATED, FULL-LENGTH GED TESTS							
Accuracy				**Consistency**			
Writing							
	REPLICATE 1				REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	19829	3893	23722	FAIL	18712	6533	25245
PASS	5416	70862	76278	PASS	6516	68239	74755
TOTAL	25245	74755	100000	TOTAL	25228	74772	100000
Social Studies							
	REPLICATE 1				REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	5921	1265	7186	FAIL	5638	2557	8195
PASS	2274	90540	92814	PASS	2584	89221	91805
TOTAL	8195	91805	100000	TOTAL	8222	91778	100000
Science							
	REPLICATE 1				REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	6968	1696	8664	FAIL	6695	3476	10171
PASS	3203	88133	91336	PASS	3473	86356	89829
TOTAL	10171	89829	100000	TOTAL	10168	89832	100000
Reading							
	REPLICATE 1				REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	2630	836	3466	FAIL	2548	2005	4553
PASS	1923	94611	96534	PASS	2013	93434	95447
TOTAL	4553	95447	100000	TOTAL	4561	95439	100000
Math							
	REPLICATE 1				REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	20088	3324	23412	FAIL	19012	5562	24574
PASS	4486	72102	76588	PASS	5472	69954	75426
TOTAL	24574	75426	100000	TOTAL	24484	75516	100000
Overall Rule							
	REPLICATE 1				REPLICATE 2		
TRUE	FAIL	PASS	TOTAL	REPLICATE 1	FAIL	PASS	TOTAL
FAIL	32527	2713	35240	FAIL	32747	6219	38966
PASS	6439	58321	64760	PASS	5982	55052	61034
TOTAL	38966	61034	100000	TOTAL	38729	61271	100000

Appendix II: Computer Code for R-Software Programs

GENERATING SIMULATION DATA FOR CHAPTER 5

COVAR0:

```
N=10000     #number of examinees
T=5   #number of tests
true<-array(0,c(N,T))

mu=c(0,0,0,0,0)
sigma = matrix(c(1,0,0,0,0,
                 0,1,0,0,0,
                 0,0,1,0,0,
                 0,0,0,1,0,
                 0,0,0,0,1),5,5)
for (i in 1:N) {
  for (j in 1:T){
  true[i,j]<-mvrnorm(N,mu,sigma) }}

obsc1<-array(0,c(N,T))  #first replicate score
obsc2<-array(0,c(N,T))  #second replicate score

SEM<-.31623    #reliability = .9

for (i in 1:N) {
  for (j in 1:T) {
  obsc1[i,j] = rnorm(1,true[i,j],SEM)  #generate first replicate score
  }}

for (i in 1:N) {
  for (j in 1:T) {
  obsc2[i,j] = rnorm(1,true[i,j],SEM)  #generate second replicate score
  }}
```

COVAR6

```
N=500000     #number of examinees
T=5   #number of tests
mu=c(0,0,0,0,0)
sigma = matrix(c(1,.6,.6,.6,.6,
                 .6,1,.6,.6,.6,
                 .6,.6,1,.6,.6,
                 .6,.6,.6,1,.6,
```

```
            .6,.6,.6,.6,1),5,5)

crit<-c(0,0,0,0,0)

true=mvrnorm(n=N,mu,sigma)

obsc1<-array(0,c(N,T))  #first replicate score
obsc2<-array(0,c(N,T))  #second replicate score

SEM<-.31623    #reliability = .9

for (i in 1:N) {
  for (j in 1:T) {
  obsc1[i,j] = rnorm(1,true[i,j],SEM)  #generate first replicate score
  }}

for (i in 1:N) {
  for (j in 1:T) {
  obsc2[i,j] = rnorm(1,true[i,j],SEM)  #generate second replicate score
  }}

for (i in 1:N) {
  for (j in 1:T) {
  obsc3[i,j] = rnorm(1,true[i,j],SEM)  #generate third replicate score
  }}

for (i in 1:N) {
  for (j in 1:T) {
  obsc4[i,j] = rnorm(1,true[i,j],SEM)  #generate FOURTH replicate score
  }}

for (i in 1:N) {
  for (j in 1:T) {
  obsc5[i,j] = rnorm(1,true[i,j],SEM)  #generate FIFTH replicate score
  }}

for (i in 1:N) {
  for (j in 1:T) {
  obsc6[i,j] = rnorm(1,true[i,j],SEM)  #generate SIXTH replicate score
  }}
```

