CORRECTION FOR GUESSING IN THE FRAMEWORK OF THE 3PL ITEM RESPONSE THEORY

by

TING-WEI CHIU

A Dissertation submitted to the
Graduate School-New Brunswick
Rutgers, The State University of New Jersey
In partial fulfillment of the requirements
For the degree of
Doctor of Philosophy
Graduate Program in Education
Written under the direction of
Gregory Camilli, Ph.D.
And approved by

Gregory Camilli, Ph.D.
Douglas A. Penfield, Ph.D.
Chia-Yi Chiu, Ph.D.
Paul Nichols, Ph.D.

New Brunswick, New Jersey

May, 2010

ABSTRACT OF THE DISSERTATION

Correction for Guessing in the Framework of the 3PL Item Response Theory

By TING-WEI CHIU

Dissertation Chair:
Gregory Camilli, Ph.D.

Guessing behavior is an important topic with regard to assessing proficiency on multiple choice tests, particularly for examinees at lower levels of proficiency due to greater the potential for systematic error or bias which that inflates observed test scores. Methods that incorporate a correction for guessing on high-stakes tests generally rely on a scoring model that aims to minimize the potential benefit of guessing. In some cases, a formula score based on classical test theory (CTT) is applied with the intention of eliminating the influence of guessing from the number-right score (e.g., Holzinger, 1924). However, since its inception, significant controversy has surrounded the use and consequences associated with classical methods of correcting for guessing.

More recently, item response theory (IRT) has been used to conceptualize and describe the effects of guessing. Yet CTT remains a dominant aspect of many assessment programs, and IRT models are rarely used for estimating proficiency with MC items –

where guessing is most likely to exert an influence. Although there has been tremendous growth in the research of formal modeling based on IRT with respect to guessing, none of these IRT approaches have had widespread application.

This dissertation provides a conceptual analysis of how the “correction for guessing” works within the framework of a 3PL model, and two new guessing correction formulas based on IRT are derived for improving observed score estimates. To demonstrate the utility of the new formula scores, they are applied as conditioning variable in two different approaches to DIF: the Mantel-Haenszel and logistic regression procedures.

Two IRT formula scores were developed using Taylor approximations. Each of these formula scores requires the use of sample statistics in lieu of IRT parameters for estimating corrected true scores, and these statistics were obtained in two different ways that are referred to as the pseudo-Bayes and conditional probability methods. It is shown that the IRT formula scores adjust the number-correct score based on both the proficiency of an examinees and the examinee’s pattern of responses across items.

In two different simulation studies, the classical formula score performed better in terms of bias statistics, but the IRT formula scores had notable improvement in bias and r^2 statistics compared to the number-correct score. The advantage of the IRT formula scores accounted for about 10% more of the variance in corrected true scores in the first quartile. Results also suggested that not much information lost due to the use of Taylor approximation. The pseudo-Bayes and conditional probabilities methods also resulted in little information loss. When applied to DIF analyses, the IRT formula scores had lower bias in both the log-odds ratios and type 1 error rates compared to the number-corrected score. Overall, the IRT formula scores decreased bias in the log-odds ratio by about 6% and in the type 1 error rate by about 10%.

Acknowledgements

I always think myself very fortunate to meet people who inspire me in every step of my life. This dissertation would not have been possible without the support from many people. To my committee members, Paul Nichols, Douglas Penfield, and Chia-Yi Chiu, I would like to thank you for your valuable time and generous advice throughout this process. Your guidance and constructive suggestions make this dissertation more complete. In particular, I would like to thank Chia-Yi for helping me clarify the estimation approach in this dissertation. You and Doug were also my calming forces. I would also express my appreciation to faculty members of the Graduate School of Education. I wish to thank Angela O’Donnel and Cindy Hmelo-Silver for their constant encouragement.

I am especially grateful to my dissertation chair and advisor Gregory Camilli whose impact will go on and on in my career. Greg, thank you for everything you have taught me. You have been supported and encouraged me over these years. Your patience and confidence in me helped me through ups and downs in this journey. You are not only a mentor but also a good friend who I can rely on.

Writing a dissertation can be a lonely journey, yet I feel so blessed to have so many great friends there for helping me get through the difficult times. To my great

friends Audrey, Stacy, Chih-Yu, Jerry, and Li-Ting, thank you all for being there sharing my happiness and sadness. You guys are not only my best friends but my family in New Jersey. Michele, thank you for your friendship and being there whenever I needed to discuss statistics. My special thanks to Jennifer for your emotional support and endless help on proof reading my writings.

Lastly, I would like to thank my family. To my sister, brother, and brother-in-law, thank you for your support and taking care of my parents. This helped to make my research less burdened while being thousands of miles from home. Most importantly, I would like to give my special thanks to my parents, Ray-Jar and Mei-Chih for your never-ending love and understanding. Your encouragement and continued support help me persevere in this process. I hope I have made you proud.

I still remember the excitement on the day I received my first hard cover book which was a collection of writings when I was 14. After nearly 20 years, I am about closing my second one, with appreciation and gratefulness to all of you who were there with me on this long journey. Thank you.

Dedication

For my parents, Ray-Jar Chiou and Mei-Chih Chen

Table of Contents

List of Tables

List of Figures

CHAPTER I. INTRODUCTION

Guessing is an important issue with regard to multiple choice (MC) tests. Examinee guessing behavior increases when examinees are encouraged to answer as many questions as possible (e.g., “Try to answer all items”), regardless of whether they know an answer. In this case, guessing is likely to increase, which in turn is likely to introduce a type of error variance distinct from classical random measurement error. Especially at the lower range of test scores, guessing is also likely to introduce a positive bias to examinee proficiency (Rowley & Traub, 1977). While the former problem can lead to incorrect interpretation of a score where there is no actual variability, the latter problem has the potential inflating average test scores. Both problems can result in incorrect interpretations of examinee proficiency relative to a proficiency classification (e.g., partially proficient, proficient, and advanced) or to examinees that do not guess. In general, guessing potentially has a number of impacts on test scores in terms of reliability and validity. For this reason, research focused on remedying the effects of guessing on test scores has a long history in the field of educational measurement.

There have been many approaches to correct or reduce the effects of guessing. A formula score based on classical test theory (CTT) is the most widely known, and is (or has been) used for major achievement test programs such as the SAT Reasoning Test, SAT Subject Tests, and the Graduate Record Examination (GRE) Subject Tests (Bridgeman, & Schmitt, 1997). The classical formula score adjusts a number-correct score by subtracting a proportion of the incorrect responses based on the number of item options. Since its inception, significant controversy has existed regarding the application of this formula score and its consequences (Roberts, 1995). More recently, modern test theory like the three-parameter (3PL) model of item response theory (IRT) has been used to conceptualize and describe the effects of guessing in obtaining examinee's proficiency by adding a pseudo-guessing item parameter (Embretson & Reise, 2000). In IRT, examinee's proficiency level is estimated using item parameters as applied to item response patterns. Both classical formula scoring methods and 3PL IRT models assume that examinees either guess at randomly or respond based on their knowledge (Holzinger, 1924; Waller, 1989). However, both methods ignore the common situation in which ordinary examinees answer questions using partial knowledge to eliminate some choices (Waller, 1989). Therefore, even with an IRT 3PL model, proficiency estimation may be less than optimal because guessing takes the form of many psychological strategies that are difficult to incorporate in a psychometric model.

In the remainder of Chapter I, a short background and basic rationale used to justify correction for guessing are given. The main utility of the classical formula score, as argued first, is actually a strategy for preventing guessing. Second, a number of criticisms of classical formula scoring are reviewed, which fall into the two general categories of behavioral prevention and post hoc statistical correction. A link between IRT and post hoc statistical corrections is then made. Given this background, the objectives of this dissertation are introduced, followed by the methodology and the potential significance of obtaining a clearer understanding of the effects of guessing.

