University of Alberta

COMPARING THE CORRECTNESS OF CLASSICAL TEST THEORY AND ITEM RESPONSE THEORY IN EVALUATING THE CONSISTENCY AND ACCURACY OF STUDENT PROFICIENCY CLASSIFICATIONS

by

Augustine Metadio Gundula

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Measurement, Evaluation and Cognition

Department of Educational Psychology

Fall 2012
Edmonton, Alberta

Dedication

I thank you Dr W. Todd Rogers for being supportive while I engaged in endeavour of completing a doctoral degree. During the five-years process, you were very understanding. You were there for me in times of my tribulations. You gave me encouragement when I needed it. You gave me hope when I was in a hopeless situation. You gave me light when there was darkness. You were not only my supervisor but you also took the role of a father to me. I love you with my heart and soul. You were truly my soul-mate. When I reflect our relationship, your love gave me strength and courage. I truly have benefited from the strength and courage that you gave me through our relationship. May God bless you.

Abstract

The purposes of this study were: 1) to compare the values of decision consistency (DC) and decision accuracy (DA) yielded by three commonly used estimation procedures: Livingston-Lewis (LL) and the compound multinomial procedure (CM) procedures, both of which are based on classical test theory approach, and Lee's IRT procedure based on item response theory approach and 2) to determine how accurate and precise these procedures are. Two population data sources were used: the Junior Reading (N = 128,103) and Mathematics (N = 127,639) assessments administered by the Education Quality and Accountability Office (EQAO) and the three entrance examinations administered by the University of Malawi (U of M; N = 6,191). To determine the degree of bias and the level of precision for both DC and DA, 100 replicated random samples corresponding to four sample sizes (n = 1,500, 3,000, 4,500, 6,000) for the EQAO populations and two sample sizes (n = 1,500, 3,000) for the U of M population were selected.

At the population level, there was an interaction between the three procedures and the four cut-scores. While the differences between the values of DC and the values of DA among the three procedures tended to be small for one or both extreme cut-scores, the differences tended to be larger when the cut-score

was closer to the population mean. The IRT procedure tended to provide the highest values for both DC and DA, followed in turn by the CM and LL procedures.

At the sample level, the estimates of DC and DA yielded by the three estimation procedures were unbiased and precise. Consequently, the findings at the population are applicable at the sample level. Therefore, based on the findings of the present study, the compound multinomial procedure should be used to determine DC and DA when classical test score theory is used to analyze a test and its items and the IRT procedure should be used to determine DC and DA when item response theory is used to analyze a test and its items.

Acknowledgements

My sincere gratitude goes to Dr W. Todd Rogers, my major supervisor, for his support and guidance extended throughout the process of conducting this study. His continued encouragement and mentoring helped me a lot during my PhD studies. I really appreciate his insights, inspiring words, and sharing his integral view on research. He was always available when I needed help. It would be impossible to place a value, on what I have learned from him. I am deeply honoured to have studied under his supervision.

I would like to thank the members of my dissertation committee: Dr George Buck, Dr William Whelton, Dr Stewart Petersen, Dr Marcel Bouffard, and Dr Barbara Plake, for their valuable suggestions, kind advice, encouragement, and fruitful discussions. I am glad that I have the opportunity to express my gratitude to all of them. Without their careful guidance, this work would not be possible. I was fortunate to have them as my committee.

I give thanks and acknowledgement to EQAO and U of M for granting me the permission to use their data for this study. Without them, it would have been impossible for me to conduct this study.

I also want to express my heartfelt thanks to my family members and friends. They have been an encouragement to me every step along the way and always provided an ear for listening. There are too many of you to name but without your support and gentle nudges along the way I would not have ever have dreamed I could achieve this goal.

Let us not give up in doing good, for in time we will reap a harvest if we do not quit.

Galatians 6:9

Table of Contents

List of Tables

List of Figures

CHAPTER 1: INTRODUCTION

Background of the Problem

The consistency and accuracy of student proficiency classifications must be evaluated if decisions from educational assessment results are to be useful and defensible. One of the most significant purposes of large-scale assessments is to determine whether or not a candidate has been classified in the correct category based on his/her score on the assessment used for this purpose (Livingston & Wingersky, 1979; Livingston & Zieky, 1982; Zhang, 2008). Using assessment results, applicants must be classified into different proficiency levels. For example, in the case of licensure decisions, candidates are either licensed or not. Likewise in public education candidates are classified into different proficiency levels, such as masters versus non-masters or Below Basic, Basic, Proficient, and Advanced (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Guo, 2006; Martineau, 2007). It is vital to note that any measurement instrument has an inherent measurement error. Consequently, the test scores used to classify students according to their abilities also include measurement error (Ercikan, 2006; Lee, 2008; Wang, Bradlow & Wainer, 2002). Therefore, it is vital for educational experts to evaluate the decision accuracy (DA) and decision consistency (DC) in the presence of measurement error so that students are classified in the correct performance-level categories.

Misclassifications usually occur when the score of a student is at the border of the cut-score (Glass, 1978; Rudner, 2005; Wang & Wilson, 2005). For instance, if the score of a student is either just above or just below the cut-score, then the score may change if a parallel form of the test is administered due to measurement error (Berk, 1980; Glaser, 1963; Lord & Novick, 1968). If the score of the student is above the cut-score on the first administration but below the cut-score on the second administration, there is high probability of committing Type II error. In this situation the placement is called false negative. On the other hand, if the score is below the cut-score on the first administration but above the cut-score on the second administration, there is high probability of committing Type I error which leads to false positive classification (Cizek, 2001; Crocker & Algina, 1986; Cronbach, Linn, Brennan & Haertel, 1997). Therefore, it is important to estimate the DA and DC of classifications so as to know the probability that correct or same decisions are made about students' performance. If the values of DA and/or DC are low, then the use of the assessment for classification decisions may not be warranted (Lee, Hanson & Brennan, 2002; Saito, 2003; Subkoviak, 1978).

Definitions of Decision Accuracy and Consistency

Decision accuracy is defined as the degree to which classifications using test scores are the same as the classifications using true scores, which are errorless (Embertson & Reise, 2000; Lee, 2010; Wilcox, 1981). Classification accuracy is used interchangeably with validity of a classification system (Hambleton & Novick, 1973; Hambleton, Swaminathan & Rogers, 1991; Lee, Hanson & Brennan, 2002). In contrast, decision consistency or reliability of classification deals with only observed scores. It refers to the degree to which examinees are placed in the same performance-level categories each time the measurement instrument is employed under similar conditions (Hambleton, Swaminathan Algina & Coulson, 1978; Hambleton, & Traub, 1973; Lord, 1965).

Factors that affect Decision Accuracy and Consistency

There are several factors that affect the values of DA and DC. First is the position of the cut-score in the score distribution relative to the mean for the score distribution (Gelman, Carlin, Stern & Rubin, 1998; Wainer & Thissen, 1996; Wan, Brennan & Lee, 2007). Given that generally a large number of students' scores are close to the mean, more students are liable to be misclassified if the cut-score is closer to the mean of students' scores due to measurement error (Cronbach, 1951; Shepard, 1980; Wainer, Bradlow & Du, 2000). The likelihood of committing a Type I error or Type II error decreases as the distance between the cut-score and the center of a distribution increases in a score distribution. The further the cut-score is from the center of a distribution of a score distribution, the higher the values of DA and DC (Feldt & Brennan, 1993; Haertel & Wiley, 1993; Huynh, 1978).

The second factor is the number of cut-scores. As the number of proficiency levels increases, the number of cut-scores increases (Bourque, Goodman, Hambleton & Han, 2004; Misley, 1984; Sireci, Thissen & Wainer, 1991). For example, the use of three proficiency levels requires two cut-scores while the use of four proficiency levels requires three cut-scores. Given the same score distribution, the distances between three cut-scores will generally be less than the difference between two cut-scores (Popham & Husek, 1969; Reschly, 1981; Resnick, 1980). When the distances between adjacent cut-scores are small, the likelihood of committing type I error or type II error is higher than when distances between adjacent cut-scores are larger (Berk, 1984; Brennan & Lee, 2006a, 2006b; Bradlow, Wainer & Wang, 1999; Traub & Rowley, 1980).

The third factor that influences DA and DC is the length of the test (Li, 2006; Linn, 1979; Rogosa, 1999). As the length of the test increases by adding more items, the error of measurement decreases, which will in turn reduce the number of misclassifications, thereby leading to higher values of DA and DC (Birnbaum, 1968; Spearman, 1910; Wainer, Wang, Skorupski & Bradlow, 2005).

Fourth, the values of indices for DA and DC are affected by the type of scores used (i.e., raw scores or scale scores) (Brennan, 2001; Brennan & Kane, 1977; Hambleton & Swaminathan, 1985). These two types of scores may not always give the same results for decision accuracy and decision consistency indices due to the lack of one-to-one correspondence between the raw and scale scores (Lee, 2005; Lord, 1980).

Estimation Procedures Based on Classical Test Score Theory

Historically, the estimation of decision consistency was considered before the estimation of decision accuracy and was based on classical test score theory (CTST). At first, KR-20 and a corrected split-half reliability were used to estimate internal consistency of examination scores before 1973 (Hambleton, & Slater, 1997). However, Hambleton and Novick (1973) recognized that these two indices were not appropriate for determining the consistency of classifications. They then defined DC as the consistency of classification of the candidate's performance in the correct category of proficiency level resulting from two administrations of the same test or parallel forms of the examination. Consequently the proportion of correct decisions, p_o, or the proportion corrected for chance, κ (Cohen, 1960), made from two administrations of the same test or the administration of parallel forms of the same test on two occasions were used. However it is very difficult to construct two parallel forms of the test that will meet the test specifications and is doubtful that some candidates administered the same test twice will not remember some of the responses to the first administration or will not be available for both administrations. As a result, several procedures for estimating DC for dichotomously-scored items using a single administration of the test were proposed (e.g., Huynh 1976; Marshall & Haertel, 1976; Peng & Subkoviak, 1980; Subkoviak, 1976).

In 1990, Hanson and Brennan proposed a procedure for estimating both DC and DA for dichotomously scored items. Subsequently, several methods for estimating both DC and DA were developed that are applicable for assessments with both dichotomously scored and polytomously-scored items (see Breyer & Lewis, 1994; Lee 2005; Lee, Brennan & Wan, 2009; Livingston & Lewis 1995).

Estimation Procedures Based on Item Response Theory

Due to the increased use of item response theory (IRT) in test development, several procedures for estimating DA and DC have been developed based on IRT. As with CTST and single test administration, the first procedures were developed for dichotomously scored items (e.g., Huynh, 1990; Schulz, Kolen & Nice-wander, 1999). Subsequently, estimation procedures were developed for both dichotomously and polytomously scored items (e.g., Lee, 2008, 2010; Lee, Hanson & Brennan, 2002; Wang, Kolen & Harris, 2000; Rudner, 2001, 2004).

However, studies comparing the various procedures (Hanson & Brennan, 1990; Lee, 2010; Lee, Hanson & Brennan, 2002) have revealed that the procedures do not necessarily lead to the same estimates of DA and DC. Of these studies, only one study (Lee, 2010) compared the correctness of the decisions using procedures based on CTST and procedures based on IRT. The three procedures that Lee considered included the Livingston-Lewis procedure (CTST), compound multinomial procedure (CTST), and the three-parameter IRT model. Lee found that the decision indices obtained using IRT procedures were generally higher than the decisions using the two CTST procedures. The study employed small sample sizes (n = 500 for a mathematics ability test and approximately

4,000 for science test). The scarcity of studies in which these three procedures are compared is regrettable because it is the sort of evidence that testing companies require when defending the accuracy and consistency of the decisions they make when classifying students based on their test scores.

Purpose of the Study

Consequently, the purpose of this study was to compare the correctness of decision accuracy and decision consistency of the following three estimation procedures: Livingston-Lewis procedure (LL) CTST approach, the compound multinomial procedure (CM) CTST approach, and the Lee IRT procedure. The following research questions were addressed:

1. Do the LL, CM, and IRT procedures yield the same results across four cut-scores and sample sizes?
2. To what extent does the cut-score location affect the magnitude of the values of DC and DA obtained using the LL, CM, and IRT procedures?
3. To what extent does the number of examinees around the cut-score affect the magnitude of the values of DC and DA obtained using the LL, CM, and IRT procedures?
4. Are the LL, CM, and IRT procedures equally consistent and accurate across four cut-scores and different sample sizes?

Delimitations

Given the scope and size of the present study, it was not possible to conduct a follow-up study to determine whether or not the placement of the students on the borderline between two performance categories was correct in terms of how much of the subject matter content these students had mastered. Such a study was not feasible in terms of time and the resources needed given the location of the students in a different province and country. However, a follow-up study would be beneficial.

Definition of Terms

Standard setting: a measurement activity in which a procedure is applied to systematically derive one or more cut-scores in the score distribution for a test (Canale & Swain, 1980; Rogers & Ricker, 2006).

Performance standard: the conceptualization of the lowest level of achievement deemed necessary to be given to a performance-level category (Livingston & Zieky, 1982; Rogers & Ricker, 2006).

Masters: Students performances whose scores are equal to or greater than a cut-score (Brennan & Lee, 2006a, 2006b; Rogers & Ricker, 2006).

Non-masters: Students performances whose scores are lower than the cut-score (Livingston & Zieky, 1982; Rogers & Ricker, 2006).

Advanced: a higher level of achievement based on performance standards. The student may also achieve at levels that exceed the grade-level standard (Cizek, 2001; Knupp, 2009).

Proficient**:** a high level of achievement based on performance standards. The student is able to perform at the level of difficult, complexity, or fluency specified in the grade-level standard (Cizek, 2001; Lee, 2005).

Basic: The lowest level of achievement based on performance standards (Brennan & Lee, 2006a, 2006b; Cizek, 2001).

Below Basic: below the lowest level of achievement based on performance standards (Livingston & Zieky, 1982; Rogers & Ricker, 2006).

Cut-score: a point in a score distribution that generates groups representing two or more states or degrees of performance. A cut-score is the numerical operationalization of a performance standard (Bachman & Palmer, 1996; Cizek, 2001; Livingston & Zieky, 1982).

Decision accuracy: the degree to which classifications based on test scores are equivalent to those that could have been made if the scores were errorless (Lee, 2010; Livingston & Zieky, 1982).

Decision consistency: the degree to which classifications based on test scores are equal to the decisions based on scores from a second, parallel form of the same test (Rogers & Ricker, 2006, Lee, 2010).

False Negative: a mistake that arises when the application of a cut-score classifies an examinee as failing when the examinee truly possesses the level of knowledge, skill or ability required for passing (Ercikan, 2006; Lee, 2008, 2010).

False Positive: a mistake that occurs when the application of a cut-score classifies an examinee as passing when the examinee truly does not possess the level of knowledge, skill or ability required for passing (Lee, Brennan & Wan, 2009; Livingston & Zieky, 1982).

Minimum Competency: The ability to display fundamental proficiencies or work effectively during the performance of life roles (Girrbach & Claus, 1982; Livingston & Zieky, 1982).

Organization of the Dissertation

Chapter 1 included a brief overview of the background to problem, the purpose and corresponding research questions to be addressed, the significance and delimitations of this study, and the operational definitions of terms. Chapter 2 begins with the role of testing in decision-making process, followed by the context in which DA and DC are needed, and a description of existing approaches to estimating DA and DC based on CTST and IRT. The chapter concludes with a summary of the literature review and how this connects to the research questions. Chapter 3 presents the procedures to be used in this study for estimating the correctness of decision consistency and decision accuracy. Chapters 4 and 5 contain the results for the population and samples and a discussion of the results for the two data sources used in this study. Chapter 6 presents the summary of the study and the findings for each data source, discussion across the two findings across the two data sources, the conclusions drawn on the basis of the overall findings, implications for practice, and recommendations for future research.

CHAPTER 2: LITERATURE REVIEW

This chapter presents a review of previous research for estimating Decision Accuracy (DA) and Decision Consistency (DC). Several methods have been developed to indicate how accurate and consistent educational decisions based on test scores are. These methods were developed using either Classical Test Score Theory (CTST) or Item Response Theory (IRT). The chapter is organized in three sections. The context in which DA and DC are needed is provided first. This is followed by a description of existing approaches to estimating DA and DC based on CTST and IRT. A summary of the literature and how the findings from the literature connect to the research questions presented in Chapter 1 concludes the chapter.

The Role of Educational Testing in Decision-making Process

Two of the purposes of large-scale testing in education are to produce reliable scores that can be validly interpreted and to make decisions about placement, remediation, and certification using these scores. It is important and fair that the decisions made are both accurate and consistent (Mislevy, 1991; Messick, 1975; Rulon, 1939; Wainer & Kiely, 1987). Accurate and consistent decisions depend upon the clarity of the descriptions of the different ordered performance levels or categories, administration of tests that yield reliable scores that can be validly interpreted in terms of the domain represented by the performance levels, and sound procedures for setting cut-scores in the test score distribution that distinguish the performance levels (Klein & Orlando, 2000; Nevitt, 1998; Skakun & Kling, 1980). For example, given these conditions are met, we could report with confidence that a student performed well in a range of activities related to numeracy but less successfully with a range of activities related to geometry. This information is useful because teachers can concentrate on those topics where a student is weak, thereby leading to improvement in the student's performance (Bradlow & Wainer, 1998; Brennan, 2001; Hagen, 1983).

Figure 1 provides a graphical representation of the steps of the decision-making process followed to determine in which of three performance levels – below basic, basic and above basic – a student should be placed based on what the student knows as demonstrated on a test.

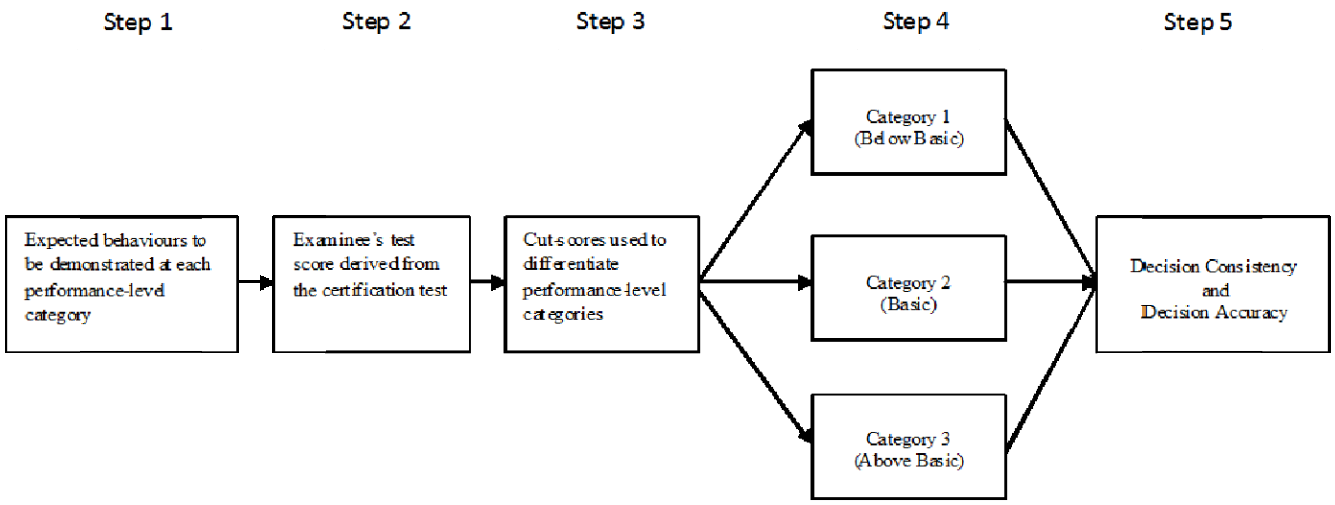
Step 1
Step 2
Step 3
Step 4
Step 5
Expected behaviours to be demonstrated at each performance-level category
Examinee's test score derived from the certification test
Cut-scores used to differentiate performance-level categories
Category 1 (Below Basic)
Category 2 (Basic)
Category 3 (Above Basic)
Decision Consistency and Decision Accuracy

Figure 1: Decision Making Process in Education

Step 1: Experts identify the expected behaviours that should be portrayed by examinees whose scores will be used to classify the examinees at each of the three performance levels. These behaviours depict the level of subject matter content and processes that an examinee should have mastered in order to be classified in a particular performance category (Howard, 1964; Skakun & Kling, 1980; Subkoviak, 1978).