COVAR9

```
N=500000     #number of examinees
T=5   #number of tests
```

```
mu=c(0,0,0,0,0)
sigma = matrix(c(1,.9,.9,.9,.9,
                 .9,1,.9,.9,.9,
                 .9,.9,1,.9,.9,
                 .9,.9,.9,1,.9,
                 .9,.9,.9,.9,1),5,5)

crit<-c(0,0,0,0,0)

true=mvrnorm(n=N,mu,sigma)

obsc1<-array(0,c(N,T))  #first replicate score
obsc2<-array(0,c(N,T))  #second replicate score

SEM<-.31623     #reliability = .9

for (i in 1:N) {
  for (j in 1:T) {
  obsc1[i,j] = rnorm(1,true[i,j],SEM)  #generate first replicate score
  }}

for (i in 1:N) {
  for (j in 1:T) {
  obsc2[i,j] = rnorm(1,true[i,j],SEM)  #generate second replicate score
  }}
```

GENERATING SIMULATION DATA FOR GED EXAMPLE

HALF-LENGTH TESTS

```
N=100000     #number of examinees
T=5   #number of tests
mu=c(-.00598, -.00551, -.00752, -.01128, -.00118)
sigma = matrix(c(.745073, .603844, .59741, .576515, .545608,
              .603844, .799204, .692102, .655146, .564010,
              .59741, .692102, .747666, .603774, .591530,
              .576515, .655146, .603774, .708481, .470365,
              .545608, .564010, .591530, .470365, .806708),5,5)


true=mvrnorm(n=N,mu,sigma)

obsc1<-array(0,c(N,T))  #first replicate score
obsc2<-array(0,c(N,T))  #second replicate score

SEMwrit<-.468    #reliability, writing = .773
SEMss<-.407      #reliability, social studies = .828
SEMsci<-.457     #reliability, science = .782
SEMread<-.477     #reliability, reading = .757
SEMmath<-.422    #reliability, math = .819

for (i in 1:N) {

  obsc1[i,1] = rnorm(1,true[i,1],SEMwrit)  #generate first replicate writing score
  obsc1[i,2] = rnorm(1,true[i,2],SEMss)    #generate first replicate social studies score
  obsc1[i,3] = rnorm(1,true[i,3],SEMsci)   #generate first replicate science score
  obsc1[i,4] = rnorm(1,true[i,4],SEMread)  #generate first replicate reading score
  obsc1[i,5] = rnorm(1,true[i,5],SEMmath)  #generate first replicate math score
  }

for (i in 1:N) {

  obsc2[i,1] = rnorm(1,true[i,1],SEMwrit)  #generate second replicate writing score
  obsc2[i,2] = rnorm(1,true[i,2],SEMss)    #generate second replicate social studies
score
  obsc2[i,3] = rnorm(1,true[i,3],SEMsci)   #generate second replicate science score
  obsc2[i,4] = rnorm(1,true[i,4],SEMread)  #generate second replicate reading score
  obsc2[i,5] = rnorm(1,true[i,5],SEMmath)  #generate second replicate math score
  }
```

FULL-LENGTH TESTS

```
N=100000     #number of examinees
T=5   #number of tests
mu=c(-.00154, -.002522, -.001804, -.004091, -.000382)
sigma = matrix(c(.86315431,.678962, .681302, .669293, .654642,
              .678962, .892009273, .784387, .759690, .635370,
              .681302, .784387, .867977039, .706269, .693413,
              .669293, .759690, .706269, .841411177, .559507,
              .654642, .635370, .693413, .559507, .896791211),5,5)


true=mvrnorm(n=N,mu,sigma)

obsc1<-array(0,c(N,T))  #first replicate score
obsc2<-array(0,c(N,T))  #second replicate score

SEMwrit<-.355952    #reliability, writing = .872
SEMss<-.304218      #reliability, social studies = .906
SEMsci<-.347286     #reliability, science = .878
SEMread<-.36702     #reliability, reading = .862
SEMmath<-.313907    #reliability, math = .901

for (i in 1:N) {

   obsc1[i,1] = rnorm(1,true[i,1],SEMwrit)  #generate first replicate writing score
   obsc1[i,2] = rnorm(1,true[i,2],SEMss)    #generate first replicate social studies score
   obsc1[i,3] = rnorm(1,true[i,3],SEMsci)   #generate first replicate science score
   obsc1[i,4] = rnorm(1,true[i,4],SEMread)  #generate first replicate reading score
   obsc1[i,5] = rnorm(1,true[i,5],SEMmath)  #generate first replicate math score
   }

for (i in 1:N) {

   obsc2[i,1] = rnorm(1,true[i,1],SEMwrit)  #generate second replicate writing score
   obsc2[i,2] = rnorm(1,true[i,2],SEMss)    #generate second replicate social studies
score
   obsc2[i,3] = rnorm(1,true[i,3],SEMsci)   #generate second replicate science score
   obsc2[i,4] = rnorm(1,true[i,4],SEMread)  #generate second replicate reading score
   obsc2[i,5] = rnorm(1,true[i,5],SEMmath)  #generate second replicate math score
   }
```

Bibliography

American Educational Research Association. (1999). *Standards for Educational and Psychological Testing.* Washington, DC.