Background for the Correction-for-Guessing

Assessments are used for a variety of purposes and a wide range of scales—from classrooms to state and nation-wide programs. The more frequently encountered purposes, such as school admissions, evaluation of teaching and learning, career placement and recruitment, and professional licensure, employ a variety of item formats (Willingham & Cole, 1997). The most common type of item format in standardized achievement testing is multiple choice (MC) because, compared to other test formats, this format is relatively cost-effective in test development and can be designed to assess many different content domains and skill levels (Ferrara & DeMauro, 2006). Multiple-choice items can also be administered in a relatively short amount of time and are easily scored relative to other item formats such as short or extended constructed responses (e.g., essays) (Ferrara & DeMauro, 2006). Even when tests are designed with both MC and constructed response items, MC items typically comprise a large portion of the total points possible.

Of particular concern with MC items is the possibility of guessing during test administration (Alnabhan, 2002). On a MC test, examinees may encounter items for which they do not recognize the correct option. While some examinees may choose to omit responses to such items, others may choose to guess from among the presented options. When examinees choose to guess, they frequently employ various strategies that are dependent on the context in which the test is administered. For example, if examines are encouraged to answer as many questions as possible, regardless of whether they know an answer, guessing is likely to increase. In general, guessing impacts on test scores in terms of reliability and validity (Burton & Miller, 1999; Ebel, 1972; Lord, 1975).

Classical Formula Scoring

The impact of corrections for guessing has been studied for decades in terms of both preventing guessing, and providing statistical methods of correction for guessing. Corrections for guessing on high-stakes tests are typically applied after administration, and the classical formula score is widely considered to eliminate the influence of guessing (e.g., Holzinger, 1924). Though classical formula scoring is a procedure ostensibly designed to reduce score inflation, it is more accurately defined as a prevention strategy because examinees receive a formula-scoring instruction prior to test administration. Therefore, if examinees responded rationally to the warning of a formula correction, they would omit items for which they do not know the correct answers. Guessing behavior is reduced during a test-taking rather than during scoring.

Illustration of Prevention

To prevent guessing behavior during a test administration, Wise, Bhola, and Yang (2006) introduced an effort-monitoring method in a low-stakes test by using a computer to monitor examinee efforts based on item response time. Because with low-stakes testing, scores carry little or no personal consequences, examinees may not have the motivation to solve the problems. They may engage in guessing by responding to items rapidly, so their test scores may underestimate their true abilities. For that reason, warning messages may prevent guessing due to rapid responses. Note that in this example, the effect of guessing is to deflate test scores, and thus formula-scoring would actually make matters worse.

Arguments for and against Classical Formula Scores

The guiding principle for classical formula scoring is that examinees with the same underlying ability should receive the same score regardless of whether they guess randomly or omit a response. Over the decades that this procedure has been in use, the formula-adjusted scores have generally been shown to have slightly higher reliabilities than uncorrected scores, yet inconsistent results have been found with respect to validity (Lord 1963, Diamond & Evans 1973, Alnabhan, 2002, and Burton, 2002). Still, a number of criticisms of classical formula scoring have been made from both psychological and statistical perspectives.

Psychological Perspective

Although classical formula score has been applied to standardized tests, significant controversy has surrounded the use and consequences associated with classical formula score since its inception (Roberts, 1995). In particular, this controversy has focused on the “invariance effect (IE) and differential effect (DE)” hypotheses (Albanese, 1988). Advocates of the IE hypothesis, such as Angoff & Schrader (1984) asserted that if examinees were forced to respond to omitted items, regardless of scoring instructions received, the chance for them to get the correct responses on those items would not exceed the chance level. They hypothesized that guessing would result in random error, and that everyone would have an equal chance of answering omitted items correctly. Thus, use of classical formula score eliminates the random error (conceptualized as an invariant effect on test scores) caused by guessing.

However, examinees usually do not choose the answer randomly when they do not know the correct option. They might use knowledge on the item to eliminate one or more options, and guess from the remaining options. Besides using partial knowledge, they may also apply different option selection strategies. As a result, the distribution of responses would not be uniform, a condition inconsistent with random guessing (Cronbach, 1984). Therefore, in contrast to Advocates of the IE hypothesis, the advocates of the DE hypothesis assert that certain examinees may omit items for which they have a greater than random chance of answering correctly, in order to avoid the scoring penalty associated with classical formula score. In this case, test scores may underestimate an examinee's true ability. Several studies have shown that when examinees are forced to respond to items they would naturally omit, they have better than chance levels of answering correctly (Bliss, 1980 & Albanese, 1988). Personality and psychological factors may affect guessing behavior (Budescu & Bar-Hillel, 1993; Burton, 2005), and under formula-scoring instruction, certain groups of examinees would be penalized.

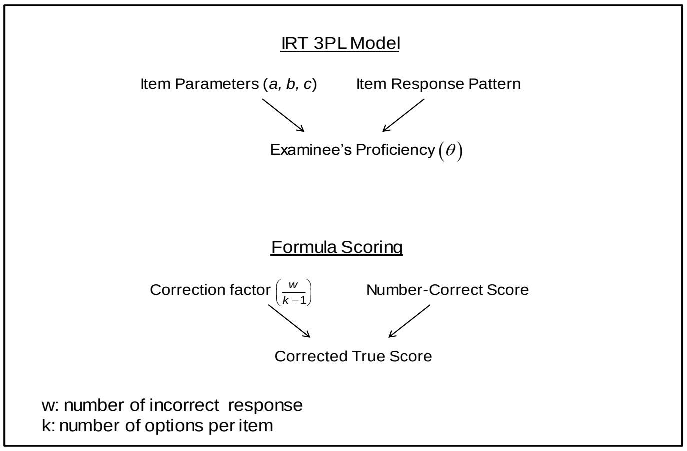
Statistical Perspective

Identical points are subtracted for each wrong response under classical formula score (given a constant number of options). Ultimately, this results in a formula score which is a simple linear transformation from the number-correct score. The classical measures of reliability and validity are identical under linear transformation; therefore, improvements in these indicators of test quality are necessarily the result of changing examinee behavior by a priori formula-scoring warnings.

Modern test theory offers several alternatives to the conceptualization of guessing. Item response theory has been used to conceptualize and describe the effects of guessing. In the context of IRT, the 3PL model (Birnbaum, 1968) is a popular choice for MC tests, because examinee’s proficiency estimates depend on both examinee’s responses pattern and item parameters that describe difficulty, item discrimination, and a lower asymptote (or pseudo-guessing). Indeed, the argument could be made that the IRT 3PL model is preferred for estimating item and individual proficiency parameters in the presence of guessing because it generally fits data better (Hambleton, Swaminathan, & Roger, 1991; Embretson & Reise, 2000).

Both classical formula score and IRT 3PL assume that examinees guess randomly, yet, the effect of guessing on examinee’s score is different. In classical formula scoring methods, examinee’s true scores depend on the correction as applied directly to the number-correct score. See Figure 1-1 for a visual description of this effect.

Figure 1-1. The effect of guessing on IRT 3PL model and formula scoring method


IRT 3PL Model
Item Parameters (a, b, c)
Item Response Pattern
Examinee's Proficiency (θ)
Formula Scoring
Correction factor $\left(\frac{w}{k-1}\right)$
Number-Correct Score
Corrected True Score
w: number of incorrect response
k: number of options per item

The IRT 3PL model adds the guessing parameter to create a nonzero lower asymptote to the item response function for MC items. If an IRT 3PL model fits item responses well, a corrected true score based on IRT scoring could be obtained that is roughly similar to the classical formula scoring. However, as shown below, in the framework of an IRT 3PL model, the effect of the lower asymptote or "guessing" parameter on an examinee's estimated proficiency is not just a function of item parameters, but also of an examinee's item response pattern relative to those parameters. So, the impression given by the classical formula score is incomplete because it is item dependent but not person dependent.

Purpose

The purpose of this dissertation was three-fold and is designed to answer the following questions:

1. How does the “correction for guessing” work within the framework of an IRT 3PL model?
2. Can IRT formula scores be constructed that improve true score estimates?
3. Do IRT formula scores have potential applications in assessment programs using traditional number-correct scores?

The first study in this dissertation was designed to answer question 1 and 2, while a second study was designed to answer question 3. The aim of this dissertation was to investigate guessing in the IRT framework, and then to determine whether IRT formula scores can produce more reliable and accurate estimates of true scores than would be obtained without guessing. Personality and psychological factors as they relate to formula-scoring methods are topics outside the scope of this dissertation. Moreover, the basic assumptions were made in this dissertation that examinees are instructed to provide answers to all questions, and that omitted items are scored as incorrect. The effects of these assumptions were not evaluated.