Step 2: Subject matter experts construct the test with items that are relevant to and representative of the expected behaviours for each performance-level category. Thereafter, proper procedures are followed during test administration and scoring to obtain reliable test scores that can be validly interpreted for each student who sat for this test (Chester, 2003; Cicchetti, & Feinstein, 1990; English Language Institute, University of Michigan, 2006). Correct decisions about students' abilities based on the performance during examinations are enhanced if the tests used are properly standardized (Hoover, Hieronymus, Frisbie, & Dunbar, 1996a, 1996b; Keats & Lord, 1962; Keller, Swaminathan, & Sireci, 2003). Standardization of a test ensures that all the students are administered the same test referenced to and representative of the performance levels, the instructions and timing are the same, and scoring rubrics for the open-response items are the same and are applied in the same way. The intent is to ensure that the testing process does not adversely influence a student's performance (Brown, 1980; Hagen, 1983; Nunnally, 1978; Wainer, Bradlow & Wang, 2007).

Step 3: Cut-scores are set that separate the students into the performance levels. Examinees' are classified into different performance levels by using cut-scores. Cut-scores, which are usually set on a continuous score scale, differentiate the examinees' performances on the test according to the prescribed performance standard at two adjacent levels (Harwell & Janosky, 1991; Heubert & Hauser, 1999; Lee, Brennan & Kolen, 2000). Consequently, when setting these cut-scores care must be taken to ensure that the cut-scores are reasonable in order to make accurate and consistent decisions. For example, examinees who have sufficient knowledge about the subject matter content and who score above the cut-score that differentiates basic from above basic performance should be able to answer most of the items on the test, whereas examinees whose scores fall below the cut-score should be able to correctly answer fewer of the items (Rogers & Ricker, 2006; Wainer, Wang, Skorupski & Bradlow, 2005).

Step 4: The cut-scores developed at Step 3 are used to place students in one of the three performance levels. If a student's test score is less than the first cut-score that separates the below basic and basic performance levels, then the student is placed in the below basic category (Fitzmaurice, 2002; Gong & Hill, 2001; Uebersax, 2003). A student with a score that is equal to or greater than the first cut-score and below the second cut-score would be placed in the basic category. Lastly, a student with a score equal to or greater than the second cut-score would be placed in the above basic category (Spray & Welch, 1990; Swaminathan, Hambleton & Algina, 1975).

Step 5: How good the decisions made at Step 4 are assessed at Step 5 by estimating decision accuracy and decision consistency.

The purpose of this dissertation, which corresponds to Step 5 in Figure 1, was to compare three estimation procedures for determining DA and DC in terms of bias and consistency for the three estimators. The fact that there is more than one estimation procedure speaks to the difficulty in determining DA and DC. For example, while it is obvious that two testing occasions are needed, it is generally not possible to test students on two different forms that are interchangeable because of difficulties in constructing the forms and examinees not being present for both testing occasions (Box & Draper, 1987; Kolen & Brennan, 2004; Liu, Bolt & Fu, 2006).

Decision Consistency and Decision Accuracy for Two or More Test Administrations

Decision Consistency

Decision consistency is discussed first because historically consistency was considered before accuracy. In estimating DC, the observed or transformed scores on the two administrations of the same test form are not errorless. For instance, some examinees may achieve higher scores on the first administration of the test than on the second administration of the same test form while other examinees may achieve equal scores on both occasions of test administrations; and still other examinees their scores on the first administration may be lower than on the second administration of the same test form (EQAO Report, 2011; Lee, Brennan & Kolen, 2000; Messick, 1975; Mislevy, 1991). The reason for these differences is due to measurement error. However, some of the differences may be very large and lead to inconsistent classifications of examinees into different performance-level categories (Kolen & Brennan, 2004; Wainer & Kiely, 1987). The problem is that we never know the size of the error of measurement for each person on each occasion of test administration, so we never know if the observed score on the first administration of the test is:

a. lower than the score on the second administration of the same test form, which would mean that examinees may be placed in the lower category with the first administration of the test but in higher category with the second administration of the same test form,
b. the same or close to the score on the second administration of the same test form, in which case their placement would be consistent, or
c. higher than the score on the second administration of the same test form, which would mean that examinees may be placed in the higher category using the first administration of the test but in a lower category when using the second administration of the same test form.

Carver (1970) was probably the first psychometrician to propose a procedure for estimating DC. He based his procedure on the proportion of examinees that were consistently placed in one of two categories: failed to meet the standard (p_{00}) and met the standard (p_{11}) (see Table 1). The sum of these two proportions, p_0 , is the total proportion of examinees consistently categorized on two interchangeable test forms or on two administrations of the same test:

$$p_0 = p_{00} + p_{11} .$$

If there is no measurement error, then identical decisions would be made and $p_0 = 1$. However, given the fallibility of measurement, measurement error will be present and $p_0 < 1$.

Table 1

Demonstration of Decision Consistency

		Observed	Scores 2	
		0	1	Row Margins
Observed	0	p_{00}	p_{01}	$p_{1.}$
Scores 1	1	p_{10}	p_{11}	$p_{2.}$
	Column Margins	$p_{.1}$	$p_{.2}$	1.00

Cohen (1960) indicated that p_0 was also influenced by chance. He therefore proposed that p_0 be corrected for chance. The formula for the probability of a correct decision by chance is:

$$p_c = p_{.1}p_{1.} + p_{.2}p_{2}.$$

where p_c is the probability of correct decisions by chance,

$p_{1\bullet}$ and $p_{\bullet 1}$ are the row and column percentages of the examinees classified in the lower proficiency category, and

$p_{2\bullet}$ and $p_{\bullet 2}$ are the row and column percentages of examinees classified in the higher proficiency categories on two interchangeable test forms (see Table 1).

The formula for coefficient kappa, κ, the corrected proportion of correct decisions for a 2 x 2 table, is given by:

$$k = \frac{p_0 - p_c}{1 - p_c},$$

where $p_0 - p_c$ is the actual gain over chance, and

$1 - p_c$ is the theoretical gain over time.

Hambleton et al. (1973) and Swaminathan et al. (1974) extended the formulas for p_0 and κ for use in situations where there were more than two performance categories. The general formula for p_0 is:

$$p_0 = \sum_{i=1}^{k} p_{ii},$$

where p_{ii} is the percentage of examinees consistently classified to the ith proficiency categories on two occasions using two interchangeable test forms, and

$k \geq 2$ is the number of proficiency-level categories.

The general formulas for p_c and κ are, respectively:

$$p_c = \sum_{i=1}^{k} p_{i\bullet} p_{\bullet i},$$

and

$$\kappa = \frac{p_0 - p_c}{1 - p_c},$$

where $p_{i\bullet}$ and $p_{\bullet i}$ are the percentage of individuals classified in the i^{th} proficiency level categories on two interchangeable test forms.

Decision Accuracy

Although test scores are not errorless due to measurement error, the corresponding true scores are errorless. Consequently, classification decisions that are made using true scores are the accurate classifications. On the contrary, classification decisions that are made using test scores are not accurate because they contain measurement error (Keats & Lord, 1962; Keller, Swaminathan & Sireci, 2003; Klein & Orlando, 2000). Sometimes, the examinees scores may be too low leading to negative measurement error, other examinees scores may be just right leading to zero measurement error and still other examinees scores may be too high and this gives rise to positive measurements error (Kolen & Brennan, 1995; Shepard, 1980; Wainer, H., Bradlow & Du, 2000). The problem is that we never know the error of measurement, so we never know if the test score is:

a. too low, which would mean that examinees may be placed in the lower category when their true score indicates they should be in the next higher category,

b. just right, in which case their placement would be valid or truthful, or

c. too high, which would mean that examinees may be placed in the higher category when their true score indicates they should be in the next lower category.

Table 2

Demonstration of Decision Accuracy

		Observed	Scores	
		0	1	Row Margins
True	0	p_{00}	p_{01}	$p_{1.}$
Scores	1	p_{10}	p_{11}	$p_{2.}$
	Column Margins	$p_{.1}$	$p_{.2}$	1.00

The inconsistent classifications, p_{01} and p_{10} (see Table 2), are likely due to the presence of error of measurement in the observed score. In Table 2, the DA is:

$$p_0 = p_{00} + p_{11},$$

where p_{00} is the proportion of students classified in the lower category using both the true score and the observed score, and

p_{11} is the proportion of students classified in the upper category using both the true score and the observed score.

The value of p_0 should be close to 1.00.

As mentioned earlier, procedures for estimating DC were developed before procedures for estimating DA. As result, the procedures developed during the 1970s and 1980s only estimated DC (e.g., Hambleton & Novick, 1973; Swaminathan, et al., 1974; Huyhn, 1976; Marshall & Haertel, 1975; Subkoviak, 1976, 1978; Peng & Subkoviak, 1980). Beginning in the early 1990s, both DC and DA were considered. The first procedure for estimating both DC and DA for dichotomously scored items was put forward by Hanson and Brennan (1990). Subsequently, Livingston and Lewis (1995) extended the Hanson and Brennan (1990) procedure for estimating DC and DA for tests containing both dichotomously scored items and polytomously scored items. To accommodate polytomously scored items they used what they called the effective test length. However, this procedure appeared to be complex in terms of mathematical computation. As a result, Lee (2005) proposed the compound multinomial procedure as an extension of Livingston and Lewis procedure that avoided the need to use a test's effective length. Lee proposed two models. The multinomial model can be used for estimating DC and DA for dichotomously scored items only and for polytomously scored items with the same score points across all items; and the compound multinomial model can be used with items with varying score points. These two developmental studies were conducted in the framework of CTST. Similarly, there were developmental studies that were carried out in the framework of IRT. For example, Huynh (1990) developed the first procedure to estimate DA and DC in the framework of IRT for tests that contained only dichotomously scored items. Schulz, Kolen, and Nice-wander (1999) put forward another procedure using dichotomously scored items in estimating DC and DA based on IRT framework. Then Wang, Kolen, and Harris (2000) extended their IRT procedure to include polytomously scored items when estimating both DC and DA. Afterwards, numerous developmental studies were completed for complex assessment in the framework of IRT to estimate both DA and DC (e.g., Lee, 2010; Lee, Hanson, & Brennan, 2002; Martineau, 2007; Rudner, 2005; Wan, 2006; Zhang, 2008).

Given that both DC and DA were considered in the present study, the following section is restricted to three procedures that estimate both consistency and accuracy: the Livingston-Lewis (LL procedure) (Livingston, & Lewis, 1995), Compound Multinomial (CM Procedure) (Lee, 2005), and IRT procedure (Lee, 2008). As mentioned above, the LL procedure and the CM procedure are based on CTST whereas Lee's procedure is based on IRT. These three procedures are the procedures that were compared in the present dissertation.

Livingston-Lewis Procedure (LL procedure)

The Livingston and Lewis (1995) procedure is based on CTST and accommodates the use of both dichotomously and polytomously scored items. Livingston and Lewis utilized the four-parameter beta distribution to estimate the examinees' true scores and the binomial distribution to estimate the conditional error of measurements. To accommodate the inclusion of polytomously scored items, they proposed what they called the effective test length. The effective test length is determined by the "number of discrete, dichotomously scored, locally independent, equally difficult test items necessary to produce total scores having the same precision as the scores being used to classify the test takers" (Livingston

& Lewis, 1995, p. 180). The steps that are followed to estimate DC and DA are delineated below:

Step 1: Determine the effective test length (H)

Estimation of the effective test length for each test is done by employing the reliability of the total scores derived from all of the items that comprise the test form. The formula for the effective test length is:

$$H = \frac{(\mu_\upsilon - U_{lowest})(U_{highest} - \mu_\upsilon) - r_{\upsilon\upsilon}\sigma_\upsilon^2}{\sigma_\upsilon^2(1 - r_{\upsilon\upsilon})} \text{ ,}$$

where H denotes the effective test length to nearest whole number,

μ_υ represents the mean of the total scores,

σ_υ^2 stands for the variance of the total scores,

$r_{\upsilon\upsilon}$ represents the test score reliability of the test, and

U_{lowest} and $U_{highest}$ are the least possible score and uppermost possible score derived from examinees' responses from the initial test form, respectively (Livingston & Lewis, 1995, p. 182).

Step 2: Adjust the observed score

Using the effective test length *H* computed in Step 1, the observed score scale is adjusted onto a new score scale ranging from 0 to *H.* The adjusted test score is:

$$L = H\frac{U - U_{lowest}}{U_{highet} - U_{lowest}} = Hp,$$

where L represents the adjusted test score to nearest whole number,

U stands for the examinee's score,

U_{lowest} and U_{higest} stand for the least possible score and uppermost possible score derived from examinees' responses from the test form respectively, and

p denotes the proportional total score for a particular examinee on the range from 0 to 1 scale.

Step 3: Determine the proportional true score (τ_p) distribution

Lord's (1965) strong true score theory is employed to determine the proportional true score distribution that is derived from the adjusted raw scores. It is assumed that the proportional true score distribution follows a four-parameter beta distribution with probability density function given by:

$$g(\tau_p / \alpha, \beta, k, l) = \left(\frac{1}{B(\alpha+1, \beta+1)} \right) \left(\frac{(\tau_p - k)^{\alpha} (l - \tau_p)^{\beta}}{(l-k)^{\alpha+\beta+1}} \right),$$

where *B* stands for beta function. The process of deriving the above function involves the use of a two-parameter beta distribution, having parameters ($\alpha+1$) and ($\beta+1$) on the scale (0, 1) and then the scores are adjusted linearly onto the interval (*k*, *l*), where 0 $\leq k<l<$ 1. The two parameters *k* and *l* are added to the function in order to make the function conformable for the computation technique by permitting zero frequencies for the lowest and highest true-scores (see Hanson & Brennan, 1990). The proportional true score equivalent to an observed score *U* on a range from 0 to 1 scale is given by:

$$\tau_p = \frac{\xi_r(U) - U_{lowest}}{U_{highest} - U_{lowest}},$$

where τ_p is the proportional true score,

$\xi_r(U)$ denotes the expected value of an examinee's observed score for transposable test forms, and

U_{lowest} and $U_{highest}$ stand for the least possible score and uppermost possible score derived from examinees' responses from the test form respectively (Livingston & Lewis, 1995, p. 182).

Step 4: Compute decision accuracy

The estimation of the agreement between true classifications and observed classifications leads to the computation of DA. This involves the use of the proportional true score distribution, that was estimated in step 3. The distribution of the hypothetical test form scores with *H* independent dichotomously-scored items conditioned on true scores for the examinees is employed to generate the binomial distribution for each performance-level category.

In order to estimate decision accuracy, it is assumed that true cut-scores are identical to observed cut-scores. Using these cut-scores, the joint distribution of the classifications derived from the true scores and the original test form scores are as shown in Table 2 and:

$$p_0 = p_{00} + p_{11}.$$

Step 5: Compute decision consistency.

Estimation of DC involves the computation of the probability of consistently classifying an examinee above prescribed cut-score or below the prescribed cut-score on both the original test form and the hypothetical test form. Using the same cut-score for both tests, the probability of consistently classifying a student based on both tests is derived from original test form and the hypothetical test form as shown in Table 1. As before, the DC is given by:

$$p_0 = p_{00} + p_{11}$$

The decision consistency due to chance is:

$$p_c = p_{0.}p_{.0} + p_{1.}p_{.1}$$

which is then used to compute coefficient kappa κ:

$$k = \frac{p_0 - p_c}{1 - p_c} \ .$$

Compound Multinomial Procedure (CM Procedure)

While still based on CTST but in contrast to the Livingston and Lewis (1995) procedure, the compound multinomial procedure employs two models depending on the nature of the items included in a test. The multinomial model is employed if the test consists of only dichotomously-scored items or only open-ended items that have the same number of total marks for each item. The compound multinomial model is used with a test in which number of marks varies across items (Lee, 2005).

Multinomial Model

Let *s* be the number of items, where each item is worth *T* possible item scores, a_1, a_2..........a_T, and $T \geq 2$ .It is assumed that:

1. the *s* items are randomly drawn from a pool of items referenced to the same domain:

2. $\vec{\omega}=\{\omega_1,\omega_2,.........\omega_T\}$ represents the proportion of items in the pool of items referenced to the same domain that an examinee can obtain marks of *a₁*, *a₂*..........*a_T*, respectively, and

3. $\sum_{g=1}^{T}\omega_g=1.$

Let *R₁* *R₂*...........*R_T* be random variables corresponding to the numbers of items in the test that examinees receive a score of *a₁*, *a₂*..........*a_T*, where $\sum_{g=1}^{T}R_g=s$ for each examinee. Then, the multinomial distribution for these randomly selected items is given by:

$$\Pr(R_1=r_1,R_2=r_2,......R_T=r_T\,/\,\vec{\omega})=\frac{s!}{r_1!r_2!..........r_T!}\omega_1^{r_1}\omega_2^{r_2},...........\omega_T^{r_T}\,.$$

The total likely combination of all possible marks for each item that examinees can get and that leads to the same aggregate mark can be expressed using the probability density function:

$$\Pr\left(Q=q/\vec{\omega}\right)=\sum_{a_1r_1+a_2r_2+.........+a_tr_t=q}\Pr(R_1=r_1,R_2=r_2,......R_T=r_T\,/\,\vec{\omega}),$$

where q is the aggregate mark obtained through different combinations of all possible marks $r_1,r_2,........,r_T$ such that $\sum_{g=1}^{T}a_gR_g=q$.

Classification and Performance-level Categories

Let the number of independent performance-level categories be *H*. This implies that there are *H*-1 cut-scores: $\lambda_1, \lambda_2, ...\lambda_{H-1}$. Lee (2005) introduced two additional "cut-scores" in order to be able to establish intervals of the form $(\lambda_{h-1} \leq H < \lambda_h)$. The two new cut-score are $\lambda_0 = \min(H)$ and $\lambda_H = \max(H)$. Let g_h represents the performance-level category in which an examinee is classified on each on two interchangeable test forms or on two occasions when the same test form is administered and *h* =1, 2, ...*H*.

In the case of a single administration, a set of scores on a hypothetical parallel form is created using the bootstrap procedure (Efron, 1982; Brennan & Wan, 2004). The bootstrap can be applied at the item level to create a full form. Alternatively, the bootstrap can be applied at the test level. In the present study, the bootstrap was applied at the test level to obtain a random set of bootstrap scores where the number of examinees equalled the number in the actual sample (Brennan, Harris, & Hanson, 1987).

Decision Consistency

The probability of a randomly selected examinee *l* with total score *Q* will be consistently placed in performance category, g_t, using the test and the hypothetical test is given by:

$$\begin{aligned}\Pr(Q_1 \in g_t, Q_2 \in g_t / \vec{\omega}) &= \Pr(Q_1 \in g_t / \vec{\omega}) \Pr(Q_2 \in g_t / \vec{\omega}) \\ &= [P(Q_1 \in g_t / \vec{\omega})]^2\end{aligned}$$

The decision consistency for examinee *l* is the probability that the examinee *l*'s scores will be placed in the same performance category, g_t, is given by:

$$\chi_l(\theta) = \sum_{t=1}^{T} \left[\Pr(Q \in g_t / \vec{\omega})\right]^2$$

Taken across the *N* examinees, the decision consistency is given by

$$p_0 = \frac{1}{N} \sum_{l=1}^{N} \sum_{t=1}^{T} \left[\Pr(Q \in g_t / \vec{\omega})\right]^2 .$$

Kappa, *k*, the correction for chance, is given by:

$$k = \frac{p_0 - p_c}{1 - p_c},$$

where p_c is the probability due to chance:

$$p_c = \sum_{t=1}^{T} \left[\frac{1}{N} \sum_{l=1}^{N} \Pr(Q \in g_t / \vec{\omega})\right]^2 .$$

Decision Accuracy

In order to estimate decision accuracy index, as with Livingston and Lewis (1995), Lee assumed that true cut-scores are identical to observed cut-scores. Using these cut-scores, examinee *l*'s true status level is given by:

$$\Gamma_l = \Pr(Q \in g_t / \vec{\omega}),$$

where examinee *l* is classified in the performance category g_t. Across the *N* examinees in a group, the decision accuracy is given by:

$$\Gamma = \frac{1}{N}\sum_{l=1}^{N} \Pr(Q \in g_t / \vec{\omega}).$$

Hanson et al. (1990) used the false positive and false negative errors to determine decision accuracy. The probability of a false positive error is given by:

$$\Gamma_l^+ = \sum_{t=\hat{t}+1}^{T} \Pr(Q \in g_t / \vec{\omega}),$$

and the probability of a false negative error is given by:

$$\Gamma_l^- = \sum_{t=1}^{\hat{t}-1} \Pr(Q \in g_t / \vec{\omega}),$$

where $\hat{t}$ represents the examinee's true category. The false positive error rates and false negative error rates for the *N* examinees who sat for this particular examination are, respectively:

$$\Gamma^+ = \frac{1}{N}\sum_{l=1}^{N}\sum_{t=\hat{t}+1}^{T} \Pr(Q \in g_t / \vec{\omega}),$$

and

$$\Gamma^{-} = \frac{1}{N}\sum_{l=1}^{N}\sum_{t=1}^{\hat{t}-1}\Pr(Q \in g_t / \vec{\omega}).$$

Compound Multinomial Model

As indicated above, the compound multinomial model allows mixed item formats. Consequently, the probability formulas for the multinomial model need to be adjusted to account for the different item formats. This is done by dividing the *s* items in the total test into *U* subtests, where the items in each subtest have the same number of score points and then working at the subtest level. However, the procedures for classification of an examinee into categories and determining decision consistency and accuracy are the same. Therefore, only the adjusted procedures are provided here. This is then followed by the presentation of a procedure for correcting for bias given the finding that the results yielded by the compound multinomial model are biased.