Bradlow, E. T. & Wainer, H. (1998). Some statistical and logical considerations when rescoring tests. *Statistica Sinica*, 8, 713-728

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

Brennan, R. L. & Kane, M. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement,* 14, 277-289.

Brennan, R. L. & Wan, L. (2004). *A bootstrap procedure for estimating decision consistency for single-administration complex assessments.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Box, G. & Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. New York: John Wiley & Son.

Carver, R. P. (1970). Special problems in measuring change with psychometric devices. In *Evaluative Research: Strategies and Methods.* Pittsburgh, PA: American Institutes for Research.

Chester, M. D. (2003). Multiple measures and high-stakes decisions: A framework for combining measures. *Educational Measurement: Issues and Practice,* 22 (2), National Council on Measurement in Education.

Cicchetti, D. V. & Feinstein, A. R. (1990). High agreement but low kappa : II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43, 551-558.

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

Cronbach, L.J., Linn, R.L., Brennan, R.L., Haertel, E.H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, 57 (3), 373-399.

Feldt, L.S. & Brennan, R. L. (1993). Reliability. In R. L. Linn (Ed.), *Educational Measurement, Third Edition.* Phoenix, AZ: Oryx Press.

Feinstein, A. R. & Cicchetti, D. V. (1990). High agreement but low kappa : I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43, 543-549.

Fitzmaurice, G. (2002). Statistical methods for assessing agreement. *Nutrition*, 18, 694-696.

GED Testing Service (2006). The Tests of General Educational Development Technical Manual. Unpublished manuscript, Washington, DC: American Council on Education.

Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (1998). *Bayesian data analysis.* New York: Chapman & Hall/CRC.

Gong, B. & Hill, R. (2001) [PowerPoint presentation] Some considerations of multiple measures in assessment and school accountability. Presentation at the Seminar on Using Multiple Measures and Indicators to Judge Schools' Adequate Yearly Progress under Title 1. Sponsored by CCSSO and US DOE, Washington, DC, March 23-24, 2001.

Haertel, E. H. & Wiley, D. E. (1993). Representations of ability structures: Implications for testing. In N. Fredericksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Heubert, J.P. & Hauser, R.M. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, D.C.: National Academy Press.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement,* 13, 253-264.

Klein, S. P, & Orlando, M. (2000). *CUNY's testing program: Characteristics, results, and implications for policy and research.* MR-1249-CAE. Santa Monica, CA. RAND.

Livingston, S. A. & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32 (2), 179-197.

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

Louisiana State Department of Education. Accessed November 27, 2005. http://www.doe.state.la.us/lde/uploads/3845.pdf

Maryland State Department of Education. Accessed November 27, 2005. http://www.marylandpublicschools.org/MSDE/testing/hsa/

Marshall, J. L. & Haertel, E. H. (1976). The mean split-half coefficient of agreement: A single administration index of reliability for mastery tests. Manuscript, University of Wisconsin.

Misley, R. (1984). Estimating latent distributions. *Psychometrika*, 49, 359-381.

Nevitt, J. (1998). Simulating univariate and multivariate nonnormal data: An implementation of the methods of Fleishman (1978) and Vale and Maurelli (1983). Department of Measurement, Statistics, and Evaluation: Technical Report.

Peng, C. J. & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement*, 17, 369-368.

R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rogosa, D. (1999). *Accuracy of individual scores expressed in percentile ranks: classical test theory calculations.* CSE Technical Report 509. National Center for Research on Evaluatio, Standards, and Student Testing.

Rudner, L. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research, and Evaluation*, 7 (14). Retrieved April 16, 2004 from http://PAREonline.net/getvn.asp?v=7&n=14.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a mastery test. *Journal of Educational Measurement*, 13, 265-276.

Subkoviak, M. J. (1980). Decision-consistency approaches. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art.* Baltimore, MD: The Johns Hopkins University Press.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement*, 25, pp. 47-55.

Sullivan, P., Yeager, M., Chudowsky, N., Kober, N., O'Brien, E., & Gayler, K. (2005). State high school exit exams: States try harder, but gaps persist.

Center on Education Policy. Downloaded from http://www.cep-dc.org/ on August 1, 2006.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement,* 11, 263-267.

Uebersax, J. Statistical methods for rater agreement. (2003). Downloaded from http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm on August 4, 2006.

Vale, C. D. & Maurelli, V. A. (1983). Simulating multivariate non-normal distributions. *Psychometrika*, 48, 465-471.

Wainer, H. & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local independence on reliability? *Educational Measurement: Issues and Practice*, 15 (1), 22-29.

Wainer, H., Wang, X. A., Skorupski, W. P., & Bradlow, E. T. (2005). A Bayesian method for evaluating passing scores: The PPOP curve. *Journal of Educational Measurement*, 42, 271-281.