The goal of this research was to derive IRT formula scores and to compare the properties of these scores to those obtained with classical formula scoring. Guessing was first examined as a conceptual analysis within the framework of an IRT 3PL model to understand how IRT proficiency estimates are adjusted for the lower asymptote (or *c* parameter). Unlike the classical formula scores in which points are subtracted from the number-correct scores based on the number of *incorrect* responses; it was shown that IRT formula scores adjust proficiency estimates for patterns of *correct* responses.

The second goal of this study was to show how IRT formula scores can be developed that provides more reliable true score estimates under certain conditions. Two IRT formula scores were developed and investigated in two simulation studies. Because these IRT formula scores take into account response patterns and item characteristics, they are not simple linear transformations of the number-correct score. Moreover, the IRT formula scores can be implemented without IRT software.

The IRT formula scores were then evaluated in terms of accuracy and accounting for true score variance compared to number-correct and classical formula scores. Previous studies have focused on overall comparisons between an examinee’s number-right score and formula score. Because the effects of guessing behavior are likely to be the strongest with examinees of lower ability (Lord, 1980), separate analyses were conducted within each quartile of the true score distribution in order to explore whether the IRT formula scores perform differently at different score levels. In particular, this study sought to determine if the IRT formula scores of lower-ability examinees improved the most.

The IRT formula scores were obtained, as described below, by a modification of the maximum likelihood method for estimating proficiency (θ). Accordingly, the log likelihood was differentiated with respect to examinee proficiency, set to zero, and the result simplified with several key assumptions. A major goal was to show *how* ability estimates are affected by *c* parameters. It could be argued that no correction in observed score units is required if ability is estimated using IRT. However, the rationale for using IRT 3PL ability estimation in the presence of guessing is not equivalent to a conceptual demonstration of the function of the *c* parameter.

The third goal of this dissertation was to demonstrate an application of IRT formula scores to differential item functioning (DIF). Because IRT formula scores were obtained without IRT parameter estimates, they may have a potential use in large-scale programs that use number-correct scores for secondary analyses. Importantly, DIF analysis is a type of validity evaluation is most often conducted in the observed score metric in most, if not all, state assessment programs, such as California (CA Department of Education, 2006), New York (NY State Department of Education, 2005), and Idaho (Hauser & Kingsbury, 2004). Observed scores are also typically used to examine linguistic issues in assessment programs (e.g., Puhan & Gierl, 2006). Testing organizations such as the Educational Testing Service (ETS) and the CTB McGraw-Hill all conduct DIF analyses based on number-correct scores to examine violations of measurement invariance for ethnic and gender groups (Bridgeman & Schmitt, 1997).

In the second study, DIF was investigated by conditioning on different formula scores as well as the number-correct score, using the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988) and logistic regression (LR) (Swaminathan & Rogers, 1990) procedure. Different factors which are likely to affect the type 1 errors are manipulated, including item parameters, sample size, and ability level (Rogers & Swaminathan, 1993; Roussos & Stout, 1996; Tian, 1999). The goal was to evaluate whether the use of IRT formula scores can improve inferences relative to those obtained with number-correct scores.

In summary, formula scoring in the framework of the 3PL IRT model is conceptually analyzed in this study. Based on this mathematical analysis, IRT formula scores are evaluated for their statistical properties. Finally, these IRT formula scores are applied as conditioning variables in DIF analysis. In the following chapters of this dissertation, a literature review is given in Chapter II on both correction-for-guessing and DIF. In Chapter III, details of the derivations of the new IRT formula scores are then given, and the simulation designs for the DIF analyses are also provided. In Chapter IV, results are presented and explained. Finally, in Chapter V, educational importance, limitation of this dissertation is discussed along with suggestions for future research.

CHAPTER II. LITERATURE REVIEW

In this chapter, a review of different scoring rules for MC tests is given, followed by a review of corrections for guessing in order to provide necessary conceptual context. Different statistical methods related to corrections for guessing are addressed from the perspective of classical test theory (CTT), followed by the perspective of item response theory (IRT) in the framework of the 3PL model. Empirical results are reviewed from different perspectives on corrections for guessing based on CTT, and several IRT investigations are examined. Because IRT formula scores are applied to differential item functioning (DIF), an overview of several current methodologies used in number-correct DIF analysis are also included. Comparisons between different methods, limitations of DIF, and empirical research results are then presented.

Correction for Guessing

A necessary but not sufficient condition for guessing is that an examinee does not have enough knowledge to answer an item correctly. Given its condition, and the fact that an examinee chooses to answer anyway, there is a nonzero probability of selecting a correct answer. The primary effect of such guessing is that both observed test scores and test variance are artificially inflated. Three different methods for scoring MC tests are discussed below: the number-correct score, the existing formula score based on a CTT perspective, and a conceptual approach based on the IRT three-parameter logistic (3PL) model.

Number-Correct Scoring Method

Typically, for a MC item there is only one correct option and each item is scored either right or wrong (wrong = 0, right = 1). Items are equally weighted and summed to a total, which is called the number-right score. In the traditional method of scoring an objective test with n items,

$$n = R + W + O, \tag{2.1}$$

where R represents the number of correct responses, W refers to the number of incorrect responses, and O represents the number of omitted responses. Number-right scoring is the most typical scoring rule and R can be expressed as the total test score for an examinee.

In general, number-correct scores remain an operational aspect of many assessment programs due to a number of factors including: the ease of implementation of statistical techniques; preferences based on historical precedents; and the greater communicative value of classical test statistics to lay audiences. The number-correct score is simple and straightforward, yet it does not adjust for the impact of guessing. This is an important issue because guessing may impart unreliability to test scores that is different from random measurement error, and can result in statistical bias (Rowley &

Traub, 1977). Moreover, the number-correct scoring method can encourage examinees to answer as many questions as possible and increase the likelihood of guessing. Encouragement of guessing might be explicit in the test-taking instructions, e.g., “Try to answer all items.” It also could be implicit; if test-wise examinees infer that there is no penalty for guessing, they may attempt to optimize their scores by answering all items. Encouraging guessing can also lead to examinees losing capacity to self-evaluate (Abu-Sayf, 1979; Kurz, 1999), and thus open the door for a host of undesirable testwiseness or irrational behaviors that affect score validity (Hopkins & Stanley, 1981, Chevalier, 1998).

Classical Test Theory (CTT) Perspective on Correction for Guessing

To reduce the effect of guessing, some testing programs employ a statistical adjustment to number-correct scores. In this case, information about scoring adjustments is given in the test instructions so that examinees understand that, for each incorrect answer, there will be a score adjustment to the total test score. If examinees respond to this information rationally, they will omit their response to any item for which they are completely unsure of the answer. The deceptively simple phrase “formula scoring” is most often used to describe these adjustments. The rationale for using the formula scoring method to correct for guessing is based on three assumptions (Rowley & Traub, 1977; Crocker & Algina, 1986): the examinee either knows the correct answer or has no knowledge at all about the item; the examinee will answer the item correctly with knowledge, or will guess or omit the item; and every incorrect response is randomly chosen by the examinee. This implicitly assumes that the degree of guessing is constant across items.

Consistent with the assumptions above, there are three scoring models used to correct the impact from guessing in the current research literatures. All models are consistent with the random-guessing assumptions above.

Reward for omitted items. The first scoring model rewards examinees additional points for not guessing. The formula can be written as

$$C_O = R + \frac{O}{k}, \tag{2.2}$$

where C_O is the corrected observed score, and k represents the number of options per item. This formula assumes that if the examinee had attempted an omitted item, the probability of answering correctly would be $1/k$, which corresponds to a random guess (Crocker & Algina, 1986; Kurz, 1999).