Again, let *s* represent the number of items in the test with *f* test item subsets. But now let *U* represent the number of subsets of items with the same number of score points a_{T_f}; and there are $s_f \ (f = 1, 2, ..., F)$ test items in each subset such that $\sum_{u=1}^{U} s_f = s$. Lastly, let Q_f represent the aggregate mark for each examinee for the *f*th subset. The compound multinomial model assumes that correlations among the errors of measurement of the *U* different subsets total scores are zero. With this assumption and letting $\vec{\omega}_f = \{\omega_{f1}, \omega_{f2}.........\omega_{fT_f}\}$ represent the joint probability density function for each examinee of Q_f s for U item subsets is given by:

$$\Pr(Q_1 = q_1,..........., Q_U = q_U / \vec{\omega}_1,......, \vec{\omega}_U) = \prod_{f=1}^{U} \Pr(Q_f = q_f / \vec{\omega}_f).$$

$\Pr(Q_f = q_f / \vec{\omega}_f)$ is computed in the same way as in multinomial model. However, the probability distribution for the total test score for each examinee $D = \sum_{f=1}^{U} v_f Q_f$ is given by:

$$\Pr(D = d / \vec{\omega}_1,....., \vec{\omega}_U) = \sum_{q_1,......q_U : \sum v_f q_f = d} \Pr(Q_1 = q_1,..........., Q_U = q_U / \vec{\omega}_1,......, \vec{\omega}_U),$$

where $q_1,......,q_U : \sum v_f q_f = d$ displays the aggregate score for all possible subsets total scores of $q_1,........,q_U$ such that the combined weighted aggregate scores is equal to the aggregate score *d*. When the correspondence between scale marks and raw marks is not one to one, then the probability density function for scale marks is given by:

$$\Pr(SC = sc / \vec{\omega}) = \sum_{q : u(q) = sc} \Pr(Q = q / \vec{\omega}),$$

where *u(q)* denotes the function used to adjust the raw marks to scale marks, and *q:u(q) = sc* is equal to the aggregate for all *q* marks that are adjusted to one scale mark *sc.*

As indicated above, the same procedure used for raw marks in estimating decision consistency and decision accuracy is also employed for scale marks.

Correction for Bias

A problem with compound multinomial procedure is that the estimates of decision accuracy and decision consistency are biased. As a result, Brennan and Lee (2006) and Wan, Brennan, and Lee (2007) proposed a bias-correction procedure for the compound multinomial procedure. The bias-correction procedure is based on the notion that the true score variance is less than the observed scored variance since the observed score is equal to true score plus error score according to classical test score theory. However, the true score variance is greater than the regressed true score variance (Kelley, 1947). Brennan and Lee's (2006) procedure for bias-correction uses weights that give the maximum estimates for decision indices. The combination of weights for raw scores is given by:

$$\frac{\sqrt{\rho^2 xx}}{1+\sqrt{\rho^2 xx}},$$

whereas the combination of weights for the regressed score is given by:

$$\frac{1}{1+\sqrt{\rho^2 xx}},$$

where $\sqrt{\rho^2 xx}$ is the index of reliability of test X.

For dichotomously scored items, the combined weighted true proportion-correct score that provides the maximum value for an examinee is given by:

$$\hat{\pi} = \frac{\overline{X}}{N} + \sqrt{\hat{\rho}^2 xx}(\frac{x}{n} - \frac{\overline{X}}{n}),$$

where $\overline{X}$ is the mean score across examinees,

n is the number of items, and

x is the examinee's correct score.

For polytomously scored items, the combined weighted true proportion-correct score that provides the maximum possible mark for each examinee is:

$$\hat{\pi}_g = \frac{\overline{X_g}}{N} + \sqrt{\hat{\rho}^2 xx}(\frac{x_g}{n} - \frac{\overline{X_g}}{n}),$$

where x_h is the observed number of items in the subset marked with a mark option *g* for the examinee, and

$\overline{X_g}$ is the average number of items in the subset marked with the mark option *h* for all examinees who sat for this test (Wan, et al., 2007, p.18).

While applying these formulas may reduce the bias, they do not totally eliminate the bias (Wan et al., 2007, p. 22). Despite this, the decision consistency and decision accuracy indices achieved by using the compound multinomial procedure are somewhat superior compared to indices attained by employing Livingston and Lewis (Wan et al., 2007 p. 23).

IRT Procedure

Lee (2008) developed a method based on the IRT for estimating DC and DA using a single-administration of the test form. The procedure may be used with the 1-, 2-, and 3-parameter IRT models. The procedure can make use of any score metric in which the cut-scores are expressed. For the purposes of this dissertation, the 3-parameter model was employed. Calibration of the IRT parameters included in the 3-parameter IRT model and derived from examinees' item responses is the first stage in the process of estimating DC using Lee's (2008) IRT procedure. The 3-parameter logistic IRT model (Hambleton, Swaminathan, & Rogers, 1991) is given by:

$$\mathrm{P}_i(\theta) = \mathrm{c}_i + (1 - \mathrm{c}_i)\frac{\mathrm{e}^{\mathrm{Da}_i(\theta - \mathrm{b}_i)}}{1 + \mathrm{e}^{\mathrm{Da}_i(\theta - \mathrm{b}_i)}},$$

where $\mathrm{P}_i(\theta)$ is the probability of a "correct" response for an individual at a given level of theta,

b_i is the item difficulty or location parameter,

a_i is the item discrimination parameter,

D is a scaling constant equal to 1.702,

c_i is the pseudo-guessing parameter. .

Estimating the examinees total raw score Q that is conditioned on ability θ using Lord and Wingersky (1984) algorithm is the second stage. Given the assumption of local independence has been met, the probability of an examinee getting a score of r on item i given ability θ is $P(R_i = r_i / \theta)$. The conditional probability for the total observed score Q is given by:

$$P(Q = q / \theta) = \sum_{q=\sum r_i} P(R_1 = r_1 / \theta) P(R_2 = r_2 / \theta) \ldots\ldots\ldots P(R_t = r_t / \theta).$$

Decision Consistency

The raw score distribution conditioned on ability θ is used when estimating examinees' probabilities of belonging to one particular performance-level category is consonant with their total scores obtained on the examination. The procedure for computing DC and DA is similar to the compound multinomial procedure. It is assumed that the examinee's scores on two interchangeable test forms or when the same test form is administered on two occasions are conditionally independent and identical. Therefore, the DC for the examinee *l*'s scores consistently classified in one performance category, g_t, is given by:

$$\begin{aligned} \Pr(Q_1 \in g_t, Q_2 \in g_t / \vec{\omega}) &= \Pr(Q_1 \in g_t / \vec{\omega}) \Pr(Q_2 \in g_t / \vec{\omega}) \\ &= [P(Q_1 \in g_t / \vec{\omega})]^2 \end{aligned}$$

where Q_1 and Q_2 represent the aggregate scores obtained from two interchangeable test forms. Therefore, the DC for an examinee *l* is the probability that the examinee's scores will be placed in the same performance category, g_t, each time the measurement instrument is employed under the similar conditions. This probability is given by:

$$\chi_l(\theta) = \sum_{t=1}^{T} \left[\Pr(Q_1 \in g_t / \vec{\omega})\right]^2.$$

The overall p_0 for all the examinees *N* is given by:

$$p_0 = \int_{-\infty}^{\infty} \sum_{t=1}^{T} \left[\Pr(Q \in g_t / \vec{\omega})\right]^2 \Gamma(\theta) d\theta,$$

where $\Gamma(\theta)$ is the ability distribution.

Kappa, k, can be computed by:

$$k = \frac{p_0 - p_c}{1 - p_c},$$

where p_c is the probability index for decision consistency due to chance given by:

$$p_c = \sum_{t=1}^{T}\left[\int_{-\infty}^{\infty}\Pr(Q \ \in g_t / \vec{\omega})d\vec{\omega}\right]^2 ,$$

where $\int_{-\infty}^{\infty}\Pr(Q \ \in g_t / \vec{\omega})d\vec{\omega}$ is the marginal category probability.

For scale scores, the conditional category probability for any two randomly selected scale scores, SC_1 and SC_2 , categorized in the same category is given by:

$$\Pr(SC_1 \in g_t, SC_2 \in g_t / \omega) \qquad = \left(\sum_{sc=z_{(t-1)}}^{z_t-1} P(SC = sc / \omega)\right)^2 .$$

The overall decision consistency for scale score is given by:

$$p_0 = \int_{-\infty}^{\infty}\left[\sum_{t=1}^{T}\Pr(SC_1 \in g_t, SC_2 \in g_t / \omega)\right]\Gamma(\theta)d\theta.$$

Decision Accuracy

Lee (2008) used the expected value of an examinee’s raw score, τ_l, as the examinee’s true score. Estimation of decision accuracy requires that the score metric for observed cut-scores is congruent with the expected score metric. Therefore, the expected scores metric derived from the cut-scores metric conditioned on θ is given by:

$$E(Q/\theta=\hat{\theta})=\sum_i\sum_j j\,\Pr(R_i = j/\theta=\hat{\theta}),$$

where $\hat{\theta}$ represents a θ cut-score;

R_i is a random variable representing examinees' responses to item *i*; and

$\Pr(R_i = j/\theta=\hat{\theta})$ denotes the conditional probability for score *j* for item *i*, conditioned on the ability θ .

The cut-scores on the theta distribution are converted to the cut-scores on the observed score metric. In view of the fact that the expected scores metric is the same as the metric for the examinees' true scores τ $(1,2,........T)$, then the probability that an examinee's score will be classified in a specific performance-level category conditioned on his/her true score is computed by:

$$\gamma\theta=\psi\theta(\tau)=\Pr(Q\in g_t/\Phi=\theta), \qquad for\ \theta\,\varepsilon\,\tau\, and\, \tau=g_t.$$

Hence, the marginal classification accuracy index, γ , is:

$$\gamma=\int_{-\infty}^{\infty}\gamma\theta\Gamma(\theta)d\theta.$$

Hanson et al., (1990) indicated that false positive error rates and false negative errors are also used as a measure for decision accuracy indices. The conditional false positive error rate is given by:

$$\gamma_\theta{}^+ = \sum_{\tau=\tau^*+1}^{T}\psi\theta(\tau), \quad for\ \ \theta\ \varepsilon\ \hat{\tau},$$

where $\hat{\tau}$ is the true cut-score associated with θ.

The conditional false negative error rate is given by:

$$\gamma_{\theta}^{-} = \sum_{\tau=1}^{\tau^{*}-1} \psi\theta(\tau), \quad for \;\; \theta \; \varepsilon \; \hat{\tau}.$$

Hence, the marginal false positive error rate, γ^{+}, is given by:

$$\gamma^{+} = \int_{-\infty}^{\infty} \gamma_{\theta}^{+} \, \Gamma(\theta) d\theta,$$

and the marginal false negative error rate, γ^{-}, is given by:

$$\gamma^{-} = \int_{-\infty}^{\infty} \gamma_{\theta}^{-} \, \Gamma(\theta) d\theta,$$

where $\Gamma(\theta)$ denotes the ability distribution. The estimation of these marginal false positive and false negative error rates is done by either using an individual distribution which involves using theta estimates for each personal, or a distribution approach which involves the use of theta quadrature points and weights.

Previous studies on DC and DA

The review of research studies that compared different procedures for determining DC and DA is focused on studies that compared the three procedures considered in the present study. Hanson and Brennan (1990) compared the correctness of DC and DA for three different beta distribution models used in the estimation of DC and DA in the procedure they developed: the two-parameter beta binomial model, the four-parameter beta binomial model, and four-parameter beta compound binomial model. Their study included two samples: 230,033 examinees who took the October 1987 ACT English, Mathematics, Social Studies and Natural Science Assessments and 151,050 examinees who took the same four tests for February 1988. The values of DC obtained from the three procedures were similar, ranging from 0.86 to 0.90. Generally the values for DC were the lowest for the two-parameter beta model, followed, in turn, by the four-parameter beta binomial model and the four-parameter beta compound binomial model. The DA values ranged from 0.88 to 0.95. Again, the values of DA were lowest for the two-parameter beta model, followed, in turn, by the four-parameter beta binomial model and the four-parameter beta compound binomial model (p.353).

Lee, Hanson and Brennan (2002) conducted a follow up study to compare the correctness of the DC and DA indices using three different distribution models used in the estimation process: the two-parameter beta binomial model; four-parameter beta binomial model, and 3-parameter logistic IRT model. Their study included 3,000 examinees for Form X of the ACT Applied Mathematics administered in fall 1997 and 19,158 examinees for Form Y of ACT Applied Mathematics administered in 1997. The decision consistency values ranged from

0.73 to 1.0. Generally, the values of DC were lowest for the two-parameter beta model, followed, in turn, by the four-parameter beta binomial model and the 3-parameter logistic model. A similar pattern was observed with DA which ranged from 0.81 to 1.0 with the lowest values for the two-parameter beta model, followed, in turn, by the four-parameter beta binomial model and the 3- parameter logistic model (p.424).

Wan (2006) compared the estimates of DC and DA using the normal approximation (NM), Breyer-Lewis (BL), Livingston-Lewis (LL), Bootstrap (BW), and compound multinomial models (CM). The sample for his study included the 650 examinees who sat for Multistate Bar Examinations developed by the National Conference of Bar Examiners. The estimates for DC were quite close, ranging from 0.83 to 0.86, with the LL providing the lowest estimates followed in turn by NM and BL. The CM and BW provided the highest estimates. The range of estimates for DA was greater, ranging from 0.70 to 0.96. Again, the LL method provided the lowest estimate of DA, followed in turn by NM and BL. The CM and BW provided the highest estimates for DA (p. 98).

Knupp (2009) compared the values of DC and DA using three procedures: normal approximation (NM), compound multinomial model (CM), and 3-parameter logistic IRT model. The sample included 3,000 grade eight examinees who sat for Iowa Tests of Basic Skills (ITBS), Form K, Level 14 1993 (p.46). The DC values for the ITBS ranged from 0.54 to 0.97. The findings revealed that the estimates of DC obtained from NM were smallest, followed by the estimates obtained from the CM which were less than the estimates from the IRT model. The estimates for DA ranged from 0.64 to 0.98. As with DC, the values of DA obtained from NM were smallest, followed by the estimates obtained from the CM, which were less than the estimates from the 3-parameter IRT model (p. 6). These findings are somewhat similar with the previous study (Lee, Hanson, & Brennan, 2002) in which IRT model provided the highest estimates for both DC and DA.

In contrast to the previous studies in which different distribution models were compared, Lee (2010) compared the correctness of the decisions using the Livingston-Lewis procedure (LL), compound multinomial procedure (CM), and six different IRT models: one-parameter logistic (1PL), two-parameter logistic (2PL), three-parameter logistic (3PL) models, partial credit (PC) model (Masters, & Wright, 1997), the generalized partial credit (GPC) model (Muraki, 1997), and graded response (GR) model (Samejima, 1997). The combinations of the dichotomous and polytomous IRT models Lee considered were: 1PL+PC, 2PL+GPC, 3PL+GPC, 1PL+GR, 2PL+ GR, and 3PL+GR. Two samples were considered. The first included 500 grade 7 examinees who sat for Iowa Tests of Basic Skills Mathematics Test. The second sample included 4,000 grade 10 examinees who sat for Science Achievement Test administered by a state government. The Mathematics test consisted of 35 multiple-choice items with 5 options scored with a two-point scale (0 and 1), and 18 open-ended items scored with a three-point scale (0, 1, and 2). For classification purposes, a cut-score was set arbitrarily to aggregate score of 38. The Science test consisted of 40 dichotomously scored multiple-choice items and 7 open-ended items scored on a four-point scale (0-3). In contrast to the Mathematics test, there were three cut-scores for the Science test: 15, 40, and 45. This provided four performance-level categories (p.6). The estimates for DC were generally higher in Mathematics, which ranged from 0.86 to 0.88, than in Science, which ranged from 0.71 to 0.75. For both tests, the estimates for DC obtained from the LL procedure were smallest, followed by the estimates obtained from the CM procedure which were less than the estimates from the IRT models, which were generally the same. A similar pattern was also found for DA. The estimates were higher for Mathematics, which ranged from 0.90 to 0.91 than for Science, which ranged from 0.79 to 0.82. Again, for both tests, the estimates for DA obtained from the

LL procedure were smallest, followed, in turn by the estimates obtained from the CM procedure and the IRT procedures.

Deficiency in the Literature

Only one study (Lee, 2010) compared the correctness of the decision indices using the Livingston-Lewis procedure (LL), compound multinomial procedure (MN), and 3-parameter IRT procedure. Lee found that the decision indices obtained using IRT procedures were generally higher than the decisions using either of the two procedures based on CTST. Lee worked with two sample sizes, a small sample size (n = 500 for the Mathematics test) and a larger sample size (n ≈ 4,000 for the Science test). Hence, it is difficult to generalize results to other samples of other sizes. The present study addressed the issue of sample size by using two different populations of students, one with approximately 127,000 examinees in Canada and the other with approximately 6,200 examinees Malawi and drawing 100 replicated random samples of different sizes from each populations. At the same time, issues of bias and precision of the estimates were addressed by comparing the mean of the 100 replications to the results obtained for the populations and computing the standard deviation of the 100 estimates.

CHAPTER 3 METHODOLOGY

The methods used to comparatively evaluate the Livingstone and Lewis, compound multinomial, and 3-parameter logistic IRT procedures for determining DA and DC are described in this chapter. First, the data sources for this study are presented and described. The factors examined are described in the second section. The replicated sampling procedure is provided in the third section. The procedures used to analyze the replicated sample data are provided in the fourth and last section.

Data Source

Two data sets were considered in this study. The first data set was provided by the Education Quality and Accountability Office (EQAO), which is an arm's-length Crown agency of the Government of Ontario, Canada. The second set was provided by the University of Malawi.

EQAO Junior Reading and Mathematics Data Sets

The EQAO is responsible for developing, administering, marking, and reporting annually standardized province-wide Reading, Writing, and Mathematics assessments at the Primary (Grade 3) and Junior (Grade 6) educational levels and the Academic and Applied Mathematics assessments at Grade 9. As well, the EQAO administers the Ontario Secondary School Literacy

Test (OSSLT), which is a high school graduation requirement administered to students in Grade 10. The data obtained for the present study from EQAO consist of the students' responses to the multiple-choice items and open response items included in the 2010 Junior Reading and Junior Mathematics Assessments. The multiple-choice items are dichotomously scored and the open-response items are polytomously scored using scoring rubrics with four scoring categories. The numbers of students who responded to these assessments was approximately 127,000.

Junior Reading

The Junior Reading assessment is designed to measure explicit and implicit information gained from reading a reading prompt and connections between the reading prompt and their own experience. The assessment, which is administered toward the end of the school year, comprises 26 multiple-choice items and 10 open-response items. The students are supposed to answer all the items. Students are expected to write this paper during a one-hour period. However, in acknowledgement of normal classroom practice, the assessments are designed to be untimed. Additional time can be provided to any student unable to complete a session in one hour. The amount of additional time per session will normally range from five to 20 minutes; however, students may take the time they need to complete the session as long as it is in one continuous sitting on the day on which the session is assigned (EQAO, 2009).

The reading booklets contain both operational and field-test items. The operational portion of the reading component contains one long reading selection (650–700 words) followed by 10 multiple-choice questions and two open-response questions and four short reading selections (300–350 words), each followed by four multiple-choice questions and two open response questions. Different sets of field test items are embedded in each operational form such that the number of forms is about 20. The field tests are embedded in the same position in each form, and the number of field test items is such that no more that 20% of the testing time is required to respond to them. Only the students' responses to the operational items are used to determine their achievement scores (EQAO, 2009).