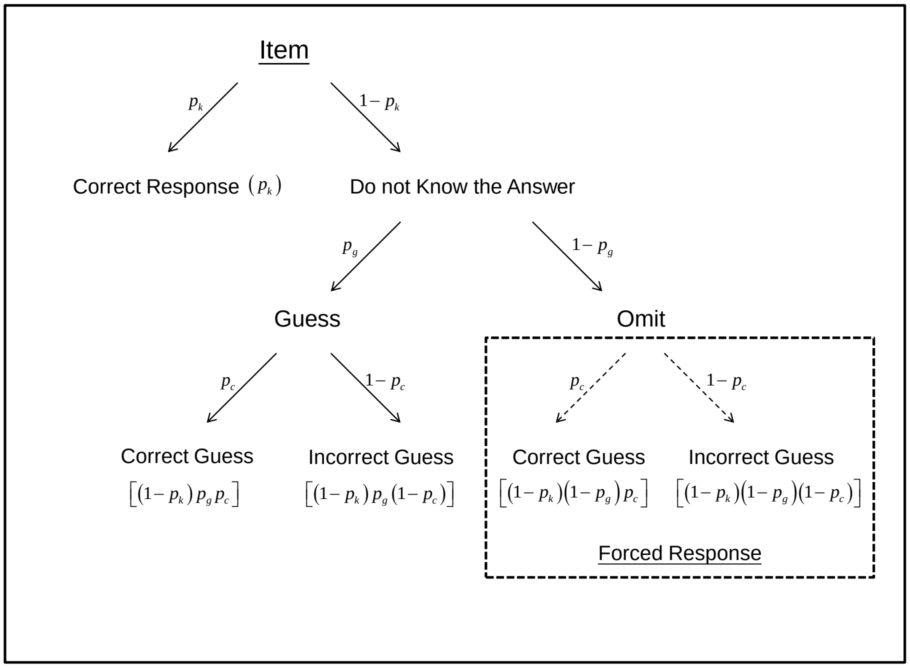
Rights minus wrongs. The second and the most widely used method is also known as the *formula score* or *negative marking* which can be expressed as

$$C_K = R - \frac{W}{k-1}, \tag{2.3}$$

where C_K represents the estimated number of correct response based on knowledge.

Higham (2007) proposed a psychological threshold model, shown in Figure 2-1, to describe how formula scoring method works in psychological terms.

Figure 2-1. The Psychological Threshold Model Implied by Classical Formula Score.



According to this schema, examinees have probability (p_k) to select the correct answer when in fact they know the answer. This probability p_k is referring to the psychological threshold of answering the item with enough knowledge. Next, when the examinee does not know the answer, the examinee decides whether to guess (p_g) or not to guess (1- p_g) on those items for which he/she does not have certain knowledge. Based on the CTT assumption, examinees select an option randomly when they do not know the answer of the items; therefore, if a guess is made, the probability to answer the item correctly is c ($p_c = c = 1/k$). The ratio of correct guessing to incorrect guessing [$c/(1-c)$ *or* $1/(k-1)$] can be used to estimate the number of correct guessing from the number of incorrect guessing. As a result, the ratio represents the portion score necessarily to be adjusted from the number of incorrect answers.

Although two scoring methods described above give numerically different value and adjustment on test scores, the resulting score is a linear transformation of the number-correct score. Furthermore, given $n=R+W+O$, Equation (2.2) can be rewritten as

$$C_O = \frac{n}{k} + \frac{k-1}{k} C_K . \tag{2.4}$$

If there are no omitted items, C_O is equal to R and is perfectly correlated with C_K. Both scoring methods provide the identical rank order of scores for fixed values of the same set of item responses.

Scharf and Baldwin method. Scharf and Baldwin (2007) proposed a third method which takes the omitted items into account in a maximum penalty equation. This method considers omitted items and items not attempted to be incorrectly answered. By replacing W with n-R, and C_K with C_M in Equation (2.3), the number of items assumed correctly answered as a result of the examinee’s knowledge can be written as

$$C_M = R - \frac{n-R}{k-1} . \tag{2.5}$$

Scharf and Baldwin (2007) compared three different methods above and concluded that the maximum penalty equation is the least justifiable; whereas the formula scoring method can be regarded as the fairest assuming that random guessing on average will be cancelled in the final score.

Empirical Research Results of Formula Scoring based on CTT

Psychological factors. The different correction methods described above can be considered as simple linear transformations of the number-correct score. Therefore, the reliability and validity should be invariant except for psychological factors involved in guessing. In fact, over three decades of research have shown that the formula score yields slightly higher reliability estimates than the uncorrected score method, but inconsistent results have been found with respect to validity (Lord 1963, Diamond & Evans 1973, Alnabhan, 2002, and Burton, 2002). Lord (1963) argued that the increased validity due to formula score occurs only with items having less than five options, the test is more difficult, and the examinees vary differently in their tendency to guess. Thus, it appears in these instances that some mild psychological effects are operative.

Personality factors. As noted by Burton (2005) and others, personality factors may affect guessing behavior. An application of the formula score is usually provided in the test administration instructions. The argument for the formula score is that examinees are encouraged in advance not to guess when they do not feel confident about answering an item. Some examinees who understand the formula scoring function will minimize their guessing during the exam. In turn, irrelevant test-score variance and bias associated with guessing will be reduced. However, examinees may have different reactions to formula-scoring instructions. Examinees that are more prone to risk-taking may be more willing to guess. Such risk-taking behaviors are a form of testwiseness and can directly impact examinees' scores.

Diamond and Evans (1973) summarized several studies of individual differences in risk-taking and concluded that risk takers are penalized less than compliers by the formula-scoring instruction on objective tests. Avila and Torrubia (2004) conducted a meta-analysis of 19 medical examinations to look at how personality factors affect examinees' answering behaviors during an exam. They found that extraversion and sensitivity to rewards and punishments (inhibition vs. disinhibition) can affect the number of incorrect responses and omitted items, even when examinees are aware that formula scoring applied. Davis (1967) recommended a test instruction to be used under formula scoring method:

> Your score will be the number of items you mark correctly minus a fraction of the number you mark incorrectly. You should answer questions even when you are

> not sure your answers are correct. This is especially true if you can eliminate one or more choice as incorrect or have a hunch or feeling about which choice is correct. However, it is better to omit an item than to guess wildly among all of the choice given. (p.43)

To reduce personality effects, it is important to ensure all examinees are informed clearly about the answering strategy which will benefit their scores (Frary, 1988).

Effects on high and low ability examinees. Angoff and Schrader (1984) conducted a study using data from the SAT and the GMAT to examine the effects of the formula scoring method. In this study, the formula scoring method was applied to both the number-right scoring instructions, and the formula-scoring instructions. The results suggested that the formula scoring method did not necessarily penalize examinees' scores, because the differences between the groups (different instructions) were small. As suggested by Lord (1980), differences due to instructions may only occur for low-ability regions of proficiency. These examinees tend to pick the attractive but wrong options more regularly, and their scores on difficult items are often worse than random guessing.

Bliss (1980) found that the formula scores tend to penalize high-ability examinees. Examinees of high-ability consider formula scoring instruction more seriously and usually hesitate to guess on items without knowing correct answers. However, this effect was not confirmed in other studies. Lord (1975) suggested that based on the stated assumptions of the formula scoring method, the number of omitted items is the major controller for improving score accuracy. He argued that the greatest improvement in accuracy should occur for lower-ability examinees who omit many items, and is insubstantial for high-ability students who know more correct answers. Crocker and Algina (1986) added that the increasing accuracy for lower-ability examinees may be due to their lack of understanding of the formula-scoring instructions. Because they do not understand the instructions, they may not properly employ the instructions and may be more likely to guess at items which they should not attempt. In this case, using the formula-scoring method can ironically ensure more reliable prediction of an examinees' true ability.

The role of omits. The number of omitted items is a critical feature of the quality of the corrected score. Ben-Shakhar & Sinai (1991) documented that females are more likely to omit questions than males even under number-correct scoring instruction. However, Grandy (1987) founds no significant difference between males and females on omitting items. Examinees from minority backgrounds tended to omit more items based on results from the GRE General Test (Bridgeman & Schmitt 1997).

Partial information and confident misinformation. One major consideration regarding formula scoring is that examinees' guessing behavior does not always comply with the random guessing assumptions. One possible violation is that the correction ignores partial knowledge. Examinees are assumed either to know the correct answer or to have no knowledge at all under formula scoring method. Yet partial knowledge can arise in at least two related forms. Some incorrect options may be more off-target than others, or an examinee may choose an option by eliminating one or more incorrect options (Rowley & Traub, 1977). From this point of view, the correction becomes a penalty for not guessing because examinees have a better chance to get an item correct. Burton (2002) suggested that when the "negative marking" is applied to true/false tests, the examiner would have to convince examinees in advance that they are more likely to get a higher score when they answer the items for which they have more than 50% certainty.

However, there are also pitfalls to number-correct scoring. Examinees who answer items incorrectly based on confident misinformation are at a particular disadvantage with number correct scoring. These examinees omit answers even if instruction specifies that no penalty for guessing is applied. Other examinees without any knowledge may prefer to guess randomly. Thus, relative to other examinees, both number-correct and formula scoring methods have the potential to penalize students whose answers are based on faulty knowledge or reasoning. Bridgeman and Schmitt (1997) suggested that for tests scored using the number-correct scoring method, examinees will unquestionably be at a disadvantage if they are reluctant to guess. Moreover, if examinees are unwilling to use an informed guess, their chance to perform well on a test using the formula scoring method may be small. Furthermore, the distinction between partial knowledge and guessing becomes particularly difficult for MC items requiring complex cognitive behaviors, such as multi-step problem solving.

Examinees of high-ability may benefit from guessing on those uncertain items because their guesses are more likely determined by accurate partial knowledge, even though it is incomplete. On the other hand, it may be a disadvantage for the low-ability examinees to guess, because their guesses are based on incorrect partial information (Angoff, 1989).

Summary of empirical results. Formula scores would seem to work the best when the three assumptions are true: Either the examinee knows the correct answer and chooses it, or the examinee does not know the answer and omits it, or the examinee select one option randomly (Frary, 1988). Muijtjens, Mameren, Hoogenboom, Evers, & van der Vleuten (1999) provided a useful discussion of these issues. Based on their research, the number-correct scoring method takes more account of partial knowledge than does the formula scoring method. They observed that, whereas the number correct scoring method tends to decrease bias, the formula scoring method tends to increase reliability. Given this tradeoff, they preferred to use the number-correct score, but they also concluded that the psychometric and the educational aspects should be weighed when choosing a scoring method and this choice may vary depending on the specific testing circumstances.

An Item Response Theory (IRT) Perspective of Correction for Guessing

Modern test theory offers several alternatives to the conceptualization of correction for guessing. Item response theory provides a statistical framework for describing how item and examinees characteristics interact in test performance. In IRT, an examinee's performance depends on an overall ability θ, and the relationship between the item performance of an examinee and traits can be described by a parametric item response function (IRF) (Hambleton, Swaminathan, & Roger, 1991). An IRF maps changes in trait level θ corresponding to changes in the probability of a correct response (Embretson & Reise, 2000). Compared with CTT, IRT ability estimates can provide a wider range of detailed predictions on unobserved testing situations given that item parameters are available. In IRT, examinees with different ability levels θ have different probabilities of answering a particular item correctly. A given model represents the probability of a discrete response to an item as a function of a person parameter and one or more item parameters. The most common models employ one proficiency and either one (1PL), two (2PL), or three (3PL) item parameters. The probability λ_i for the examinee with a certain ability level (θ) to answer a particular item right based on 3PL can be represented as

$$\lambda_i\left(u_i=1|\theta,a_i,b_i,c_i\right)=c_i+\left(1-c_i\right)P_i, \tag{2.6}$$

where

$$P_i=\frac{\exp\left[Da_i\left(\theta-b_i\right)\right]}{1+\exp\left[Da_i\left(\theta-b_i\right)\right]}. \tag{2.7}$$

The symbol u_i represents the scored response (0 or 1) of an examinee to item i, and the parameters a_i, b_i, c_i are indices of item discrimination, item difficulty, pseudo-chance-level (guessing) parameters, respectively. A scaling constant D = 1.7 is included in the model. The item difficulty parameter, b_i, represents the point on the ability scale where an examinee has 50% chance of giving a correct response when c_i = 0 or $\left(1+c_i\right)/2$ chance otherwise. The item discrimination parameter, a_i, represents item difference in discrimination and is proportional to the slope of the IRF at the point where the ability scale equals b_i. The parameter c_i represents the probability that an examinee with infinitely low ability answering the item correctly. It is assumed that examinees either randomly guess or answer on the basis of knowledge.

To determine which IRT model to use, several rules can be applied to make the decision. The Rasch (1PL) model is favored if each item is equally weighted for scoring. On the other hand, if the goal is to model the existing date with more flexible parameter estimates, the 2PL or 3PL models may be used (Embretson & Reise, 2000). The 3PL model is a common choice because it generally fits MC data better than the 1PL or 2PL models with *c* parameters (Hambleton, Swaminathan, & Roger, 1991; Embretson & Reise, 2000). There are two solutions to define a guessing parameter and add into models: 1) to define a fixed value with $c = 1/k$, where k represents the number of options per item, and 2) to use an identical guessing value for all items which is estimated from the data (San Martin, del Pino, and De Boeck, 2006). After adding a guessing parameter included in the 1PL or 2PL model, the probability for the examinee to answer a particular item right will be similar to Equation (2.6). Because of their flexibility, efficiency, and comprehensiveness, IRT models are widely used in large-scale assessment testing programs in different forms (Yen & Fitzpatrick, 2006).

Lord (1980) suggested that the formula scoring method may be used to estimate examinees' true score for tests designed with any IRT model. According to this method, the formula score correction would be applied directly to the estimated true score based on Equation (2.6). The two critical assumptions of the use of the formula score in IRT are that examinees answer items based on their ability on the specific latent trait only, and that examinees understand and follow the formula-scoring instructions. Lord (1980) suggested that the practice can be used to estimate an examinee’s score even when there are omitted items, as long as the examinee finishes all test items. He also argued that if examinees exhibit different patterns in omitted items or do not finish the test, a modification of this model will be needed.

Modern test theory offers several alternatives to the conceptualization of guessing. Informal approaches to IRT analysis have been attempted in which guessers are identified and excluded from the data set before item parameter estimation with a 2PL model. A second approach is based on the idea that the presence of noise in test score data, such as guessing or other different response strategies, leads to difficulty in the estimation of proficiencies. One solution to this problem is robust estimation as reported by Wainer and Wright (1980). They employed a jackknife scheme for estimating proficiency (θ) based on a Rasch model. In order to compute jackknife pseudo-values, each item was omitted sequentially and θ was re-estimated. Their results indicated that in the jackknife estimates, the effects of unusual item responses (including items that appeared to be answered by guessing) were reduced. Some criticisms of this work were given by Divgi (1986) and Dimitrov (2004) because the procedure can not estimate ability if the score is near zero or perfect. Dimitrov (2004) also suggested methods for improving the jackknife approach on ability estimation.

In contrast, one other formal measurement approach to guessing treats examinees as having a probabilistic membership in latent classes. Yamamoto (1989) formulated a mixture model in which one group (or latent class) of examinees are random guessers, and a second group responds to an item according to the Rasch model. Xie (2002) found that the estimate of item difficulties from the mixture model was closer to the true item difficulties than from a simple Rasch model and in further simulation work, showed that the mixture model provides more accurate estimates than the 3PL model of both item and person parameters (the model was also successful in retrieving the mixture proportions). San Martin et. al. (2006) proposed an ability-based guessing model. They conducted a simulation study with a 3PL model, which guessing was modeled as a function of examinee proficiency θ. They applied the model to different tests in language and mathematics and concluded that the c parameters seemed to depend on proficiency for the reading test, but not for the mathematics test. They concluded that partial knowledge plays more of a role in reading, that is, examinees use their ability to guess to a greater degree on the language test. In another innovative application, Wise and DeMars (2006) proposed the effort-moderated IRT model which takes into account item response time in the estimation of proficiency and item parameters. Their proposed model reduced the effects of rapid guessing which results in better model fit. The effort- moderated IRT model also improved accuracy of item parameters estimates and yielded proficiency estimates with higher convergent validity.