Junior Mathematics

Five mathematical strands are assessed in the Junior Mathematics assessment: Number Sense and Numeration; Measurement; Geometry and Spatial Sense; Patterning and Algebra; and Data Management and Probability. The operational portion of the mathematics booklet contains 28 multiple-choice and eight open-response questions. As with Reading, one hour with extra time is allowed to complete the form and field-test items are embedded in the same position in approximately 20 operational forms. Likewise, the students' responses to the operational items are used to determine their scores (EQAO, 2009).

University of Malawi Data Set

The second data set was obtained from the University of Malawi. The University of Malawi administers entrance examinations in the areas of Verbal Reasoning, Numerical Reasoning, and English Language. The data from the University of Malawi consist of the applicants' responses to the multiple-choice items and the open-response items contained in these three examinations. Approximately 6,200 students sat these examinations in 2009.

Verbal Reasoning

The Verbal Reasoning examination has two parts: Part A: 30 multiple-choice items and Part B: 10 open-response items. It is a 2-hour paper. The applicants are expected to answer all the items. The items in Part A are dichotomously scored and the items in Part B are scored using a 3 point scoring rubric. It is expected that examinees would take less time for each multiple-choice item and more time for each written response item (U of M, 2009).

Numerical Reasoning

The Numerical Reasoning paper has 40 multiple-choice items in Section A and 10 open-response items in section B. The administration time for this paper is two hours. As for Verbal Reasoning, it is expected that examinees will take less time for each multiple-choice item and more time for each open-response item. Applicants are expected to answer all the items in both sections. The score points for section A is 40 points and for section B 40 points (each of the 10 open-response items was scored on scale from 0 – 4 points) (U of M, 2009).

English Language

The English Language examination has two parts: Part A: Multiple Choice and Part B: Open-ended Response items. The examination consists of 40 multiple choice items and 10 open-ended response items. The test is for 2 hours. It is expected that examinees would take less time for each multiple choice item and more time for each open-ended response item. The applicants are expected to answer all the items. The total score points for this test is 80 with 40 points for the multiple-choice items and 40 points for the open-ended response items (each response to an open-response item is scored using a five point (0-4) rubric) (U of M, 2009).

Cut-Scores

EQAO Junior Reading and Mathematics Assessments

The students who write the EQAO Junior Reading and Mathematics assessments are placed in one of five performance categories: 0, 1, 2, 3, and 4. The performance category descriptors are presented below:

Category 4: The student has demonstrated the required knowledge and skills. Achievement surpasses the provincial standard.

Category 3: The student has demonstrated most of the required knowledge and skills. Achievement is at the provincial standard.

Category 2: The student has demonstrated some of the required knowledge and skills. Achievement approaches the provincial standard.

Category 1: The student has demonstrated some of the required knowledge and skills in limited ways. Achievement falls much below the provincial standard.

Category 0: "Not enough evidence for Level 1." The student has not demonstrated enough evidence of knowledge and understanding to be assigned Level 1. (EQAO, 2009).

As well, a two category system is used – met standard (categories 3 and 4) and did not meet standard (categories 0. 1, and 2). The cut-scores corresponding to these performance categories are set on the theta score distribution created using the 3-parameter IRT model with a fixed pseudo-guessing parameter of 0.20 (EQAO, 2009) and the PARSCALE computer program (Muraki**,** & Bock, 2003). The cut-score values in ascending order are: -3.0671, -1.8559, -0.7191 and 1.0282 for Junior Reading and -3.0896, -1.5770, -0.4279 and 0.9875 for Junior Mathematics (Michael Kozlow, Personal Communication, June 8, 2011)

The applicants' scores obtained from the University of Malawi Entrance Examinations are classified into five categories: Failure, Pass, Credit, Marginal Distinction, and Undoubted Distinction. The performance category descriptors are presented below:

Undoubted Distinction: The applicant demonstrates outstanding knowledge and superior ability.

Marginal Distinction: Performance of the applicant is excellent.

Credit: Performance is considerably above the expected minimum level for an applicant

Pass: Performance is at the minimum level expected for an applicant.

Failure: Performance is unacceptably low for an applicant to be admitted into the University of Malawi.

The four cut-scores corresponding to these performance categories are set on the raw score distribution. The cut-scores associated with the data from University of Malawi for all the three papers (i.e., Verbal Reasoning, Numerical Reasoning, and English Language) are presented in percentages: 50%, 60%, 70%, and 75% (U of M, 1985).

Procedure

Decision consistency and accuracy were determined using the cut-scores identified above for the Junior Reading and Mathematics data sets from the EQAO and data set from the U of M. The values of DC and DA obtained from these two populations served as the population parameters to be recovered by each of the three procedures examined and sample sizes considered in this study.

Replicated samples were selected from each of the EQAO population and the U of M population. In the case of the EQAO Junior assessments, four sample sizes were considered: 1,500, 3,000, 4,500, and 6,000. In the case of the U of M, two samples sizes were considered: 1,500 and 3,000. The difference in the sample sizes between the EQAO and the U of M is because the population data from U of M is smaller (6,200 applicants vs. 127,000 students). For each sample size, 100 replicates were generated (Harwell, Stone, Hsu, & Kirisci, 1996; Vale, &

Maurelli, 1983). The DC and DA were estimated for each replicate for the LL, CM and IRT procedures.

Dependent Variables

The bias of DC and DA provided by the LL, CM and IRT procedures was equal to the difference between the population values of DC and DA at each cut-score from the mean of 100 replicates of DC and DA for each selected sample size for each cut-score. The following formula was used for the standard error for each cut-score/sample size condition Glass & Hopkins, 1996, p. 321):

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} \text{ ,}$$

where σ_p is the standard error of the percentage of students below the cut-score,

π is the population percentage of the students below the cut-score, and

n is the sample size.

A procedure is unbiased if the difference between the estimate and the parameter value is zero (Bose, 2001). But the estimate is subject to sampling error. Thus, the values may differ from zero when the true bias is zero. Thus a rule had to be established about how large the difference between the estimate and the parameter value could be and still claim that the estimate was unbiased. Examination of the data from the replications revealed the majority of values of the bias were close to zero. Thus, taking the ratio of the bias to its standard error led to values that did not reflect the small values of the bias, the vast majority of which were within one percent. For example, consider the results for the replicated samples for Reading presented in Tables 5 to 8 in the next chapter. Of the 144 estimates of bias across the four cut-scores and four sample sizes, the bias for 137 (95.1%) estimates was within 1% of their corresponding parameter values. Of the remaining seven, four were 1.1%, one was 1.2%, one was 1.7% and the last was 1.9%. All but one was for cut-score 2/3, and all seven were for the IRT procedure. Essentially the same results were obtained for the other four assessment considered. Of the 144 standard error estimates for the replicated samples, all but two were less than one percent, and the two that were not 1.1%. Thus, with perhaps one or two cases, the bias was essentially zero and the estimates were precise.

Software

The following programs and software packages were used: BB-CLASS (Brennan.2004) for the LL procedure, MULT-CLASS, version 3.0 (Lee, 2008) for the CM procedure, and IRT-CLASS (Lee & Kolen, 2008) for the IRT procedure. The programs that were employed to estimate item and ability parameters were the same as the programs used by the respective sources of data: PARSCALE

(Muraki & Bock, 2003) for the EQAO Junior Reading and Junior Mathematics assessments and MULTILOG (Thissen, 1991) for the University of Malawi examinations. The SSPS and SAS computer packages were used for other programming procedures to facilitate the computation of decision indices.

Chapter 4: Results and Discussion

The EQAO Assessments

The results obtained from the data analyses for the research questions presented in the Chapter 1 are provided and discussed in the present chapter for the two EQAO assessments and in the next chapter for the University of Malawi examinations. The present chapter is organized in two parts, one for the EQAO Junior Reading assessment and the other for the EQAO Junior Mathematics assessment. Each part is divided into four sections. In the first section, population parameters for the EQAO assessments are provided. The values of DC and DA obtained using the Livingston-Lewis, compound multinomial, and IRT procedures are provided in the second section, and the values for bias and standard error of DC and DA for 100 replicates associated for each sample size are provided for each cut-score for each procedure in the third section. A discussion of the full set of results for the assessment is then provided in the fourth section.

EQAO Junior Reading

Population Parameters

This dataset included 128,103 student responses for the Junior (Grade 6) Reading assessment. The population parameters for EQAO Junior Reading assessment are reported in Table 3 and the frequency distribution of ability estimates is provided in Figure 2.

The scores, mean, standard deviations, and cut-scores are expressed in terms of logits. As shown in Table 3, the population mean was -0.5695, the standard deviation was -3.3725, and the distribution was negatively skewed (γ_1 = -0.54) and slightly leptokurtic (γ_2 = 0.23) (see Figure 2). The internal consistency was 0.87, and the cut-scores progressed from -3.07 (0/1) to 1.03 (3/4).

Population DC and DA

The values for DC and DA for the population for Reading are reported in Table 4 for each cut-score and over all the cut-scores. The 2/3 cut-score is marked with an asterisk since students that score at or above the 2/3 cut-score have met the provincial standard. Whereas, the values for DC and DA over all the cut-scores yielded by the IRT procedure were the highest, followed in turn by the CM procedure and the LL procedure, there is an interaction between procedure and cut-score. As the position of the cut-score move towards the extremes of the score distributions, the differences among the values of DC and of DA yielded by the three procedures become smaller, but in different ways. The results for each cut-score are presented first followed by a discussion.

Table 3

Psychometric Properties for EQAO English Reading, N =128,103

μ_θ	σ_θ	γ_1	γ_2	α	$cs_{0/1}$	$cs_{1/2}$	$cs_{2/3}$	$cs_{3/4}$
-0.5695	-3.3725	-0.544	0.228	0.87	-3.067	-1.855	-0.719	1.0282

Note: μ_θ is the population mean; σ_θ is the standard deviation; γ_1 is the population skewness; γ_2 is the population kurtosis; α is the internal consistency (Cronbach's alpha), $cs_{0/1}$ is cut-score 0/1; $cs_{1/2}$ is cut-score 1/2; $cs_{2/3}$ is cut-score 2/3; and $cs_{3/4}$ is cut-score 3/4.

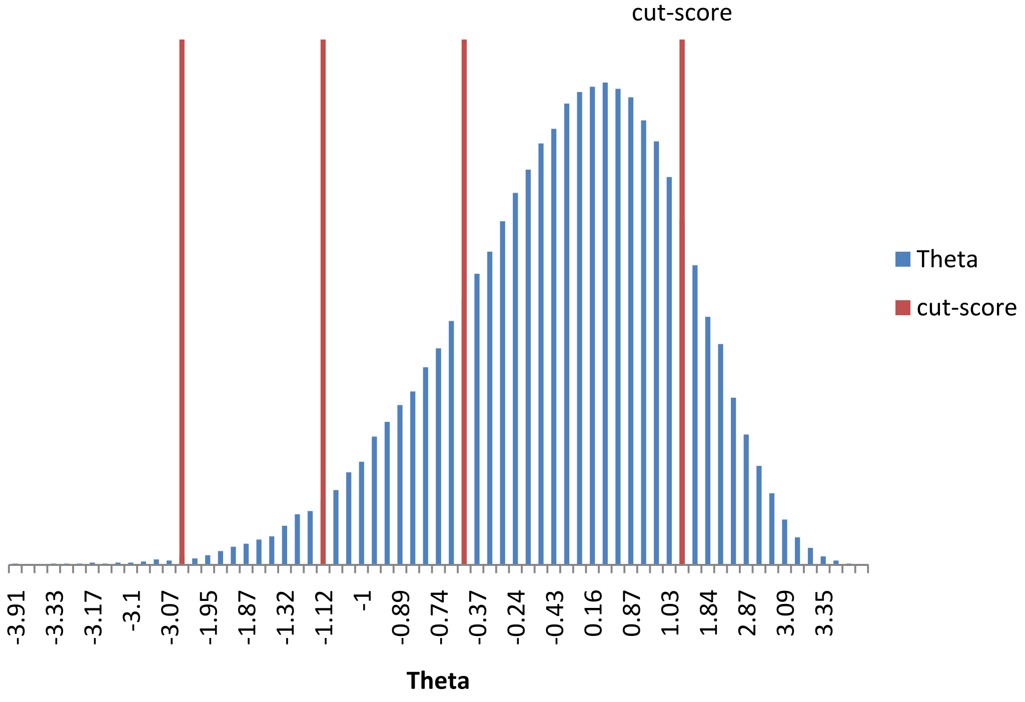

Figure 2: Frequency Distribution of Thetas for EQAO Junior Reading

0/1 cut-score

The three values of DC yielded by the three procedures are within 0.001 of each other and the values of the DA are within 0.002 of each other at the 0/1 cut-score. All values are greater than 0.998. This finding is attributable to the large difference between the 0/1 cut-score and the population mean and the small number of students around cut-score 0/1 (see Figure 2).

Table 4

Decision Consistency and Decision Accuracy Using LL, CM, and IRT Models Conditioned on Cut-Scores for EQAO Junior Reading Scores

	Procedure					
	LL		CM		IRT	
Separately using each Cut-Score	DC	DA	DC	DA	DC	DA
0/1	0.998	0.998	0.999	1.000	0.998	0.999
1/2	0.978	0.979	0.994	0.996	0.998	0.997
2/3*	0.909	0.928	0.980	0.979	0.970	0.972
3/4	0.883	0.915	0.885	0.918	0.910	0.945
Overall	0.774	0.822	0.849	0.892	0.895	0.923

1/2 cut-score

The three values of DC are within 0.004 of each other and the values of the DA are within 0.018 of each other at the 1/2 cut-score. All values are greater than 0.994. The values for DC and DA are close to the values of DC and DA for the 0/1 cut-score. This is likely due to the large difference between the 1/2 cut-score and the population mean, and as shown in Figure 2, while larger than for the 0/1 cut-score, the relatively small number of students around the 1/2 cut- score.

2/3 cut-score

The discrepancy among the values of DC and DA are more pronounced for the 2/3 cut-score than that observed for the 0/1 and 1/2 cut-scores. This is due principally to the LL procedure. The values for the LL procedure, 0.909 and 0.928, are the lowest, while the values for the CM and IRT procedures are higher and closer together, 0.980 and 0.979 and 0.970 and 0.972, respectively. Further, the values for both DC and DA for each procedure are lower than the values observed for the 0/1 and 1/2 cut-scores. This finding is attributable to the fact that the cut-score 2/3 is closer to population mean and the much greater number of students around the 2/3 cut-score (see Figure 2). Hence, there is higher probability of misclassifications and cut-score 2/3 than the two previous cut-scores and this leads to lower values for DC and DA (Lee, 2010).

3/4 cut-score

As for the first two cut-scores, the three values of DC yielded by the three procedures are within 0.027 of each other and the values of the DA are within 0.030 of each other at the 3/4 cut-score. However, the values of DC and DA are the lowest at the 3/4 cut-score, ranging from 0.883 to 0.945. In contrast to the 2/3 cut-score, the values of DC and DA for the LL and CM procedures are close and lower than the corresponding values for the IRT procedure. As for the 0/1 and 1/2 cut-scores and in contrast to the 2/3 cut-score, the distance between the 3/4 cut-score and the population mean is large. However, in contrast to the 0/1 and 1/2 cut-scores and like the 2/3 cut-score, the number of students around this cut-score is large, in fact the largest. The large number of examinees led to a greater number of misclassifications (Lee, 2010).

Sample Results

Bias and Standard Error

Tables 5 to 8 show the values of DC and of DA for the mean of the 100 replicated samples for DC and DA for each cut-score, the bias, and the standard error of the mean, which in the present case is the standard deviation of the 100 sample means. The four tables correspond respectively to the four sample sizes considered: 6,000, 4,500, 3,000, and 1,500. The bias was determined as the difference between the population values of DC and DA at each cut-score from the mean of 100 replicates of DC and DA for each selected sample size for each cut-score. For example, the bias was zero and the standard error was 0.001 for DC for the LL procedure, n = 6, 000 (see Table 5).

Examination of Tables 5 to 8 reveals that the values of bias and the corresponding standard errors are small and also similar. The largest bias for both DC and DA occurred at the 2/3 cut-score. Further, at this cut-score, the largest bias occurred for the IRT procedure. However, the bias values were less than 0.02 (2%). The remaining bias values for the LL and CM procedures at this cut-score and for the three procedures at the other cut-scores were less than or equal to 0.01 (1%). The largest standard error, 0.011, was obtained for DA, IRT procedure, sample sizes 3,000 and 1,500. The remaining standard errors were less than or equal to 0.01 (1%). For example, the bias and standard error for the IRT procedure were, respectively, 0.000 and 0.002 for DC and 0.000 and 0.002 for DA, n = 6,000. Thus, it appears that the three estimation procedures produce unbiased and precise estimates.

Table 5

Decision indices for LL, CM, and IRT models using sample size 6000 for each cut-score for EQAO Junior Reading Scores

		Procedure					
		LL		CM		IRT	
Cut-Scores	Statistics	DC	DA	DC	DA	DC	DA
0/1	Mean	0.988	0.999	0.999	0.999	0.998	0.998
	Bias	0.000	0.000	0.000	0.000	0.000	0.000
	SE	0.001	0.001	0.000	0.000	0.002	0.002
1/2	Mean	0.978	0.979	0.993	0.994	0.999	0.999
	Bias	0.000	0.000	0.001	0.002	-0.001	-0.002
	SE	0.001	0.001	0.000	0.000	0.003	0.002
2/3	Mean	0.906	0.925	0.962	0.969	0.959	0.960
	Bias	0.003	0.005	0.008	0.010	0.011	0.012
	SE	0.005	0.008	0.002	0.002	0.003	0.008
3/4	Mean	0.882	0.918	0.881	0.913	0.899	0.934
	Bias	0.001	-0.003	0.004	0.006	0.009	0.011
	SE	0.002	0.003	0.001	0.001	0.002	0.004

Table 6

Decision Indices for LL, CM, and IRT Models Using Sample Size 4500 for each Cut-Score for EQAO Junior Reading Scores

		Procedure					
		LL		CM		IRT	
Cut-Scores	Statistics	DC	DA	DC	DA	DC	DA
0/1	Mean	0.998	0.999	0.999	0.999	0.998	0.998
	Bias	0.000	0.000	0.000	0.000	0.000	0.000
	SE	0.000	0.001	0.000	0.000	0.001	0.002
1/2	Mean	0.979	0.981	0.995	0.997	0.998	0.998
	Bias	0.001	-0.002	-0.001	-0.001	-0.000	-0.001
	SE	0.001	0.004	0.000	0.000	0.003	0.002
2/3	Mean	0.909	0.926	0.979	0.974	0.943	0.961
	Bias	0.000	0.002	-0.009	0.005	0.007	0.011
	SE	0.002	0.006	0.002	0.001	0.004	0.007
3/4	Mean	0.882	0.916	0.887	0.916	0.906	0.939
	Bias	0.001	-0.001	0.002	0.004	0.004	0.006
	SE	0.003	0.004	0.001	0.001	0.001	0.005

Table 7

Decision Indices for LL, CM, and IRT Models Using Sample Size 3000 for each Cut-Score for EQAO Junior Reading Scores

		Procedure					
		LL		CM		IRT	
Cut-Scores	Statistics	DC	DA	DC	DA	DC	DA
0/1	Mean	0.998	0.999	0.999	0.999	0.998	0.998
	Bias	0.000	0.000	0.000	0.000	0.000	0.000
	SE	0.000	0.001	0.000	0.000	0.001	0.001
1/2	Mean	0.980	0.981	0.990	0.991	0.990	0.991
	Bias	-0.002	-0.002	0.004	0.005	-0.008	-0.006
	SE	0.003	0.006	0.002	0.002	0.004	0.010
2/3	Mean	0.906	0.923	0.961	0.972	0.960	0.953
	Bias	0.002	0.005	0.009	0.007	0.010	0.019
	SE	0.004	0.017	0.003	0.004	0.006	0.011
3/4	Mean	0.881	0.918	0.880	0.913	0.902	0.937
	Bias	0.001	-0.003	0.005	0.006	0.008	0.008
	SE	0.002	0.005	0.001	0.001	0.005	0.009

Table 8

Decision Indices for LL, CM, and IRT Models Using Sample Size 1500 for each Cut-Score for EQAO Junior Reading Scores

		Procedure					
		LL		CM		IRT	
Cut-Scores	Statistics	DC	DA	DC	DA	DC	DA
0/1	Mean	0.998	0.999	0.999	0.999	0.998	0.998
	Bias	0.000	0.000	0.000	0.000	0.000	0.000
	SE	0.002	0.002	0.000	0.000	0.003	0.007
1/2	Mean	0.980	0.983	0.989	0.990	0.991	0.989
	Bias	-0.002	0.003	0.005	0.007	0.007	0.009
	SE	0.002	0.007	0.000	0.000	0.004	0.008
2/3	Mean	0.905	0.931	0.962	0.971	0.959	0.965
	Bias	0.003	-0.004	0.008	0.008	0.011	0.017
	SE	0.003	0.009	0.003	0.003	0.006	0.011
3/4	Mean	0.881	0.913	0.881	0.914	0.901	0.938
	Bias	0.001	0.002	0.004	0.005	0.008	0.007
	SE	0.002	0.003	0.001	0.001	0.005	0.010

Discussion

As expected, the findings reveal that the values of DC were never greater than the values of DA. Where differences were observed, DC < DA for all three procedures. The difference is due to the use of an observed score and an estimated observed score in the case of DC and an observed score and the corresponding estimated true score in the case of DA. Since observed scores contain error of measurement, there are two sources of error in the case of DC and only one source in the case of DA.

The findings also reveal that the three procedures were differentially influenced by the distance between the cut-score and the population mean and the number of students around the cut-score. The absolute values of the differences and the relative numbers of students around each cut-score are:

- for 0/1: 2.49766 largest difference and smallest number of students around cut-score;
- for 1/2: 1.2864 second largest difference and third largest number of students around cut-score;
- for 2/3: 0.1496 smallest difference and second largest number of students around cut-score; and

- for 3/4: 0.4587 third largest difference and largest number of students around cut-score.

Based only on the distance between the cut-score and the mean, the values of DC and DA should be the highest at 0/1, followed by 1/2, and then 3/4, and lastly 2/3. Based only on the number of students around the cut-score, the values of DC and DA should be highest at 0/1, followed in turn by 1/2, 2/3, and 3/4. Thus, both the distance between the cut-score and the mean and the number of students around the cut-score are important. However there is an interaction between the two factors.

As shown in Table 4, the values for DC and for DA for all three models were highest and essentially the same for cut-score 0/1. Then next highest values for DC and DA were for cut-score 1/2 for all three models. However, differences among the three models emerged: both the DC and DA values for the LL procedure were less than the corresponding values for the CM and IRT procedures, which were quite similar. At cut-score 3/4, the values for all three procedures were lower than at cut-score 2/3. Again, the differences between values DC and DA values for the LL procedure were less, but not by as much, than the corresponding values for the CM and IRT procedures, which again were similar. Lastly, the values of DC and DA were the lowest for all three models at cut-score 3/4. However, and in contrast to the results at cut-scores 1/2 and 2/3, the values of DC and DA for the LL and CM procedure were essentially the same but lower, but by not as much as that observed above for the LL procedure, than the values of DC and DA for the IRT procedure. Thus, it would appear that the number of students around a cut-score may be more important than the distance between the cut-score and the mean.

Another factor identified in the literature is the difference in the assumptions regarding the nature of the test forms. Whereas, the IRT model assumes that the test forms are strictly parallel because the item parameters are the same across an infinite number of replicates of test forms, the compound multinomial model and Livingston-Lewis model state the assumption made is that the test forms are randomly parallel.

Given these assumptions, the expectation would be that the values for DC and DA would be more similar for the LL and CM procedures than for the IRT procedure. However, this was not the case. As shown in Table 4, when the DC and DA values for the LL and CM procedures differed, the differences were larger than when the DC and DA values for the CM and IRT differed. It would appear that the CM procedure, which respects the nature of the way sets of items are scored, is more similar to the IRT which works at the item level than to the LL procedure that creates an effective test length.

Despite the differences noted among the LL, CM, and IRT procedures, the estimates of DC and DC were unbiased and precise for each of the four sample sizes considered. Thus, the discussion presented above for the population is applicable at the sample level.

EQAO Junior Mathematics

Population Parameters

This dataset included 127,639 student responses for the Junior (Grade 6) Mathematics assessment. The population parameters for EQAO Mathematics assessment are reported in Table 9 and the frequency distribution of ability estimates is provided in Figure 3. As for Reading, the scores, mean, standard deviation, and cut-scores are expressed in terms of logits. As shown in Table 9, the population mean was -0.3867, the standard deviation was -3.7185, and the distribution was slightly negatively skewed (γ_1= -0.0664), and Platykurtic (γ_2 = -0.8046) (see Figure 3). The internal consistency was again high, 0.87, and the cut-scores progressed from -3.09 (0/1) to 0.99 (3/4).

Table 9

Psychometric Properties for EQAO Mathematics, N =127,639

μ_θ	σ_θ	γ_1	γ_2	α	$cs_{0/1}$	$cs_{1/2}$	$cs_{2/3}$	$cs_{3/4}$
-0.3867	-3.7185	-0.0664	0.8046	0.89	-3.0896	-1.5770	-0.4279	0.9875

Note: μ_θ is the population mean; σ_θ is the standard deviation; γ_1 is the population skewness; γ_2 is the population kurtosis; α is the internal consistency (Cronbach's alpha), $cs_{0/1}$ is cut-score 0/1; $cs_{1/2}$ is cut-score 1/2; $cs_{2/3}$ is cut-score 2/3; and $cs_{3/4}$ is cut-score 3/4.

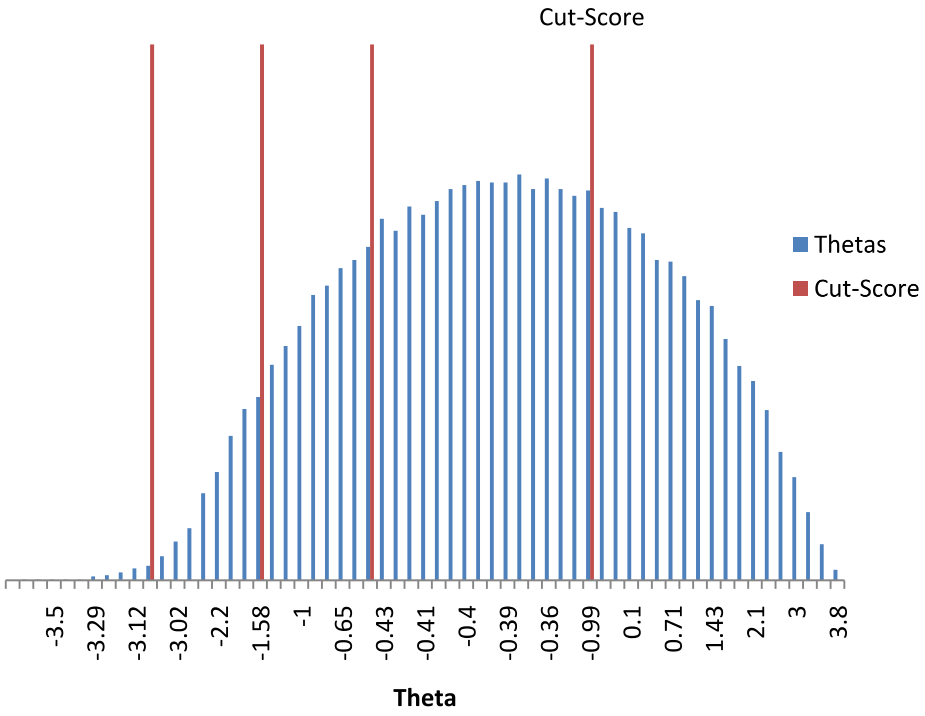

Figure 3: Frequency Distribution of Thetas for EQAO Junior Mathematics

Population DC and DA

The values for DC and DA for the population for Mathematics are reported in Table 10 for each cut-score and overall the cut-scores. The 2/3 cut-score is marked with an asterisk since students that score at or above the 2/3 cut-score have met the provincial standard. Again, as for Reading, whereas the values for DC and DA over all the cut-scores yielded by the IRT procedure were the highest, followed in turn by the CM procedure and the LL procedure, there is an interaction between procedure and cut-score. As the position of the cut-score move towards the extremes of the score distributions, the differences among the values of DC and of DA yielded by the three procedures become smaller, but in different ways. The results for each cut-score are presented first followed by a discussion.

0/1 cut-score.

The three values of DC yielded by the three procedures are within 0.006 of each other and the values of the DA are within 0.003 of each other at the 0/1 cut-score. All values are greater than 0.992. This finding is attributable to the large difference between the 0/1 cut-score and the population mean and the small number of students around cut-score 0/1 (see Figure 3).

Table 10

Decision Consistency and Decision Accuracy Using LL, CM, and IRT Models Conditioned on Cut-Scores for EQAO Junior Mathematics

	Procedure					
	LL		CM		IRT	
Separately using each Cut-Score	DC	DA	DC	DA	DC	DA
0/1	0.992	0.996	0.998	0.999	0.998	0.999
1/2	0.945	0.962	0.987	0.992	0.997	0.976
2/3*	0.865	0.904	0.914	0.939	0.953	0.975
3/4	0.924	0.946	0.867	0.904	0.946	0.970
Overall	0.729	0.808	0.771	0.835	0.868	0.919

1/2 cut-score

The three values of DC yielded by the three procedures are within 0.052 of each other and the values of the DA are within 0.030 of each other at the 1/2 cut-score. All values are greater than 0.945, with the lowest values for both DC and DA yielded by the LL procedure. This is likely due to the large difference between the 1/2 cut-score and the population mean, and the still small number of students around the 1/2 cut-score (see Figure 3).

2/3 cut-score

The discrepancy among the values of DC and DA are more pronounced for the 2/3 cut-score than that observed for the 0/1 and 1/2 cut-scores. However, somewhat in contrast to Reading, while the values for DC and DA yielded by the LL procedure were again the lowest, 0.87 and 0.90, respectively, the values for the CM and IRT procedures differed more for Mathematics, 0.91 and 0.94 vs. 0.95 and 0.98, respectively. However, like Reading, the values for both DC and DA for all three procedures were lower than the values observed for the 0/1 and 1/2 cut-scores. This latter finding is attributable to the fact that the cut-score 2/3 is closer to population mean and the much greater number of students around the 2/3 cut-score (see Figure 3). Hence, there is high probability of misclassifications and this leads to lower values for DC and DA (Lee, 2010).

3/4 cut-score

The three values of DC yielded by the three procedures are within 0.079 of each other and the values of the DA are within 0.066 of each other at the 3/4 cut-score. The values of DC and DA for the CM procedure were lower than the corresponding values for the LL procedure which were lower than the corresponding values for the IRT procedure. Further, the values of DC and DA are the lowest at the 3/4 cut-score especially for the CM and IRT procedures, ranging from 0.87 to 0.97, whereas, the values of DC and DA for LL procedure are higher than values of DC and DA yielded by the CM procedure but lower than values of DC and DA yielded by IRT procedures. As for the 0/1 and 1/2 cut-scores and in contrast to the 2/3 cut-score, the distance between the 3/4 cut-score and the mean is large. However, in contrast to the 0/1 and 1/2 cut-scores and like the 2/3 cut-score, the number of students around this cut-score is large, in fact the largest. The large number of examinees at the 3/4 cut-score can lead to a greater number of misclassifications (Lee, 2010).

Sample Results

Bias and Standard Error

Tables 11 to 14 contain the values of DC and DA for the mean of the 100 replicated samples for DC and DA for each cut-score, the bias, and the standard error of the mean, which in the present case is the standard deviation of the 100 sample means. As with Reading, the four tables correspond respectively to the four sample sizes considered: 6,000, 4,500, 3,000, and 1,500. The bias was determined as the difference between the population values of DC and DA at each cut-score from the mean of 100 replicates of DC and DA for each selected sample size for each cut-score. For example, the bias was zero and the standard error was 0.002 for DC for the LL procedure, n = 6, 000 (see Table 11).

As was the case for Reading, examination of Tables 11 to 14 reveals that the values of bias and the corresponding standard errors for Mathematics are small and also similar. The largest bias for both DC and DA occurred at the 2/3 cut-score. Further, at this cut-score, the largest bias occurred for the IRT procedure. However, the bias values were less than 0.02 (2%). The remaining bias values for the LL and CM procedures at this cut-score and for the three procedures at the other cut-scores were less than or equal to 0.01 (1%). The largest standard error, 0.011, was obtained for DA, IRT procedure, sample sizes 3,000 and 1,500. The remaining standard errors were less than or equal to 0.01 (1%). For example, the bias and standard error for the IRT procedure were, respectively, 0.000 and 0.002 for DC and 0.000 and 0.002 for DA, n = 6,000. Thus, it appears that the three estimation procedures produce unbiased and precise estimates.

Table 11

Decision Indices for LL, CM, and IRT Models Using Sample Size 6000 for each Cut-Score for EQAO Junior Mathematics

		Procedure					
		LL		CM		IRT	
Cut-Scores	Statistics	DC	DA	DC	DA	DC	DA
0/1	Mean	0.992	0.996	0.998	0.999	0.998	0.999
	Bias	0.000	0.000	0.000	0.000	0.000	0.000
	SE	0.002	0.001	0.000	0.000	0.003	0.001
1/2	Mean	0.945	0.961	0.986	0.991	0.994	0.973
	Bias	0.000	0.001	0.001	0.001	-0.003	-0.003
	SE	0.001	0.002	0.000	0.000	0.003	0.002
2/3	Mean	0.868	0.893	0.913	0.946	0.949	0.963
	Bias	0.003	-0.005	0.008	0.007	0.005	0.012
	SE	0.006	0.007	0.002	0.001	0.003	0.008
3/4	Mean	0.922	0.941	0.868	0.906	0.958	0.977
	Bias	0.002	-0.003	0.004	0.006	0.009	0.011
	SE	0.001	0.003	0.001	0.001	0.006	0.004

Table 12

Decision Indices for LL, CM, and IRT Models Using Sample Size 4500 for each Cut-Score for EQAO Junior Mathematics

		Procedure					
		LL		CM		IRT	
Cut-Scores	Statistics	DC	DA	DC	DA	DC	DA
0/1	Mean	0.992	0.994	0.998	0.999	0.998	0.998
	Bias	0.000	0.000	0.000	0.000	0.000	0.001
	SE	0.001	0.001	0.000	0.000	0.001	0.002
1/2	Mean	0.946	0.962	0.986	0.993	0.998	0.970
	Bias	0.000	0.000	-0.001	-0.001	-0.001	0.006
	SE	0.002	0.004	0.000	0.001	0.003	0.006
2/3	Mean	0.863	0.901	0.900	0.934	0.945	0.961
	Bias	0.003	0.003	0.004	0.005	0.007	0.014
	SE	0.002	0.004	0.002	0.001	0.003	0.005
3/4	Mean	0.923	0.945	0.871	0.908	0.960	0.978
	Bias	0.001	0.000	0.004	0.004	0.014	0.006
	SE	0.004	0.005	0.001	0.001	0.002	0.005

Table 13

Decision Indices for LL, CM, and IRT Models Using Sample Size 3000 for each Cut-Score for EQAO Junior Mathematics

		Procedure					
		LL		CM		IRT	
Cut-Scores	Statistics	DC	DA	DC	DA	DC	DA
0/1	Mean	0.992	0.996	0.998	0.999	0.995	0.997
	Bias	0.000	0.000	0.000	0.000	0.003	0.002
	SE	0.001	0.001	0.000	0.000	0.001	0.001
1/2	Mean	0.941	0.960	0.987	0.991	0.989	0.970
	Bias	-0.004	-0.002	0.004	0.001	-0.008	-0.006
	SE	0.004	0.005	0.001	0.002	0.005	0.010
2/3	Mean	0.867	0.909	0.909	0.937	0.928	0.956
	Bias	0.002	0.005	0.005	0.002	0.025	0.019
	SE	0.003	0.014	0.003	0.003	0.003	0.016
3/4	Mean	0.922	0.938	0.945	0.907	0.966	0.981
	Bias	0.002	-0.008	0.012	0.003	0.0013	0.011
	SE	0.003	0.004	0.001	0.001	0.003	0.005

Table 14

Decision Indices for LL, CM, and IRT Models Using Sample Size 1500 for each Cut-Score for EQAO Junior Mathematics

		Procedure					
		LL		CM		IRT	
Cut-Scores	Statistics	DC	DA	DC	DA	DC	DA
0/1	Mean	0.992	0.996	0.998	0.999	0.998	0.999
	Bias	0.000	0.000	0.000	0.000	0.000	0.000
	SE	0.001	0.002	0.000	0.000	0.004	0.007
1/2	Mean	0.945	0.962	0.988	0.992	0.985	0.975
	Bias	0.000	0.000	0.011	0.000	0.012	0.001
	SE	0.002	0.007	0.000	0.000	0.004	0.008
2/3	Mean	0.870	0.904	0.917	0.949	0.965	0.982
	Bias	0.005	0.000	0.003	0.010	0.012	0.017
	SE	0.002	0.007	0.001	0.002	0.005	0.010
3/4	Mean	0.922	0.944	0.870	0.909	0.932	0.966
	Bias	0.002	0.002	0.003	0.005	0.012	0.004
	SE	0.001	0.004	0.001	0.001	0.004	0.009

Discussion

As expected, the findings reported above reveal that the values of DC were never greater than the values of DA. Where differences were observed, DC < DA for all three procedures. The difference is due to the use of an observed score and an estimated observed score in the case of DC and an observed score and the corresponding estimated true score in the case of DA. Since observed scores contain error of measurement, there are two sources of error in the case of DC and only one source in the case of DA.

The findings reported above also reveal that the three procedures were differentially influenced by the distance between the cut-score and the population mean and the number of students around the cut-score. The absolute values of the differences and the relative numbers of students around each cut-score are:

- for 0/1: 2.7029 largest difference and smallest number of students around cut-score;
- for 1/2: 1.1903 second largest difference and third largest number of students around cut-score;
- for 2/3: 0.0412 smallest difference and second largest number of students around cut-score; and

- for 3/4: 0.6008 third largest difference and largest number of students around cut-score.

Based only on the distance between the cut-score and the mean, the values of DC and DA should be the highest at 0/1, followed by 1/2, and then 3/4, and lastly 2/3. Based only on the number of students around the cut-score, the values of DC and DA should be highest at 0/1, followed in turn by 1/2, 2/3, and 3/4. Thus, both the distance between the cut-score and the mean and the number of students around the cut-score are important. However there is an interaction between the two factors.

As shown in Table 10, the values for DC and DA for all three models were highest at cut-score 0/1. Then next highest values for DC and DA were for cut-score 1/2 for all three models. At cut-scores 0/1, 1/2, and 2/3, IRT procedure yielded the highest values for DC and DA followed in turn by CM procedure and LL procedure yielded the lowest values for DC and DA. However, there was a slight change at cut-score 3/4 in which the values of DC and DA for the LL and IRT procedures were similar, and CM yielded the lowest values for DC and DA. Thus, it would appear that the number of students around a cut-score may be more important than the distance between the cut-score and the mean.

As mentioned in the case of Reading assessment, another factor identified in the literature and thought to account for differences among the DC and DA values is the difference in the assumptions regarding the nature of the test forms. Whereas, it is assumed that the test forms are strictly parallel for the IRT model because the item parameters are the same across an infinite number of replicates of test forms, it is assumed that test forms are randomly parallel for the Livingston-Lewis and compound multinomial models. Given these assumptions, the expectation would be that the values for DC and DA would be more similar for the LL and CM procedures than for the IRT procedure. However, this was not the case. As shown in Table 10, when the DC and DA values for the LL and CM procedures differed, the differences were larger than when the DC and DA values for the CM and IRT differed. It would appear that the CM procedure, which respects the nature of the way sets of items are scored, is more similar to the IRT which works at the item level than to the LL procedure that creates an effective test length.

Despite the differences noted among the LL, CM, and IRT procedures, the estimates of DC and DC were unbiased and precise for each of the four sample sizes considered. Thus, the discussion presented above for the population is applicable at the sample level.

Chapter 5: Results and Discussion

The University of Malawi Examinations

The results obtained from the data analyses for the research questions presented in Chapter 1 are provided and discussed in the present chapter for the University of Malawi (U of M) entrance examinations. The present chapter is organized in three parts: the first part is for the U of M Verbal Reasoning examination, the second part is for the U of M English Language examination, and the third part is for the U of M Numerical Reasoning examination. As was the case in the previous chapter, each part is divided into four sections. The population parameters for U of M for the entrance examination presented and discussed in the part are provided in the first section. The values of DC and DA obtained using the Livingston-Lewis, compound multinomial, and IRT models are provided in the second section, followed by presentation of the values for bias and standard error of DC and DA for the 100 replicates associated for each sample size in the third section. A discussion of the full set of results for the assessment is then provided in the fourth section.