In sum, there has been tremendous growth in the research of formal modeling with respect to guessing. It is obvious that many debates on the application of formula scoring stem from the lack of sensitivity to partial knowledge, and the inconsistency of psychological effects due to formula-scoring instructions. Some research on correction for guessing has been done in IRT theory; however, none of the new IRT approaches have widespread application in formal testing programs.

Differential Item Functioning

In this section, a brief introduction to differential item functioning (DIF) is given. This provides some context for the application of the two IRT formula scores to DIF analysis. The IRT formula scores after development are applied to a non-IRT method of DIF. Thus, after a brief review of some topics in IRT framework for DIF, two major non-IRT methods of DIF analysis are discussed (the Mantel-Haenzel and logistic regression approaches).

Along with the development of testing theories, an issue of great importance to the public is test fairness. In the last two decades, there has been considerable attention in the measurement community to detecting items that may lead to the misestimation of proficiency for particular groups of examinees (Embretson & Reise, 2000). This area of research is known as differential item functioning (DIF), which is defined by psychometricians as follows: “An item shows DIF if individuals having the comparable ability, but from different groups, do not have the same probability of getting the item right” (Hambleton, Swaminathan, & Roger, 1991). Racial, ethnic, and gender differences are the most common groups in DIF research, but other groupings such as social class, age, and geographic region have also been considered (Camilli & Shepard, 1994). The different groups are usually referred to as the focal group, which is the particular group of interest (usually the minority group), and reference group, white is usually a baseline group.

In the past decades, psychometricians have developed many parametric and nonparametric techniques to assess DIF based on classical measurement theory and IRT. Researchers initially focused on group differences in item difficulty, calculated as *p*-values, and then relative differences in *p*-values. However, subsequent research indicated that these methods provide biased estimates of DIF under certain conditions, e.g., when the reference and focal groups truly differ in ability (Cole & Moss, 1989;

Hunter, 1975; Shepard, 1981; Angoff, 1982). In this case, biased type 1 error levels can arise from ignoring item discrimination (Lim & Drasgow, 1990; Angoff, 1993; Camilli & Shepard, 1994).

Compared to CTT, IRT estimates of DIF are based on item response functions (IRF), which describe the probability of answering an item correctly based on the characteristics of the item parameters and underlying ability levels. The definition of DIF then can be stated as “when the IRFs across two subgroups are not identical, the item shows DIF” (Hambleton et. al., 1991). There are two categories of DIF based on the IRT perspective, uniform DIF and nonuniform DIF (Mellenbergh, 1982). An item with uniform DIF is defined as group differences in the probability of answering the item correctly are constant across all ability levels. In other words, the IRFs of the two groups are not identical, but do not cross throughout the range of ability. Nonuniform DIF occurs when an item favors one group members at certain ability levels and favors the other group at other ability levels (assuming two groups). Nonuniform DIF can be observed when the 2PL or 3PL model is used (Camilli & Shepard, 1994; Kristjansson, Aylesworth, McDowell, & Zumbo, 2005). Camilli and Shepard (1994) summarized two different IRT approaches used for detecting DIF: IRT measurement of DIF, and IRT tests for DIF. There are four methods to measure the size of DIF: 1) simple area indices, 2) probability difference indices, 3) *b* parameter difference, and 4) IRF method for small samples. Five methods designed to do statistical test for DIF: 1) test of *b* difference, 2) item drift method, 3) Lord's chi-square, 4) empirical sampling distributions for DIF indices, and 5) model comparison measures.

In typical DIF studies, non-IRT methods are used due to their relative ease of implementation. Moreover, the number of examinees in the focal group (usually from minority) is usually small and with limited ability range (Hambleton et. al., 1991; Camilli, 2006). More flexible IRT models (2PL and 3PL) are more difficult to calibrate in this situation, even thought an argument can be made for employing IRT models with strong assumptions, such as the 1PL. The inevitably poorer parameter estimates for the focal group drive most criticism of these IRT methods. In any case, it may not be possible to conduct a DIF analysis on a relatively small sample.

Because of the potential problems associated with parametric approaches, which may primary be a problem of expert labor, nonparametric methods to detect DIF using observed scores are widely accepted. Several statistical methods have been developed to detect DIF for MC items. The most widely studied and applied methods include the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988), logistic regression (LR) (Swaminathan & Rogers, 1990), the simultaneous item bias test (SIBTEST) (Shealy &

Stout, 1993), and the standardization approach (Dorans & Kulick, 1986). Among these procedures, the MH procedure and the LR procedure are the two most popular.

Mantel-Haenszel Procedure

The MH procedure was designed and used in medical research by Mantel and Haenszel (1959), and applied to psychometrics by Holland and Thayer (1988) in order to inspect item bias on dichotomously scored items. The MH procedure identifies DIF by considering between-group differences in the odds of a correct response, after matching (or conditioning) on observed test scores of the reference and focal groups. The characteristic design of this method is based on a contingency table with a 2 (groups)-by-2 (item scores)-by-*M* (score categories) design that provides the frequencies of item responses (correct and incorrect) of different groups (focal and reference groups) with possible number-correct categories ($m = 1, 2, 3\ldots, M$) as a matching variable. The null hypothesis maintains that, under the conditioning on the observed test score, the odds of correct response will be equal for the focal and reference groups and the odd-ratio will be equal to 1, which is no DIF. The odds ratio for score level m is defined as

$$\alpha_m = \frac{P_{Rm}/Q_{Rm}}{P_{Fm}/Q_{Fm}} = \frac{P_{Rm}Q_{Fm}}{P_{Fm}Q_{Rm}} \tag{2.8}$$

where P_{Rm} and P_{Fm} represent the population proportions of correct responses for the reference and focal groups at the m^{th} score level, and Q_{Rm} and Q_{Fm} represent the corresponding population proportions of incorrect responses. However, when the matching variable is zero or *M* (perfect score), the MH odd ratio will be indeterminate and the odds ratio cannot be calculated. Therefore, for a *M*- item test, the index *m* runs from 1 to *M*-1. The Mantel and Haenszel (1959) procedure also assumes all α_m to be a constant value, and the combined estimate across *m* of the odds ratio α is given by

$$\hat{\alpha}_{MH} = \frac{\sum_m \left[\frac{R_{Rm} W_{Fm}}{N_{Tm}} \right]}{\sum_m \left[\frac{R_{Fm} W_{Rm}}{N_{Tm}} \right]}, \qquad (2.9)$$

where R_{Rm} and R_{Fm} refer to the frequencies of having a correct response to the item in the reference and focal groups, $W_{Rm} = N_{Tm} - R_{Rm}$ and $W_{Fm} = N_{Tm} - R_{Fm}$, and N_{Tm} refers to the total number of responses from both reference and focal group examinees. This odds ratio is an estimate of the DIF effect size and indicates there is no DIF when the value equals to 1. If the ratio is greater than 1, item is said to favor the reference group. On the contrary, if the value is less than 1, the item favors the focal group (Dorans & Holland, 1993; Penfield & Camilli, 2007). Nonetheless, the estimated odds ratio $\hat{\alpha}_{MH}$ is not very useful for DIF interpretation because of its asymmetric distribution. Holland and Thayer (1988) proposed a transformation of $\hat{\lambda}_{MH}$ as delta scores (MH D-DIF) obtained through a transformation to $\hat{\lambda}_{MH} = -2.53 \ln\left(\hat{\alpha}_{MH}\right)$ leading to a symmetric and more useful index for interpretation. When this value differs from 0, DIF and therefore potential bias exist. The converted MH D-DIF has been used as an index of relative item difficulty (Dorans & Holland, 1993; Camilli & Penfield, 1997; Camilli, 2006).