U of M Verbal Reasoning Examination

Population Parameters

This dataset included 6,191 student responses for the University of Malawi Entrance Examination for Verbal Reasoning. The population parameters for the U of M Verbal Reasoning examination are reported in Table 15 and the frequency distribution of examinees scores is provided in Figure 4. In contrast to the EQAO assessments discussed in Chapter 4, the scores, mean, standard deviation, and cut-scores are expressed in terms of percentages. As shown in Table 15, the population mean was 60, the standard deviation was 10, and the distribution was slightly negatively skewed (γ_1 = -0.31) and but mesokurtic (γ_2 = -0.02) (see Figure 4). The internal consistency was 0.89 and the cut-scores progressed from 50% (0/1) to 75% (3/4).

Population DC and DA

The values for DC and DA for the population for Verbal Reasoning are reported in Table 16 for each cut-score and overall. As for the two EQAO assessments, whereas the overall values for DC and DA yielded by the IRT procedure were the highest, followed in turn by the CM procedure and the LL procedure, there is an interaction between procedure and cut-score.

Table 15

Psychometric Properties for U of M Verbal Reasoning Examination, N =6,191

μ_θ	σ_θ	γ_1	γ_2	α	$cs_{0/1}$	$cs_{1/2}$	$cs_{2/3}$	$cs_{3/4}$
60	10	-0.31	-0.02	0.89	50	60	70	75

Note: μ_θ is the population mean; σ_θ is the standard deviation; γ_1 is the population skewness; γ_2 is the population kurtosis; α is the internal consistency (Cronbach's alpha), $cs_{0/1}$ is cut-score 0/1; $cs_{1/2}$ is cut-score 1/2; $cs_{2/3}$ is cut-score 2/3; and $cs_{3/4}$ is cut-score 3/4.

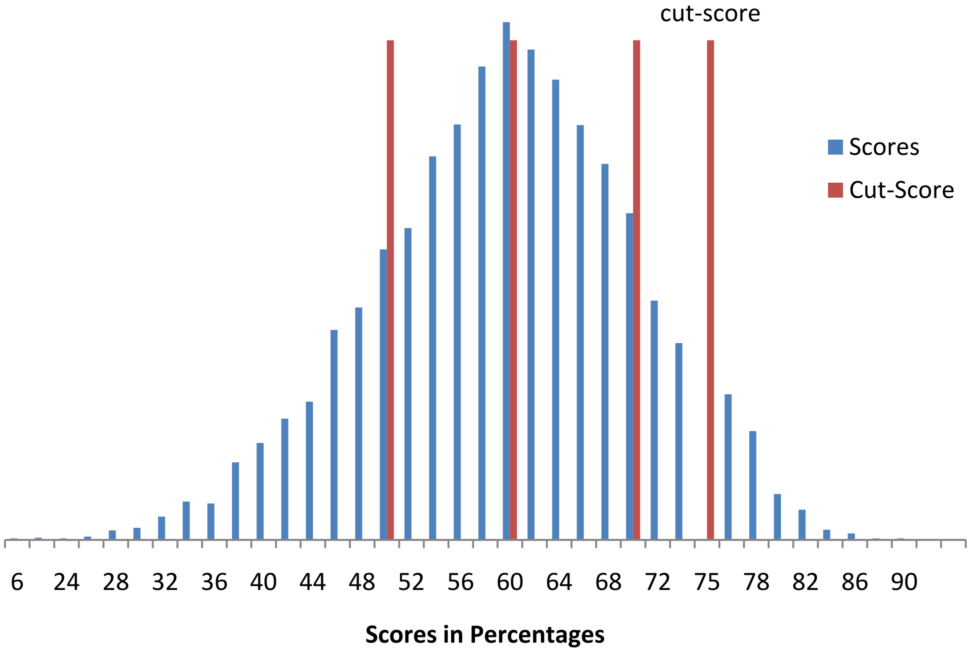

Figure 4: Frequency Distribution for U of M Verbal Reasoning Examination

As the position of the cut-score move towards the extremes of the score distributions, the differences among the values of DC and of DA yielded by the three procedures become smaller, but in different ways. The results for each cut-score are presented first followed by a discussion.

0/1 cut-score

The three values of DC are within 0.009 of each other and the values of the DA are within 0.007 of each other at the 0/1 cut-score. All values, which are greater than 0.823, are the second highest values across the four cut-scores. This is attributable to fact the difference between the 0/1 cut-score and the population mean is the second largest. However, the values for DC and DA are lower than that observed for the EQAO assessments at the 0/1 cut-score presented in Chapter 4 due to the fact that there are a greater number of U of M students around the 0/1 cut-score (cf., Figures 1 and 2 (Ch. 4) and Figure 4).

Table 16

Decision Consistency and Decision Accuracy Using LL, CM, and IRT Models Conditioned on Cut-Scores for U of M Verbal Reasoning Examination

	Procedure					
	LL		CM		IRT	
Separately using each Cut-Score	DC	DA	DC	DA	DC	DA
0/1	0.824	0.867	0.823	0.869	0.832	0.874
1/2	0.706	0.761	0.731	0.792	0.732	0.819
2/3	0.799	0.842	0.814	0.871	0.829	0.885
3/4	0.909	0.923	0.910	0.923	0.911	0.935
Overall	0.529	0.523	0.545	0.551	0.598	0.652

1/2 cut-score

The discrepancy among the values of DC and DA are more pronounced for the 1/2 cut-score than that observed for the 0/1 cut-score. The values for DC and DA for the LL procedure, 0.706 and 0.761 respectively, are the lowest, while the values for the CM and IRT procedures are higher and closer together, 0.731 and 0.792 vs. 0.732 and 0.819, respectively. The values for both DC and DA for each procedure are lower than the values observed for the 0/1 cut-score. The lower values are attributable to the fact that the cut-score 1/2 is at the population mean, and as shown in Figure 4, there is a greater number of students around the

1/2 cut-score than around the 0/1 cut-score. Hence, there is higher probability of misclassifications at 1/2 cut-score than around the 0/1 and this leads to lower values for DC and DA (Lee, 2010).

2/3 cut-score

The three values of DC are within 0.030 of each other and the values of the DA are within 0.043 of each other at the 2/3 cut-score. All values are greater than 0.799. The values for DC and DA are also larger than the values of DC and DA for the 1/2 cut-score, but less than the values at the 0/1 cut-score. These findings are likely due to the difference between population mean and the 2/3 cut-score and the smaller number of students around the 2/3 cut-score than around the 1/2 cut-score and the larger number of students around the 2/3 cut-score than around the 0/1 cut-score (see Figure 4).

3/4 cut-score

The values of DC are the within 0.002 of each other and the values of the DA are within 0.012 of each other at the 3/4 cut-score. Further, the values of DC and DA are the highest at the 3/4 cut-score, with all values greater than or equal to 0.909. The distance between the 3/4 cut-score and the mean is largest and the number of students around this cut-score is the smallest (see Figure 4). Hence, the values of DC and DA at the 3/4 cut-score are the highest (Lee, 2010).

Sample Results

Bias and Standard Error

Tables 17 and 18 show, respectively, the mean values of DC and DA for the 100 replicated samples for DC and DA for each cut-score, the bias, and the standard error of the mean, which in the present case is the standard deviation of the 100 sample means. The two tables correspond respectively to the two sample sizes considered: 3,000, and 1,500. The bias was determined as the difference between the population values of DC and DA from the mean of 100 replicates of DC and DA for each selected sample size for each cut-score. For example, the bias was 0.002 and the standard error was 0.004 for DC for the LL procedure, n = 3,000 (see Table 17).

Examination of Tables 17 and 18 reveals that the values of bias and the corresponding standard errors are small and similar. For example, the bias and standard error for the IRT procedure were, respectively, 0.010 and 0.007 for DC and 0.012 and 0.012 for DA, n = 3,000. The largest difference between the population DC and DA and the corresponding sample mean DC and DA occurred at the 1/2 cut-score. At this cut-point, the largest differences occurred for the IRT procedure. However, the values of bias were less than 0.015 (1.5%). The largest standard error, 0.013, was obtained for DA using the IRT procedure and with n = 1,500. Thus, it appears that the three estimation procedures produce unbiased and precise estimates.

Discussion

As expected, the findings reported above reveal that the values of DC were never greater than the values of DA. Where differences were observed, DC < DA for all three procedures. The difference is due to the use of an observed score and an estimated observed score in the case of DC and an observed score and the corresponding estimated true score in the case of DA. Since observed scores contain error of measurement, there are two sources of error in the case of DC and only one source in the case of DA.

Table 17

Decision Indices for LL, CM, and IRT Models Using Sample Size 3000 for each Cut-Score for U of M Verbal Reasoning Examination

		Procedure					
		LL		CM		IRT	
Cut-Scores	Statistics	DC	DA	DC	DA	DC	DA
0/1	Mean	0.822	0.864	0.828	0.863	0.827	0.866
	Bias	0.002	0.003	-0.005	0.006	0.005	0.008
	SE	0.004	0.006	0.002	0.003	0.006	0.009
1/2	Mean	0.702	0.756	0.729	0.983	0.743	0.807
	Bias	0.004	0.005	0.007	0.009	0.010	0.012
	SE	0.006	0.008	0.003	0.005	0.007	0.012
2/3	Mean	0.799	0.841	0.814	0.868	0.825	0.882
	Bias	0.000	0.001	0.002	0.003	0.004	0.003
	SE	0.003	0.002	0.001	0.002	0.004	0.005
3/4	Mean	0.909	0.923	0.91	0.923	0.911	0.935
	Bias	0.000	0.000	0.000	0.000	0.000	0.000
	SE	0.002	0.001	0.001	0.0001	0.002	0.003

Table 18

Decision Indices for LL, CM, and IRT Models Using Sample Size 1500 for each Cut-Score for U of M Verbal Reasoning Examination

		Procedure					
		LL		CM		IRT	
Cut-Scores	Statistics	DC	DA	DC	DA	DC	DA
0/1	Mean	0.822	0.864	0.829	0.862	0.827	0.866
	Bias	0.002	0.003	-0.004	0.005	0.005	0.008
	SE	0.002	0.007	0.001	0.002	0.006	0.007
1/2	Mean	0.702	0.757	0.730	0.983	0.743	0.806
	Bias	0.004	0.04	0.006	0.009	0.01	0.013
	SE	0.006	0.008	0.003	0.003	0.007	0.010
2/3	Mean	0.799	0.841	0.814	0.869	0.825	0.880
	Bias	0.000	0.001	0.002	0.002	0.004	0.005
	SE	0.002	0.002	0.001	0.002	0.004	0.005
3/4	Mean	0.909	0.923	0.91	0.923	0.911	0.935
	Bias	0.000	0.000	0.000	0.000	0.000	0.000
	SE	0.002	0.001	0.001	0.0001	0.002	0.004

The findings reported above also reveal that the three procedures were differentially influenced by the distance between the cut-score and the population mean and the number of students around the cut-score. The absolute values of the differences and the relative numbers of students around each cut-score are:

- for 0/1: 10. second largest difference and third largest number of students around cut-score;
- for 1/2: 0, smallest difference and largest number of students around cut-score;
- for 2/3: 10 second largest difference and second largest number of students around cut-score; and
- for 3/4: 15, largest difference and smallest number of students around cut-score.

Based only on the distance between the cut-score and the mean, the values of DC and DA should be the highest at 3/4, followed by a tie at 0/1 and 2/3, and lastly 1/2. Based only on the number of students around the cut-score, the values of DC and DA should be highest at 3/4, and the lowest at cut-score 1/2. While the number of students around cut-score 2/3 is higher than the number of students around cut-score 0/1, the difference between the two numbers is not that large. Hence, it is not as clear about whether the values of DC and DA would be greater at 0/1 than at 2/3 or vice-versa. Despite this, both the distance between the cut-score and the mean and the number of students around the cut-score are important factors to consider. However there is an interaction between the two factors.

As shown in Table 16, the values for DC and for DA for all three procedures were highest and essentially the same for cut-score 3/4, which was furthest from the mean and had the fewest number of students around it. The next highest values for DC and DA were for cut-scores 0/1 and 2/3 for all three procedures. However, differences among the three procedures emerged: both the DC and DA values for the LL procedure were less than the corresponding values for the CM and IRT procedures, which were more similar in value. Further the difference between the DC and DA values at the 0/1 cut-score and the 2/3 cut-score were larger for the LL procedure than differences for the CM and IRT procedures, which again were similar. While these two cut-scores were equidistant from the mean, there were fewer students were around cut-score 0/1 than around cut-score 2/3. It appears that the LL procedure is more sensitive to the number of students around the cut-score than the CM and IRT procedures. Lastly, the values of DC and DA were the lowest for all three procedures at cut-score 1/2 and followed the same pattern observed at cut-score 2/3 but not cut-score 0/1.

Thus, it would appear that the number of students around a cut-score may be more important than the distance between the cut-score and the mean.

As mentioned in the previous chapter, another factor identified in the literature is the difference in the assumptions regarding the nature of the test forms. Whereas, the IRT model assumes that the test forms are strictly parallel because the item parameters are the same across an infinite number of replicates of test forms, the compound multinomial model and Livingston-Lewis model state the assumption made is that the test forms are randomly parallel. Given these assumptions, the expectation would be that the values for DC and DA would be more similar for the LL and CM procedures than for the IRT procedure. However, this was not the case. As shown in Table 16, when the DC and DA values for the LL and CM procedures differed, the differences were larger than when the DC and DA values for the CM and IRT differed. It would appear that the CM procedure, which respects the nature of the way sets of items are scored, is more similar to the IRT which works at the item level than to the LL procedure that creates an effective test length.

Despite the differences noted among the LL, CM, and IRT procedures, the estimates of DC and DC were unbiased and precise for each of the four sample sizes considered. Thus, the discussion presented above for the population is applicable at the sample level.

U of M English Language Examination

Population Parameters

This dataset included 6,191 student responses for the U of M Entrance Examination for English Language. The population parameters for this examination are reported in Table 19 and the frequency distribution of ability estimates is provided in Figure 5. As for Verbal Reasoning, the scores, mean, standard deviation, and cut-scores are expressed in terms of percentages. As shown Table 19, the population mean was 50, the standard deviation was 12, and the distribution was essentially symmetric (γ_1 = 0.07) and slightly mesokurtic (γ_2 = -0.23) (see Figure 5). The internal consistency was again high, 0.89, and the cut-scores progressed from 50% (0/1) to 75% (3/4).

Table 19

Psychometric Properties for U of M English Language Examination, N =6,191

μ_θ	σ_θ	γ_1	γ_2	α	$cs_{0/1}$	$cs_{1/2}$	$cs_{2/3}$	$cs_{3/4}$
50	12	0.07	-0.23	0.89	50	60	70	75

Note: μ_θ is the population mean; σ_θ is the standard deviation; γ_1 is the population skewness; γ_2 is the population kurtosis; α is the internal consistency (Cronbach's alpha), $cs_{0/1}$ is cut-score 0/1; $cs_{1/2}$ is cut-score 1/2; $cs_{2/3}$ is cut-score 2/3; and $cs_{3/4}$ is cut-score 3/4.

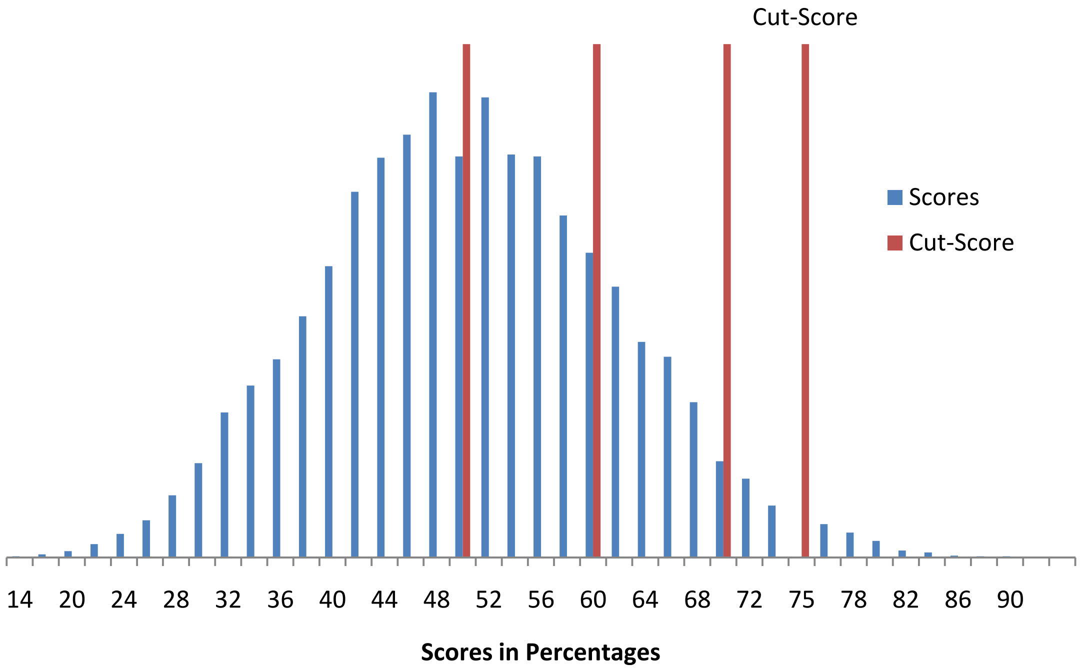

Figure 5: Frequency Distribution for U of M English Language Examination

Population DC and DA

The values for DC and DA for the population for English Language examination are reported in Table 20 for each cut-score and overall. Again, as with two EQAO assessments and the U of M Verbal Reasoning examination, whereas the overall values for DC and DA yielded by the IRT procedure were the highest, followed in turn by the CM procedure and the LL procedure, there is an interaction between procedure and cut-score. As the position of the cut-score move towards the extremes of the score distributions, the differences among the values of DC and of DA yielded by the three procedures become smaller, but in different ways. The results for each cut-score are discussed first followed by a discussion.

0/1 cut-score

The three values of DC are within 0.003 of each other and the values of the DA are within 0.001 of each other at the 0/1 cut-score. Further, the values of DC and DA are the lowest at cut-score 0/1 than at the three other cut-scores. This finding is attributable to the fact that the cut-score 0/1 is at the population mean and the greatest number of students around the 0/1 cut-score (see Figure 5).

Hence, there is high probability of misclassifications and this leads to lower values for DC and DA (Lee, 2010).

Table 20

Decision Consistency and Decision Accuracy Using LL, CM, and IRT Models Conditioned on Cut-Scores for U of M English Language Examination

	Procedure					
	LL		CM		IRT	
Separately using each Cut-Score	DC	DA	DC	DA	DC	DA
0/1	0.763	0.827	0.761	0.827	0.764	0.828
1/2	0.825	0.860	0.829	0.864	0.826	0.877
2/3	0.909	0.927	0.943	0.940	0.956	0.954
3/4	0.946	0.973	0.968	0.975	0.959	0.983
Overall	0.519	0.625	0.532	0.638	0.558	0.643

1/2 cut-score

The three values of DC are within 0.004 of each other and the values of the DA are within 0.017 of each other at the 1/2 cut-score. All values are greater than 0.825. The values for DC and DA are larger than the values of DC and DA for the 0/1 cut-score but smaller at the higher two cut-scores. This is likely due to the difference between the 1/2 cut-score and the population mean and the second largest number of students around the 1/2 cut-score (see Figure 5).

2/3 cut-score

The discrepancy among the values of DC and DA are more pronounced for the 2/3 cut-score than that observed for the 0/1 and 1/2 cut-scores. The value for the LL procedure, 0.91 and 0.93, are the lowest, while the values for the CM and IRT procedures are higher and closer together, 0.94 and 0.94 vs. 0.96 and 0.95. Further, the values for each procedure are higher than the values observed for the 0/1 and 1/2 cut-scores. This finding is likely more attributable to the larger difference between the population mean and the 2/3 cut-score and the somewhat smaller number of students around the 2/3 cut-score than the difference for and number of students around the 0/1 and 1/2 cut-scores (see Figure 5).

3/4 cut-score

The values of DC are the within 0.022 of each other and the values of the DA are within 0.010 of each other at the 3/4 cut-score. Further, the values of DC and DA are the highest at the 3/4 cut-score, with all values greater than or equal to 0.946. The reasons for this latter finding is the fact the 3/4 cut-score is furthest away from the population mean and the smallest number of students around the 3/4 cut-score (see Figure 4). Hence, there is a smaller number of misclassifications of examinees at this cut-score than at cut-score 2/3 (Lee, 2010).

Sample Results

Bias and Standard Error

Tables 21 and 22 show, respectively, the mean values of DC and DA for the 100 replicated samples for DC and DA for each cut-score, the bias, and the standard error of the mean, which in the present case is the standard deviation of the 100 sample means. The results reveal that the values of bias and the corresponding standard errors are small and similar. For example, the bias and standard error for the IRT procedure were, respectively, 0.013 and 0.008 for DC and 0.016 and 0.010 for DA, n = 3,000. The largest difference between the population DC and DA and the corresponding sample mean DC and DA occurred at the 0/1 cut-score. At this cut-point, the largest differences occurred for the IRT procedure. However, the values of bias were less than 0.02 (2%). The largest standard error, 0.013, was obtained for DA using the IRT procedure and with n = 1,500. Thus, it appears that the three estimation procedures produce unbiased and precise estimates.

Table 21

Decision Indices for LL, CM, and IRT Models Using Sample Size 3000 for each Cut-Score for U of M English Language Examination

		Procedure					
		LL		CM		IRT	
Cut-Scores	Statistics	DC	DA	DC	DA	DC	DA
0/1	Mean	0.766	0.831	0.770	0.835	0.777	0.844
	Bias	0.002	0.004	0.009	0.008	0.013	0.016
	SE	0.005	0.012	0.003	0.004	0.008	0.010
1/2	Mean	0.825	0.857	0.833	0.870	0.832	0.884
	Bias	0.001	-0.003	0.004	0.006	0.006	0.007
	SE	0.002	0.004	0.001	0.001	0.005	0.008
2/3	Mean	0.907	0.925	0.946	0.944	0.947	0.948
	Bias	-0.002	-0.002	0.003	0.004	-0.005	-0.006
	SE	0.003	0.003	0.002	0.002	0.003	0.06
3/4	Mean	0.946	0.973	0.968	0.975	0.959	0.983
	Bias	0.000	0.000	0.000	0.000	0.000	0.000
	SE	0.001	0.003	0.001	0.001	0.003	0.002

Table 22

Decision Indices for LL, CM, and IRT Models Using Sample Size 1500 for each Cut-Score for U of M English Language Examination

		Procedure					
		LL		CM		IRT	
Cut-Scores	Statistics	DC	DA	DC	DA	DC	DA
0/1	Mean	0.766	0.824	0.768	0.835	0.777	0.847
	Bias	0.003	-0.003	0.007	0.008	0.013	0.019
	SE	0.003	0.006	0.003	0.003	0.007	0.013
1/2	Mean	0.823	0.862	0.832	0.869	0.833	0.885
	Bias	-0.002	0.002	0.003	0.005	0.007	0.008
	SE	0.002	0.004	0.001	0.002	0.004	0.006
2/3	Mean	0.910	0.928	0.946	0.943	0.958	0.959
	Bias	0.001	0.001	0.003	0.003	0.002	0.005
	SE	0.001	0.002	0.001	0.001	0.005	0.003
3/4	Mean	0.946	0.973	0.968	0.975	0.959	0.983
	Bias	0.000	0.000	0.000	0.000	0.000	0.000
	SE	0.001	0.001	0.0001	0.0001	0.002	0.002

Discussion

As expected, the findings reported above reveal that the values of DC were never greater than the values of DA. Where differences were observed, DC < DA for all three procedures. The difference is due to the use of an observed score and an estimated observed score in the case of DC and an observed score and the corresponding estimated true score in the case of DA. Since observed scores contain error of measurement, there are two sources of error in the case of DC and only one source in the case of DA.

The findings reported above also reveal that the three procedures were differentially influenced by the distance between the cut-score and the population mean and the number of students around the cut-score. The absolute value of the differences and the relative numbers of students around each cut-score are:

- for 0/1: 0, smallest difference and largest number of students around cut-score;
- for 1/2: 10, third largest difference and second largest number of students around cut-score;
- for 2/3: 20, second largest difference and third largest number of students around cut-score; and

- for 3/4: 25, largest difference and smallest number of students around the cut-score.

Based only on the distanced between the cut-score and the mean, the values of DC and DA should be the highest at 3/4, followed by 2/3, and then 1/2, and lastly 0/1. Based only on the number of students around the cut-score, the values of DC and DA should be highest at 3/4, followed in turn by 2/3, 1/2, and 0/1. This was the case. However there was an interaction between the cut-scores and the three procedures.

As reported in Table 20 the values for DC and for DA for all the three procedures were essentially the same for the cut-scores 0/1 and 1/2. In contrast, they differed at the other two cut-scores. The difference between the LL procedure and the CM and IRT procedures was more pronounced for cut-scores 2/3 in which the value of DC and for DA for the LL procedure was the lowest, followed in turn by the CM procedure and then the IRT procedure However, this effect is somewhat different at cut-score 3/4. At cut-score 3/4, for the value of DC for the LL procedure was lowest, followed by the value for the IRT procedure and then the value for the CM procedure. In the case of DA, the values for the LL and CM procedures were similar but less than the value for the IRT procedure. Thus, it would appear that the number of students around a cut-score may be more important than the distance between the cut-score and the mean.

Another factor identified in the literature is the difference in the assumptions regarding the nature of the test forms. Whereas the IRT model assumes that the test forms are strictly parallel because the item parameters are the same across an infinite number of replicates of test forms, the compound multinomial model and Livingston-Lewis model state the assumption made is that the test forms are randomly parallel. Given these assumptions, the expectation would be that the values for DC and DA would be more similar for the LL and CM procedures than for the IRT procedure. However, this was not the case. As shown in Table 20, when the DC and DA values for the LL and CM procedures differed, the differences were larger than when the DC and DA values for the CM and IRT differed. It would appear that the CM procedure, which respects the nature of the way sets of items are scored, is more similar, to the IRT procedure which works at the item level than to the LL procedure that creates an effective test length.

Again, despite the differences noted among the LL, CM, and IRT procedures, the estimates of DC and DA were unbiased and precise for each of the four sample sizes considered. Thus, the discussion presented above for the population is applicable at the sample level.

U of M Numerical Reasoning Examination

Population Parameters

This dataset included 6,191 student responses for the University of Malawi Entrance Examination for Numerical Reasoning. The population parameters for the U of M Numerical Reasoning examination are reported in Table 23 and the frequency distribution of ability estimates is provided in Figure 6. As for Verbal Reasoning, and English Language examinations, the scores, mean, standard deviation, and cut-scores are expressed in terms of percentages. As shown in Table 23, the population mean was 40, the standard deviation was 15, and the distribution was slightly positively skewed (γ_1 = 0.45) and essentially mesokurtic (γ_2 = 0.03; see Figure 6). The internal consistency was once more high, 0.89, and the cut-scores progressed from 50% (0/1) to 75% (3/4).

Table 23

Psychometric Properties for U of M Numerical Reasoning Examination, N =6,191

μ_θ	σ_θ	γ_1	γ_2	α	$cs_{0/1}$	$cs_{1/2}$	$cs_{2/3}$	$cs_{3/4}$
40	15	0.45	0.03	0.89	50	60	70	75

Note: μ_θ is the population mean; σ_θ is the standard deviation; γ_1 is the population skewness; γ_2 is the population kurtosis; α is the internal consistency (Cronbach's alpha), $cs_{0/1}$ is cut-score 0/1; $cs_{1/2}$ is cut-score 1/2; $cs_{2/3}$ is cut-score 2/3; and $cs_{3/4}$ is cut-score 3/4.

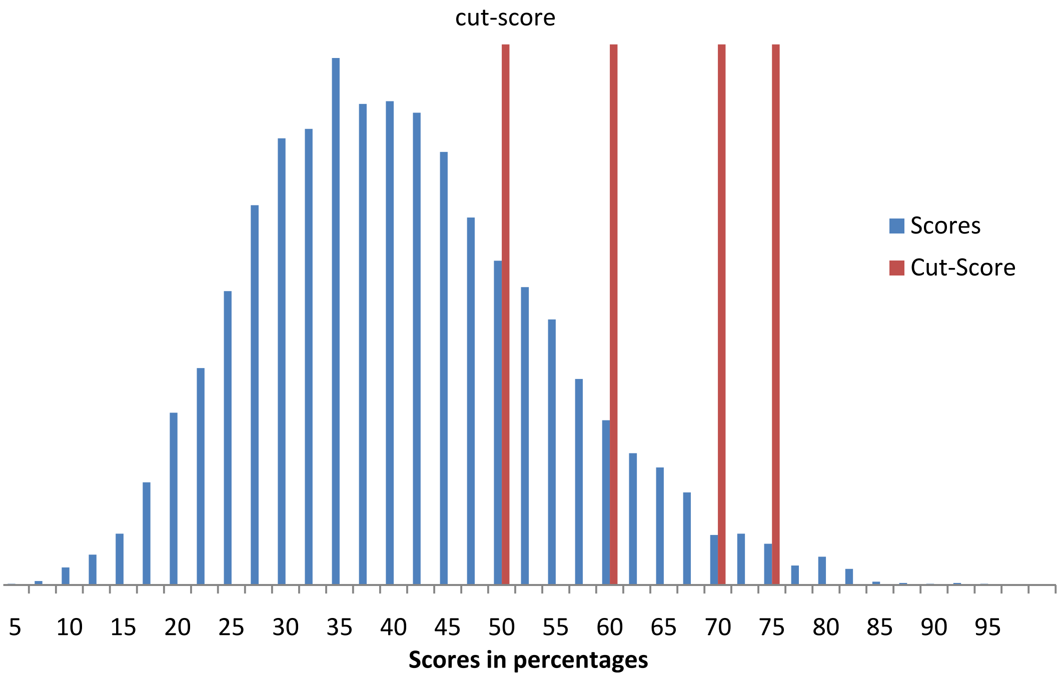

Figure 6: Frequency Distribution for Numerical Reasoning

Population DC and DA

The values for DC and DA for the population for Numerical Reasoning examination are reported in Table 24 for each cut-score and overall. Over again, as with two EQAO assessments and the U of M Verbal Reasoning, and English Language examinations, whereas the overall values for DC and DA yielded by the IRT procedure were the highest, followed in turn by the CM procedure and the LL procedure, there is an interaction between procedure and cut-score. As the position of the cut-score move towards the extremes of the score distributions, the differences among the values of DC and of DA yielded by the three procedures become smaller, but in different ways. The results for each cut-score are discussed first followed by a discussion.

0/1 cut-score

The three values of DC are within 0.023 of each other and the values of the DA are within 0.024 of each other at the 0/1 cut-score. All values are greater than 0.786. This finding is attributable to the large difference between the 0/1 cut-score and the population mean, and there were a few number of students who scored around the 0/1 cut-score (see Figure 6).

Table 24

Decision Consistency and Decision Accuracy Using LL, CM, and IRT Models Conditioned on Cut-Scores for U of M Numerical Reasoning

	Procedure					
	LL		CM		IRT	
Separately using each Cut-Score	DC	DA	DC	DA	DC	DA
0/1	0.786	0.856	0.797	0.855	0.809	0.832
1/2	0.840	0.877	0.839	0.879	0.843	0.879
2/3	0.897	0.908	0.918	0.915	0.926	0.937
3/4	0.928	0.949	0.929	0.952	0.929	0.954
Overall	0.560	0.637	0.573	0.689	0.615	0.700

1/2 cut-score

The three values of DC are within 0.004 of each other and the values of the DA are within 0.002 of each other at the 1/2 cut-score. All values are greater than 0.839. The values for DC and DA are larger than the values of DC and DA for the 0/1 cut-score. This is likely due to the larger difference between population mean and the 1/2 cut-score, and as shown in Figure 6, while larger than for the 0/1 cut-score, the relatively small number of students around the 1/2 cut- score.

2/3 cut-score

The three values of DC are within 0.029 of each other and the values of the DA are within 0.019 of each other at the 2/3 cut-score. All values are greater than 0.897. The values for DC and DA are also larger than the values of DC and DA at cut-scores 0/1 and 1/2. These findings are likely due to the larger difference between population mean and the 2/3 cut-score and the smaller number of students around the 2/3 cut-score than around the1/2 cut-score, and also the smaller number of students around the 2/3 cut-score than around the 0/1 cut-score (see Figure 6).

3/4 cut-score

The values of DC are the within 0.001 of each other and the values of the DA are within 0.005 of each other at the 3/4 cut-score. Further, the values of DC and DA are the highest at the 3/4 cut-score, with all value greater than or equal to 0.93. As for the 0/1 and 1/2 cut-scores and in contrast to the 2/3 cut-score, the distance between the 3/4 cut-score and the mean is large. However, in contrast to the 0/1 and 1/2 cut-scores and like the 2/3 cut-score, the number of students around this cut-score is small, in fact the smallest as shown in Figure 6. The small number of examinees around 3/4 cut-score and combined with the fact that the 3/4 cut-score is furthest away from the population mean, lead to fewer number of misclassifications and hence high values of DC and DA (Lee, 2010).

Sample Results

Bias and Standard Error

Tables 25 and 26 show, respectively, the mean values of DC and DA for the 100 replicated samples for DC and DA for each cut-score, the bias, and the standard error of the mean, which in the present case is the standard deviation of the 100 sample means. The two tables correspond respectively to the two sample sizes considered: 3,000, and 1,500. The bias was determined as the difference between the population values of DC and DA from the mean of 100 replicates of DC and DA for each selected sample size for each cut-score. For example, the bias was 0.003 and the standard error was 0.003 for DC for the LL procedure, n = 3,000 (see Table 25).

Examination of Tables 25 and 26 reveals that the values of bias and the corresponding standard errors are small and similar. The largest difference between the population DC and DA and the corresponding sample mean DC and DA occurred at the 0/1 cut-score. At this cut-point, the largest differences occurred for the IRT procedure. However, the values of bias were less than 0.015 (1.5%). The largest standard error, 0.014, was obtained for DA using the IRT procedure and with n = 1,500. Thus, it appears that the three estimation procedures produce unbiased and precise estimates.

Discussion

As expected, the findings reported above reveal that the values of DC were never greater than the values of DA. Where differences were observed, DC < DA for all three procedures. The difference is due to the use of an observed score and an estimated observed score in the case of DC and an observed score and the corresponding estimated true score in the case of DA. Since observed scores contain error of measurement, there are two sources of error in the case of DC and only one source in the case of DA.

The findings reported above also reveal that the three procedures were again differentially influenced by the distance between the cut-score and the population mean and the number of students around the cut-score. The absolute values of the differences and the relative numbers of students around each cut-score are:

- for 0/1: 10, smallest difference and largest number of students around cut-score;
- for 1/2: 20, third largest difference and second largest number of students around cut-score;

- for 2/3: 30, second largest difference and third largest number of students around cut-score;
- for 3/4: 35, largest difference and smallest number of students around cut-score;

Based only on the distance between the cut-score and the mean, the values of DC and DA should be the highest at 3/4, followed by 2/3, and then 1/2, and lastly 0/1. Based only on the number of students around the cut-score, the values of DC and DA should be highest at 3/4, followed in turn by 2/3, 1/2, and 0/1. Thus, both the distance between the cut-score and the mean and the number of students around the cut-score are important. However there is an interaction between the two factors.

As shown in Table 24, the values for all the three models were essentially the same for the cut-score 1/2, and 3/4 but differed at the other two cut-scores. The difference between the LL, CM and IRT procedures was more pronounced for cut-score 2/3. In contrast, the values of DC for the CM and IRT procedures are essentially the same at the 1/2, and 3/4 cut-scores, and while the values differ at cut-score 2/3, the difference is not as great as that observed for the LL procedure at this point.

Table 25

Decision Indices for LL, CM, and IRT Models Using Sample Size 3000 for each Cut-Score for U of M Numerical Reasoning

		Procedure					
		LL		CM		IRT	
Cut-Scores	Statistics	DC	DA	DC	DA	DC	DA
0/1	Mean	0.789	0.871	0.786	0.862	0.821	0.853
	Bias	0.003	0.005	0.006	0.007	0.007	0.014
	SE	0.003	0.006	0.003	0.004	0.006	0.010
1/2	Mean	0.837	0.875	0.844	0.885	0.834	0.867
	Bias	-0.001	-0.002	0.005	0.006	-0.007	-0.012
	SE	0.004	0.005	0.002	0.002	0.005	0.008
2/3	Mean	0.898	0.906	0.922	0.920	0.934	0.944
	Bias	0.001	-0.002	0.002	0.003	0.005	0.007
	SE	0.002	0.003	0.001	0.001	0.004	0.004
3/4	Mean	0.928	0.949	0.929	0.952	0.929	0.954
	Bias	0.000	0.000	0.000	0.000	0.000	0.000
	SE	0.0001	0.002	0.001	0.001	0.002	0.002

Table 26

Decision Indices for LL, CM, and IRT Models Using Sample Size 1500 for each Cut-Score for U of M Numerical Reasoning

		Procedure					
		LL		CM		IRT	
Cut-Scores	Statistics	DC	DA	DC	DA	DC	DA
0/1	Mean	0.789	0.852	0.785	0.863	0.821	0.855
	Bias	0.003	-0.004	0.008	0.008	0.012	0.015
	SE	0.003	0.005	0.003	0.003	0.006	0.014
1/2	Mean	0.838	0.879	0.844	0.885	0.85	0.889
	Bias	-0.002	0.002	0.005	0.006	0.007	0.008
	SE	0.002	0.004	0.002	0.001	0.004	0.006
2/3	Mean	0.898	0.909	0.92	0.918	0.930	0.941
	Bias	0.001	0.001	0.002	0.003	0.003	0.004
	SE	0.002	0.002	0.000	0.000	0.003	0.007
3/4	Mean	0.928	0.949	0.929	0.952	0.929	0.954
	Bias	0.000	0.000	0.000	0.000	0.000	0.000
	SE	0.002	0.002	0.001	0.001	0.003	0.004

Again, the differences between values DC and DA values for the LL procedure were less, but not by as much, than the corresponding values for the CM and IRT procedures, which again were similar. The values of DC and DA were the highest for all three models at cut-score 3/4. Finally, and in contrast to the findings at cut-scores 1/2, and 3/4, values of DC yielded by the IRT procedure were the highest, followed in turn by CM procedure, and LL procedure yielded the least values for DC. Thus, it would appear that the number of students around a cut-score may be more important than the distance between the cut-score and the mean.

As before, another factor identified in the literature is the difference in the assumptions regarding the nature of the test forms. Whereas, the IRT model assumes that the test forms are strictly parallel because the item parameters are the same across an infinite number of replicates of test forms, the compound multinomial model and Livingston-Lewis model state the assumption made is that the test forms are randomly parallel. Given these assumptions, the expectation would be that the values for DC and DA would be more similar for the LL and CM procedures than for the IRT procedure. However, this was not the case. As shown in Table 24, when the DC and DA values for the LL and CM procedures differed, the differences were larger than when the DC and DA values for the CM and IRT differed. It would appear that the CM procedure, which respects the nature of the way sets of items are scored, is more similar, to the IRT which works at the item level than to the LL procedure that creates an effective test length.

These factors appear to be more dominant than the differences between the two models. Whereas, the IRT model assumes that the test forms are strictly parallel because the item parameters are the same across an infinite number of replicates of test forms, the compound multinomial model and Livingston-Lewis model state the assumption made is that the test forms are randomly parallel. Given the parallel condition is a subset of the randomly parallel condition and the assessments are built to the same specifications and are equated for difficulty, it likely that differences due to the differences in assumption were small and dominated by cut-score placement and numbers of students around the cut-scores.

Again, despite the differences noted among the LL, CM, and IRT procedures, the estimates of DC and DA were unbiased and precise for each of the four sample sizes considered. Thus, the discussion presented above for the population is applicable at the sample level.

Chapter 6: Summary, Conclusion and Recommendations

This chapter is organized in five sections. In the first section, the research questions stated in Chapter 1 together with a summary of the procedures followed is provided. A summary of the findings for decision consistency (DC) and decision accuracy (DA) is presented for the three procedures in relation to the research questions, followed by a discussion of the findings in light of previous research findings in the second section. Limitations of the study are provided in the third section followed by the conclusions drawn in light of the limitations. Implications for practice and recommendations for future research are presented in the last two sections.

Research Questions and Procedures

The purpose of this study was to compare the correctness of decision accuracy and decision consistency of the following three estimation procedures: Livingston-Lewis procedure (LL) CTST approach, the compound multinomial procedure (CM) CTST approach, and Lee IRT procedure. The specific research questions (RQ) addressed included:

1. Do the LL, CM, and IRT procedures yield the same results across four cut-scores and sample sizes?

2. To what extent does the cut-score location affect the magnitude of the values of DC and DA obtained using the LL, CM, and IRT procedures?
3. To what extent does the number of examinees around the cut-score affect the magnitude of the values of DC and DA obtained using the LL, CM, and IRT procedures?
4. Are the LL, CM, and IRT procedures equally consistent and accurate across four cut-scores and different sample sizes?