The Mantel-Haenszel chi-square ($MH\text{-}\chi^2$) has a test distribution of chi-square with 1 degree of freedom. It provides the most powerful and uniformly statistical unbiased test of no DIF under the null hypothesis of uniform bias (Holland & Thayer, 1988). As an alternative to $MH\text{-}\chi^2$, the log-odds ratio can be divided by its standard error to obtain a test statistic (Holland & Thayer, 1988). Rules used to measure degrees of DIF were also developed and categorized by ETS regarding both the absolute value of MH D-DIF and the significant test results (Zieky, 1993). Camilli and Shepard (1994) suggested a way to conceptualize the MH odds ratio in the framework of IRT in order to detect DIF. In the IRT 2PL model ($c = 0$), the log odds ratio conditional on θ can be expressed as

$$\begin{aligned}\lambda_{MH-2PL} &= \ln\left(\frac{\exp\left[Da_R\left(\theta-b_R\right)\right]}{\exp\left[Da_F\left(\theta-b_F\right)\right]}\right) \\ &= D\theta\left(a_R-a_F\right)+D\left(a_Fb_F-a_Rb_R\right)\end{aligned} \quad . \tag{2.10}$$

If the item discrimination parameter a is invariant for reference and focal group, Equation (2.10) can be simplified as $\lambda_{MH\text{-}2PL} = Da(b_F - b_R)$. The effect size, $\lambda_{MH\text{-}2PL}$, is then proportional to the difference between item difficulty parameters in the reference and focal group (uniform DIF). Holland and Thayer (1988) emphasized that this method gives an unbiased estimate of DIF under the Rasch model (1PL, with $a = 1$) with the assumptions that all items included in matching variable, all other items are measurement invariant across groups, and data are random samples from both groups. However, if a parameters are different in two groups, $\lambda_{MH\text{-}2PL}$ is no longer proportional to the difference between the b parameters (i.e., nonuniform DIF).

The MH log-odds ratio (LOR) procedure is not designed to detect nonuniform DIF, and a number of alternative procedures have been suggested. For example, Roussos, Schnipke and Pashley (1999) proposed a general formula of the MH DIF population parameter which is appropriate for any IRT model and is also applicable for either uniform DIF or nonuniform DIF. However, the findings from this research suggested that more attention is needed to applying the procedure with 3PL data, because guessing can affect the MH DIF estimate for relatively difficult items, especially when the focal group has significantly lower mean proficiency. However, there is little evidence to suggest nonuniform DIF is prevalent, and even in this case, the MH procedure provides a useful index for screening test items for bias.

Logistic Regression Procedure

The logistic regression procedure (LR) is another popular method for detecting DIF due to its ability to take into account the continuous nature of ability levels, and its capability to detect uniform as well as nonuniform DIF. Swaminthan and Rogers (1990) were the first to apply LR procedure on DIF analysis. The LR procedure models the probability of observing each dichotomous item response (0 or 1) as a function of independent variables, which includes a group indicator (G), a matching variable (X, usually the observed total score), and a group-by-ability (*GX*) interaction. The LR procedure employs the assumption that the examinee's ability is well represented by his/her observed total score, and the probability of the individual answering the item correctly is linearly proportional to the examinee's ability (Camilli and Shepard, 1994). The LR model can be written as

$$\begin{gathered} P\left(Y_i=1\right)=\frac{e^{Z_i}}{1+e^{Z_i}} \\ Z_i=\beta_0+\beta_1 X_i+\beta_2 G_i+\beta_3 X_i G_i \end{gathered}, \qquad (2.11)$$

where $P(Y_i = 1)$ represents the probability for individual i to answer the studied item correctly. The coefficient β_1 corresponds to the effect on performance of ability level; whereas β_2 and β_3 correspond to the effects of group and the ability-by-group interaction.

The full model mentioned in Equation (2.11) can be simplified depending upon three different situations: no DIF, uniform DIF, and nonuniform DIF. Camilli and Shepard (1994) summarized stepwise selection of model testing using likelihood ratio statistics. First, conditioned on observed totals score, the presence of nonuniform DIF is evaluated by comparing $Z_i = \beta_0 + \beta_1 X_i + \beta_2 G_i + \beta_3 X_i G_i$ to $Z_i = \beta_0 + \beta_1 X_i + \beta_2 G_i$. Next, to test the uniform DIF, comparison between $Z_i = \beta_0 + \beta_1 X_i + \beta_2 G_i$ and $Z_i = \beta_0 + \beta_1 X_i$ is conducted. A chi-square statistic is used to evaluate model differences. In addition, this 2-step procedure can be used to compare differences among multiple groups with the addition of dummy codes (Camilli, 2006). The estimate of β_2 is an effect-size measure of DIF and is usually similar in value to MH LOR ($\hat{\lambda}_{MH}$) when the group-by-ability interaction is not included in the model. The coefficients can be estimated by maximum likelihood estimation (Swaminathan & Rogers, 1990).

The coefficient β_2 and coefficients β_3 indicate uniform and nonuniform DIF. If both β_2 and β_3 equal 0, then DIF does not exist. When β_2 shows a statistically significant difference from 0, it suggests that the odds of getting the item correct from two groups are different. The estimate of β_2 is an effect-size measure of DIF and is usually similar in value to MH LOR ($\hat{\lambda}_{MH}$) when the group-by-ability interaction is not included in the model. The case of nonuniform DIF is indicated when β_3 is significantly different from 0. Unsurprisingly, β_1 is almost significantly different from zero; since the examinees with a higher level of ability (or higher observed total score) tend to have a better chance of answering the item correctly. The coefficients can be estimated by maximum likelihood estimation (Swaminathan & Rogers, 1990).

Comparison between the MH and the LR Procedure

Swaminathan and Rogers (1990) designed a simulation study that varied different sample size, test length, and the nature of the DIF when comparing the LR and MH procedures. They concluded that LR is as powerful as MH in detecting uniform DIF and is more powerful than MH in detecting nonuniform DIF, which is not surprising given the assumption of a uniform LOR across score categories. The LR procedure was also found to have slightly higher false positive error (type 1 error) than the MH procedure, and it contained more inconsistent classifications of DIF items (Swaminathan & Rogers, 1990; Narayanan & Swaminathan, 1996; Huang, 1998). Rogers and Swaminathan (1993) extended their study under different conditions (including 2PL, 3PL models) to compare the performance of the LR and the MH procedures. The LR procedure did not function well for very difficult and highly discriminating items. Li and Stout (1996) provided a possible explanation for this result. They pointed out that the presence of pseudo guessing was associated with the inflated type 1 error rates.

Given the similar power in detecting uniform DIF, the MH procedure is relatively easier to implement. According to Rogers and Swaminathan (1993), the LR procedure takes three to four times more computing time in conducting a DIF analysis than the MH procedure. However, if researchers would like to incorporate different variables into the explanation, the LR procedure is preferable (Kristjansson et al., 2005; Swaminathan & Rogers, 1990, Mazor, Kanjee, & Clauser, 1995). In any case, the MH procedure is the most frequently used DIF procedure in practice.

Limitations of DIF

For all of the DIF methods above, it is important to understand that the presence of DIF does not necessarily mean the item is biased. A DIF index only provides an indicator of potentially bias. Moreover, measurement error associated with DIF procedures can include both type 1 error and type 2 errors. It is well known that type 1 errors and type 2 errors are impossible to minimize simultaneously. More false occurrences of the flagged items (type 1 error) implies fewer undetected potential biased items (type 2 error) and vice versa. Most statistical models focus on the reduction of type 1 error; especially from the test developers' and researchers' points of view. However, from the examinee's point of view, the presence of type 2 errors would seem to be a more serious problem.

Camilli and Shepard (1994) suggested that DIF can be detected by examining the content of each item and identifying patterns of significant DIF in similar items. This is because DIF indices may signal multidimensionality in the test (Camilli and Shepard, 1994). Multiple dimensions, as defined by Shealy and Stout (1993), are the essential characteristics of an item that can have an effect on the probability of a correct response.

One of the common assumptions of IRT models is unidimensionality. However, most tests to some degree assess a number of skill dimensions. In characterizing such items, the primary dimension is referred to as the target trait measured by the item, whereas the secondary dimension is referred to the confounding trait. If a secondary dimension is significantly related to a test item, then DIF indices may reflect multidimensionality, and not bias. An interpretation of bias would require the judgment that the secondary dimensions leading to group differences are irrelevant to the test construct.

To ensure that the items included in the test have the smallest DIF possible, most test developers and testing organizations evaluate DIF at the pretest stage. Bridgeman and Schmitt (1997) suggested that DIF analyses may be conducted after the pretest, before score reporting, and after score reporting. Penfield and Camilli (2007) presented a 6-step procedure for DIF analyses to conduct a more comprehensive and reliable DIF analyses.