To address these questions, two population data sources were used: the Junior Reading (N = 128,103) and Mathematics (N = 127,639) assessments administered by the Education Quality and Accountability Office (EQAO) and the three entrance examinations administered by the University of Malawi (U of M; N = 6,191). Each assessment had four cut-scores corresponding to different levels of performance. To determine the degree of bias and the level of precision for both DC and DA, 100 replicated random samples corresponding to four sample sizes (n = 1,500, 3,000. 4,500, and 6,000) for the two EQAO populations and two sample sizes (n = 1,500 and 3,000) for the U of M population were selected.

Summary and Discussion of the Findings

The major findings at the population level were:

a) The IRT procedure tended to provide the highest values for both DC and DA, followed in turn by the CM procedure and the LL procedure across the four cut-scores for both the EQAO assessments and U of M assessments. However, there was an interaction between procedure and cut-score; whereas the differences among the values of DC and DA yielded by the three procedures tended to be smaller for one or both extreme cut-scores, the values of DC and DA tended to be larger for the two middle cut-scores. The results for the LL procedure tended to vary the most, while the results for the CM and IRT procedures tended to be more stable and similar. (RQ 1)

b) Generally, while the differences between the values of DC and DA yielded by the three procedures tended to be smaller for one or both extreme cut-scores, they tended to be larger when the cut-score is closer to the population mean or at the population mean with LL procedure yielding the lowest values and the CM and IRT procedures yielding comparable values for DC and DA. (RQ2).

c) The values of DC and DA increased as the number of examinees around the cut-score decreased for both the EQAO and U of M assessments. When the number of examinees around the cut score was the largest, there was high probability of misclassifications. Hence this led to low values of DC and DA. However, as the number of examinees decreased, the values of DC and DA increased because the number of examinees around the cut-score decreased which gave rise to fewer misclassifications and hence high values for DC and DA (RQ 3).

d) At the sample level, despite the differences noted among the LL, CM, and IRT procedures at the population level, the estimates of DC and DC were unbiased and precise for each of the four sample sizes considered at each cut-score for the EQAO and two sample sizes considered at each cut-score for the U of M assessments (RQ4). Consequently, the findings presented above for the population are applicable at the sample level.

The findings presented above generally agree with findings reported in the literature (Huynh, 1976; Lee, 2010; Wan, Brennan, & Lee, 2007).The reason for the differences among the three models in estimating DC and DA is due *in part* to the differences in the assumptions made for each model. Whereas for the IRT procedure it is assumed that the test forms are strictly parallel because the item parameters are the same across an infinite number of replicates of test forms, for the CM procedure and LL procedure it is assumed that the test forms are randomly parallel. Thus, the assumptions are stronger for the IRT procedure which in turn theoretically leads to the higher values of DC and DA than for CM and LL. The expectation would be that the values for DC and DA would be more similar for the LL and CM procedures than for the IRT procedure. However, this was not the case in the present study. When the DC and DA values differed, the values yielded by the LL procedure were lower than the values for the CM and IRT procedures which were more similar. It would appear that the CM procedure, which respects the nature of the way sets of items are scored, is more similar to the IRT which works at the item level than to the LL procedure that creates an effective test length to accommodate open-response items.

However, the interaction between the distance between a cut-score and the number of examinees around the cut-score appears to be more influential than the distance between a cut-score and the population mean. For example, the values of DC and DA obtained using the LL, CM and IRT procedures were essentially the same when the distance was large and the number of examinees was small. But as the distance between the cut-score and population mean decreased and/or the number of examinees increased, differences among the procedures appeared, but not in a consistent way. For instance, at cut-score 2/3 for the EQAO Reading, while the distance between the population mean and the cut-score was small, there was a large number of students around this cut-score. The values of DC and DA yielded by the CM procedure were slightly higher than the values of DC and DA yielded by the IRT procedure, while the values of DC and DA yielded by LL were quite a bit lower. However, at cut-score 3/4 the distance between the population mean and the cut-score was large as was the number of students around the cut-score. In this case the values of DC and DA yielded by CM were lower than the values of DC and DA yielded by IRT and closer to the values yielded by LL procedure, which were still the lowest (see Table 4, Chapter 4). A second example of the lack of consistency can be seen with the U of M examinations. In the case of the Verbal Reasoning examination, the values for DC and DA for all three procedures were highest and essentially the same for cut-score 3/4, which was furthest from the mean and had the fewest number of students around it. The next highest values for DC and DA were for cut-scores 0/1 and 2/3 for all three procedures. However, differences among the three procedures emerged: both the DC and DA values for the LL procedure were less than the corresponding values for the CM and IRT procedures, which were more similar. Further the difference between the DC and DA values at the 0/1 cut-score and the 2/3 cut-score were larger for the LL procedure than differences for the CM and IRT procedures, which again were similar. While these two cut-scores were equidistant from the mean, there were fewer students were around cut-score 0/1 than around cut-score 2/3. Lastly, the values of DC and DA were the lowest for all three procedures at cut-score 1/2 and followed the same pattern observed at cut-score 2/3 but not cut-score 0/1. Thus, it would appear that the number of students around a cut-score may be more important than the distance between the cut-score and the mean and that the LL procedure is more sensitive to the number of students around the cut-score than the CM and IRT procedures. The reason for the first finding is that large numbers of examinees around cut-scores lead to more misclassifications that give rise to lower values for DC and DA (Huynh, 1976; Lee, 2005; Wan, Brennan, & Lee, 2007). As indicated earlier, the reason for the second finding is that the LL procedure does not fully recognize the nature of open-response items while the CM and IRT do, but in different ways.

Turning to the use of samples, such as in a pilot-testing situation, the results revealed that regardless of cut-score or the number of students in the sample, the three estimation procedures produce unbiased and precise estimates. Therefore, the values of DC and DA found and the pattern of values found at the population level exist at the sample level.

Limitations of the Study

The findings of the present study were confined to single-administration procedures for estimating DC and DA. Consequently, the findings cannot be generalized to a situation where direct estimates of DC and DA are obtained when two parallel test forms are administered or a single test is administered twice.

The smallest samples size considered in the present study was 1,500. While the LL, CM, and IRT procedures yielded unbiased and precise estimates, sample sizes less than 1,500 need to be considered.

Conclusion

Based on the findings of the present study, the compound multinomial procedure should be used to determine DC and DA when classical test score theory is used to analyze a test and its items and the IRT procedure should be used to determine DC and DA when item response theory is used to analyze a test and its items. These procedures can be used with both a population of examinees or a sample of examinees of at least 1,500 students given the sample estimates of DC and DA are unbiased and precise. Lastly, regardless of procedure, the distance between a cut-score and the number of examinees around the cut-score must be taken into account when interpreting the decision accuracy and precision indices.

Implications for Practice

In view of the findings from this study, testing agencies should use the estimation procedure in agreement with test theory – classical or item response – that the agency uses in order to estimate the item parameter and obtain estimates of ability. If the agency uses classical test score theory to conduct an item analysis and to obtain total scores, then the CM procedure should be used to determine the values of DC and DA. If IRT theory is used to estimate item parameters and ability estimates, then the IRT procedure should be used to determine the values of DC and DA. In so doing, the full set of analyses will be in harmony, with the same definition of error at each point in the analyses.

Recommendations for Future Research

Based on the finding of the present study, three topics for future research emerged:

a) When a single administration is used to determine DC and DA, it is necessary to estimate, respectively, a second observed score of a true score. Replication of this study involving two parallel forms or the same test administered on two occasions to confirm the findings of the present study.

b) As indicated above, there was an interaction between the distance between the population mean and the cut-score and the number of examinees around the cut-score. Use of simulation procedures in which the distance between cut-scores and the mean and the number of examinees around a cut-score are systematically varied and how the variations influence the values of DC and DA would help to better understand the nature of this interaction.

c) As pointed out in the limitation of the study and in the conclusion, the smallest sample size considered in this study was 1,500 students. While the sample estimates were unbiased and precise, further study is needed to see if the estimates are unbiased and precise with smaller samples.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, UK: Oxford University Press.

Berk, R. A. (1980). A consumers' guide to criterion-referenced test reliability. *Journal of Educational Measurement*, *17*(4).

Berk, R. A. (1984). Selecting the index of reliability. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 231-266). Baltimore: Johns Hopkins University Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores*, (pp. 397–472), Reading, MA: Addison-Wesley.

Bose J. (2001). Achieving Data Quality in a Statistical Agency: a Methodological

Perspective. *Proceedings of Statistics Canada Symposium 2001*. National Center for Education Statistics, 1990 K St. NW, Washington DC 20006, USA.

Bradlow, E.T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model fortestlets. *Psychometrika*, *64*(*2*), 153–168.

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

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. L. (2004). *BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy (Version 1.0)* (CASMA Research Report No. 9). [Computer software and manual]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma).

Brennan, R. L., & Lee, W. (2006a). *Correcting for bias in single-administration decision consistency indexes* (CASMA Research Report No. 18). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma).

Brennan, R. L., & Lee, W. (2006b). *Some perspectives on KR-21* (CASMA Technical Note No.2). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma).

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

Brennan, R. L., & Wan, L. (2004). *Bootstrap procedures for estimating decision consistency for single-administration complex assessments* (CASMA Research Report No. 7). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma).

Breyer, F. J., & Lewis, C. (1994). *Pass-fail reliability for tests with cut scores: A simplified method* (ETS Research Report No. 94 (39). Princeton, NJ: Educational Testing Service.

Brown, F. G. (1980). *Guidelines for test use: A commentary on the Standards for Educational and Psychological Tests.* National Council on Measurement in Education.

Bourque, M. L., Goodman, G., Hambleton, R. K., & Han, N. (2004). *Reliability* *estimates for the ABTE tests in elementary education, professional teaching knowledge, secondary mathematics and English/language arts* (Final Report). Leesburg, VA: Mid- Atlantic Psychometric Services

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

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(*1*), 1–47.

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 Frame work 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.

Cizek, G. (2001). (Ed.). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, N.J.: Erlbaum.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

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.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*. *16*(*3*), 297–334.

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

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans.* Philadelphia: Society for Industrial and applied Mathematics.

Embertson, S. E., & Reise, S. E. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

English Language Institute, University of Michigan. (2006). *Examination for the Certificate of Proficiency in English 2004–05 annual report*. Ann Arbor: English Language Institute, University of Michigan.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link,

V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement,35*, 137-154.

EQAO. (2009). School Board Report on the Assessments of Reading, Writing and Mathematics, Primary Division (Grades 1–3) and Junior Division (Grades 4–6). *Education Quality and Accountability Office,* Ontario, Canada.

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

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rded.). New York: American Council on Education and Macmillan.

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.

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

Girrbach, C. J., & Claus, R. N. (1982). *Competency Testing: A Review of the Literature*. Evaluation Report. Saginaw Public Schools, Mich. Dept. Of Evaluation Services

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist*, *18*, 519-521.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in eduation and psychology* (3rd ed.). Boston, MA: Allyn & Bacon.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15,* 237-261.

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.

Guo, F. M. (2006). Expected classification accuracy using the latent distribution. *Practical Assessment, Research & Evaluation*, *11*(*6*).

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.

Hagen, R. L., (1983). *An exploration of decision consistency indices for one form test.* (Doctoral Dissertation). Retrieved from ProQuest Dissertations & Theses.

Hambleton, R.K, & Novick, M. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, *10*(3), 159-170.

Hambleton, R.K, & Slater, S. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. *Applied Measurement in Education*, *10*(1), 19-38.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer

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

Hambleton, R.K., & Traub, R. (1973). Analysis of empirical data using two logistic latent trait models. *Br. J. math. Statist. Psychol*, *26,* 195-211.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B.. (1978).

Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48*, 1-47.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27*, 345-359.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small data sets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15*(***x***), 279–291.

Harwell, M., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo Studies in Item Response Theory. *Applied Psychological Measurement:* 20, 101.

Headrick, T. C., & Sawilowsky, S. S. (1999). Simulating correlated multivariate non normal 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.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1996a). Iowa Tests of Basic Skills: Form M: Levels 13-14. Itasca, IL: Riverside Publishing.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1996b).

Constructed- Response Supplement to The Iowa Tests Form 1: Levels 13-14. Itasca, IL: Riverside Publishing.

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

Huynh, H. (1978). Reliability of multiple classifications. *Psychometrika, 43,* 317 - 325.

Huynh, H. (1990). Computation and statistical inference for decision consistency indexes based on the Rasch model. *Journal of Educational Statistics, 15*, 353-368.

Kang, T., Cohen, A. S., & Sung, H.-J. (2005, April). *IRT model selection methods for polytomous items*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika, 27*.

Keller, L.A., Swaminathan, H., & Sireci, S.G. (2003). Evaluating scoring procedures for context dependent item sets1. *Applied Measurement in Education*, *16*(*3*), 207–222.

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.

Kelley, T. L. (1947). *Fundamentals of statistics.* Cambridge: Harvard University Press.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating methods and practices*. New York: Springer-Verlag.

Knupp, T. L. (2009). *Estimating decision indices based on composite scores*. (Doctoral Dissertation). Retrieved from ProQuest Dissertations & Theses.

Lee, W. (2005). *Classification consistency under the compound multinomial model* (CASMA Research Report No. 13). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma).

Lee, W. (2007). Multinomial and compound multinomial error models for tests with complex item scoring. *Applied Psychological Measurement, 31*, 255-274.

Lee, W. (2008). *MULT-CLASS: A computer program for multinomial and* *compound multinomial classification consistency and accuracy (Version 3.0)*. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma).

Lee, W. (2010). Classification Consistency and Accuracy for Complex Assessments Using Item Response Theory. *Journal of Educational Measurement: 47*, 1-17.

Lee, W., & Kolen, M. J. (2008). *IRT-CLASS: A computer program for item response theory classification consistency and accuracy (Version 2.0).* Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma).

Lee W., Brennan R. L., & Wan L., (2009). Classification Consistency and Accuracy for Complex Assessments Under the Compound Multinomial Model: *Applied Psychological Measurement 33: 374.*

Lee, W., Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. *Journal of Educational Measurement, 37*, 1-20.

Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, *26*, 412-432.

Li, H.-H., & Stout, W. F. (1995, April). *Assessment of unidimensionality for mixed polytomous and dichotomous item data: Refinements of Poly DIMTEST*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Linn, R. L. (1979). Issues of validity for measurement in competency-based programs. In M. A. Bunda, & J. R. Saunders (Eds.) *Practices and problems in competency-based education.* Washington, D.C.: National Council on Measurement in Education.

Liu, Y., Bolt, D.M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *1*(*1*), 3–21.

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

Livingston, S. A., & Wingersky, M. S. (1979). Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measurement, 16,* 247-260.

Livingston,S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting* *standards of performance on educational and occupational tests.* Princeton, N.J.: Educational Testing Service.

Lord, F. N. (1965). A strong true score theory, with applications. *Psychometrika, 30,* 239-270.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum Associates, Inc.

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equating." *Applied Psychological measurement*, *8*, 452-461.

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.

Martineau, J. A. (2007). An Expansion and Practical Evaluation of Expected Classification Accuracy: *Applied Psychological Measurement*; 31; 181.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp.153-185). New York: Springer-Verlag.

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

Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177- 196.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American psychologist*, *30*, 955-966.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153-185). New York: Springer-Verlag.

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4: IRT item analysis and test scoring for rating- scale data*. [Computer program]. Chicago, IL: Scientific Software International, Inc.

Muraki, E., & Bock, R. D. (1986). PARSCALE: IRT Item Analysis and Test Scoring for Rating Scale Data. [Computer program]. Chicago, IL: Scientific Software International, Inc.

Nevitt, J. (1998). Simulating univariate and multivariate non normal data: An implementation of the methods of Fleishman (1978) and Vale and

Maurelli (1983). Department of Measurement, Statistics, and Evaluation: Technical Report.

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

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Rosenbaum, P.R. (1988). Item bundles. *Psychometrika*, *53*(*3*), 349–359.

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,* 359-368.

Popham. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6,* 1-9.

Reschly, D. J. (1981). Psychological testing for educational classification and placement. *American Psychologist, 36,* 1021-1034.

Resnick, D. (1980). Minimum competency testing historically considered. In Berliner (Ed.), *Review of Research in Education, 8,* 1-29

Rogers W. T., & Ricker, K. L. (2006). Establishing Performance Standards and Setting Cut-Scores. *The Alberta Journal of Educational Research; Vol. 52, No. 1,* 16-24.

Rogosa, D. (1999). *Accuracy of individual scores expressed in percentile ranks:*

classical test theory calculations. CSE Technical Report 509.National Center for Research on Evaluation, Standards, and Student Testing.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment Research & Evaluation*, 7(14). Available online: http://pareonline.net/getvn.asp?v=7&n=14.

Rudner, L.M. (2004). *Expected classification accuracy*. Paper presented at the annual meeting of the National council on Measurement in Education. San Diego, CA.

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment & Evaluation*, 10(13). Available online: http://pareonline.net/getvn.asp?v=10&n=13.

Rulon, P.J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, *9*, 99-103.

Samejima, F. (1997). Graded response model. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer-Verlag.

Saito, Y. (2003). Investigating the construct validity of the cloze section in the examination for the Certificate of Proficiency in English. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *2*, 39–82.

Schulz, E. M., Kolen, M. J., & Nice-wander, W. A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement, 23*, 347-362.

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

Shepard, L. (1980). Technical issues in minimum competency testing. In D. Berliner (Ed.), *Review of Research in Education, 8,* 30-82.

Skakun, E. N., & Kling, S. (9180). Comparability of methods for setting standards. *Journal of Educational Measurement, 17,* 229-235.

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology*, *3*,271–295.

Spiegelhalter, D. Thomas, A., & Best, N. (2003). WinBUGS version 1.4 [computer program]. Robinson Way, Cambridge CB2 2SR, UK: MRC Biostatistics Unit, Institute of Public Health.

Spray, J. A., & Welch, C. J. (1990). Estimation of classification consistency when the probability of a correct response varies. *Journal of Educational Measurement.27*(1).

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*, 293-325.

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

Subkoviak, M. J. (1978). Empirical investigation of procedures for estimating reliability for mastery tests. *Journal of Educational Measuremen*t, *15*(*2*).

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

Swaminathan, H., Hambleton, R. K., & Algina, J. (1975). A Bayesian decision-theoretic procedure for use with criterion-referenced tests. *Journal of Educational Measurement, 12*, 87-98.

Thissen, D. (1991). MULTILOG 6.3 [Computer program]. Mooresville, IN: Scientific Software.

Thissen, D., Steinberg, L., & Mooney, J.A. (1989). Trace lines for testlets: A use of multiple categorical-response models. *Journal of Educational Measurement*, *26*(*3*), 247–260.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*, 39-49.

Traub, R. E. (1994). *Reliability for the Social Sciences*. Thousand Oaks, California: Cage Publications.

Traub, R., & Rowley, G. (1980). Reliability of test scores and decisions. *Applied Psychological Measurement*, 4, 517-545.

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

U of M. (2009). University of Malawi Entrance Examination: *University Office Zomba, Malawi.*

U of M. (1985). Report on University of Malawi Standard Setting Meeting. *University Office Zomba, Malawi.*

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., Bradlow, E.T., & Du, Z. (2000). Testlet response theory: an analog for the 3-PLuseful in testlet-based adaptive testing. In W.J. van der Linder & C.A.W. Glass (Eds.), *Computerized adaptive testing: theory and practice*, 245–270. Boston, MA: Kluwer-Nijhoff.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory*. New York: Cambridge University Press.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized-adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*(*3*), 185–201.

Wainer, H., Wang, X., Skorupski, W. P., & Bradlow, E. T., (2005). A Bayesian method for evaluating passing scores: the PPoP curve. *Journal of Educational Measurement*, *2*(*3*), 271–281.

Wan, L. (2006*). Estimating Classification consistency for single-administration complex assessment using non-IRT procedures*. (Doctoral Dissertation). Retrieved from ProQuest Dissertations & Theses.

Wan, L., Brennan, R. L., & Lee, W. (2007). Estimating Classification Consistency For Complex Assessments. (*CASMA Research Report No. 22).* Iowa City, IA: University of Iowa.

Wang, W. C., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, *29*(*4*), 296–318.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: theory and application. *Applied Psychological Measurement*, (*26*) No. 1, 109–128.

Wang, X., Bradlow, E. T., & Wainer, H. (2004). User's guide for SCORIGHT (version 3.0): A computer program for scoring tests built of testlets including a module for covariate analysis. Research Report 04–49. Princeton, NJ: Educational Testing Services.

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational Measurement, 37*, 141-162.

Wilcox, R. R. (1981). A review of the beta-binomial model and its extensions. *Journal of Educational Statistics, 6*, 3-32.

Zhang, B. (2008). Investigating Proficiency Classification for the Examination for the Certificate of Proficiency in English (ECPE). *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 6*, 57–75.