Summary

Test scores are widely used as criteria for decisions regarding placement, promotion, and licensure. Because MC tests are prevalent in assessment programs, there is a concern that systematic error due to guessing can lead to incorrect interpretations of examinee proficiency or bias statistical estimates from secondary analyses of test information (e.g., DIF). The measurement error involved is different from random error which pushes observed scores up or down randomly; guessing behavior can result in consistently higher observed scores and inflated test variance. Therefore corrections for guessing, applied via scoring methods, have the potential to enhance interpretations of test scores.

Although modern test theory has more flexibility in predicting examinees' performance, a more sophisticated understanding of how guessing affects proficiency estimation in 3PL IRT models is yet to be developed. Furthermore, because guessing represents a systematic error, it could result in statistical bias in analyses using observed total score. In particular, DIF procedures such as the MH procedure and the LR procedure depend on the accuracy of observed total score (as the matching variable). If the effects of guessing behavior are more likely in one group (focal or reference), then the observed total score is less useful as a matching variable. Therefore, the development of IRT-based corrections for observed scores may potentially be useful in observed-score DIF analysis.

CHAPTER III. METHODOLOGY

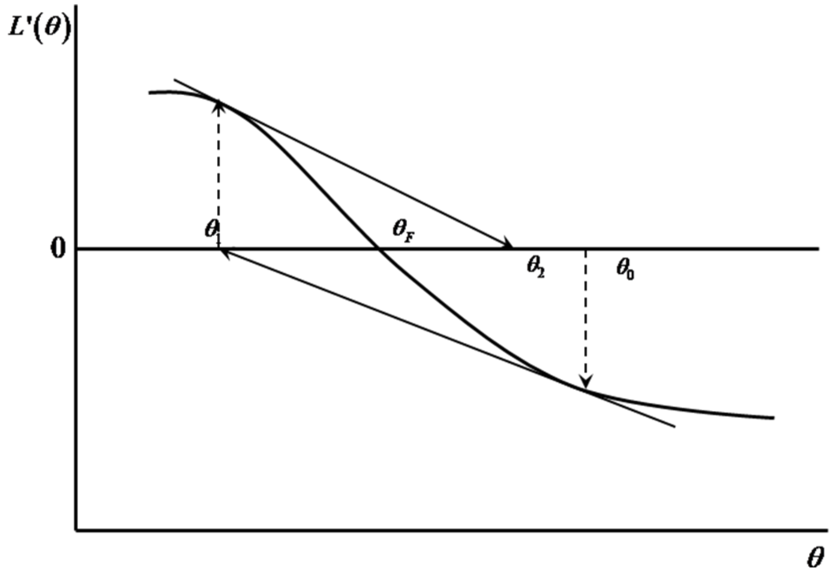
In this chapter, research questions and assumptions are addressed. Then a comprehensive conceptual and statistical framework on different correction for guessing methods is presented. Formula scores were described based on the CTT perspective, followed by the IRT 3PL model. Next, two new methods motivating the uses of the 3PL IRT model are derived. Two simulation studies are then conducted. In the first, the accuracy of the IRT formula scores is assessed. In the second, the MH and LR DIF procedures are carried out matching on the number-correct score and alternatively matching on the IRT formula scores. The results are then compared in terms of type 1 errors and bias.

Research Questions and Assumptions

To date, IRT models for MC items have been developed that model the probability of an examinee answering an item correctly. To model the effects of guessing, a fixed lower-asymptote parameter can be added to the 1PL or 2PL IRT models, or the full 3PL model can be chosen. Although IRT has been used to estimate ability, number-correct scores are more prevalent in operational psychometric data processing. In part, the goal of this dissertation was to develop a new correction-for-guessing based on the 3PL IRT model with practical application to DIF analysis and other analyses based on number-correct scores.

In IRT, maximum-likelihood estimation (MLE) is a procedure used to estimate the ability (*θ*) levels of examinees as well as item parameters. Finding *θ* requires maximizing the likelihood (or log likelihood) of an examinee's item response pattern with respect to a set of fixed item parameters (Embretson & Reise, 2000). The Newton-Raphson procedure is a common iterative procedure used for MLE. The algorithm is applied to find the mode of an examinee's proficiency likelihood function. It requires the first and second derivatives of the log-likelihood function to update *θ* estimates iteratively. The logic of the Newton- Raphson procedure is illustrated below in Figure 3-1.

Figure 3-1. Illustration of the Logic of Newton-Raphson Procedure



In Figure 3-1, the first derivative of the log-likelihood function of *θ* is graphed against ability (θ) level. The starting value, θ_0 in this case, is a guess of an examinee's possible trait level. The projected second derivative then gives the updated θ_1 estimate, and in turn, θ_1 leads to θ_2. The iterations end when the second derivative is zero (Embretson & Reise, 2000; Veerkamp, 2000). One basic method of this dissertation is to derive an expression of the true score when the second derivative of the 3PL log-likelihood is zero.

The MLE provides an unbiased estimate of θ; however, it has some problems. The major problem is that with MLE, no θ can be obtained for perfect or zero score (Embretson & Reise, 2000). The other alternative to estimate θ, the expected a posteriori (EAP) estimation, offers finite θ estimation for perfect score or for the patterns with all incorrect responses. In EAP, information from the examinees' response pattern and information about the population are combined. The EAP is a Bayesian estimator from the mean of the posterior distribution (Embretson & Reise, 2000). One drawback on EAP estimation is that an estimate of θ is regressed toward the mean of the prior distribution unless the number of items is relatively large (Meijer & Nering, 1999; Embretson & Reise, 2000).

In this dissertation, the essential approach to understanding the effects of c parameters was to 1) approximate the log-likelihood function as a Taylor series expansion around a guessing parameter c, and 2) examining the implications of the model when the approximate likelihood is maximized. This provided the link between the 3PL IRT model proficiency estimate and a corrected-for-chance observed score. One main goal of this study was to understand the effect of guessing within the IRT framework.

The second purpose of this dissertation was to develop two IRT formula scores based upon using the 3PL model. Though ideally undesirable effects of guessing should be prevented, the IRT formula scores provided a post hoc statistical correction that is not a function of the number-correct score. These IRT formula scores conceptually illustrated the mechanism by which the 3PL IRT model adjusts for guessing, and provided estimates of proficiency that may improve analyses traditionally carried out with number-correct scores. In the next section, different scoring methods were detailed and discussed from a mathematical point of view.

Scoring Methods

Formula Score based on CTT

The most widely used method is the formula scoring method. For a test of n items, the number of correct responses (R) for an examinee may be expressed as

$$R = C_K + C_G, \tag{3.1}$$

where C_K and C_G represent the number of correct responses with knowledge and the number of correct responses by guessing, respectively. To determine the number of correct responses with knowledge, Equation (3.1) can be re-written as

$$C_K = R - C_G \,. \tag{3.2}$$

Assuming no omitted items, the expected number of items which an examinee answers by guessing (n_G) is the difference between the total number of items and the number of correct responses. This can be represented as

$$n_G = n - C_K = R + W - C_K \,, \tag{3.3}$$

where W is the number of incorrect responses. The highest number of correct responses, based on random guessing, with k options per item is

$$C_G = k^{-1} n_G = k^{-1}\left(R + W - C_K\right), \tag{3.4}$$

therefore, substituting the right-hand side of Equation (3.4) for C_G in Equation (3.2) results in

$$\begin{aligned} C_K &= R - k^{-1}\left(R + W - C_K\right) \\ &= R - (k-1)^{-1} W \,. \end{aligned} \tag{3.5}$$

This correction method penalizes examinees for guessing by subtracting partial points from the number-right score based on the number of incorrect responses.

IRT 3PL Model

In a 3PL IRT model, the probability λ_i for the examinee with a certain ability level (θ) to answer a particular item right can be represented as

$$\lambda_i\left(u_i = 1 \middle| \theta, a_i, b_i, c_i\right) = c_i + \left(1 - c_i\right) P_i , \tag{3.6}$$

where

$$P_i = \frac{\exp\left[Da_i\left(\theta - b_i\right)\right]}{1 + \exp\left[Da_i\left(\theta - b_i\right)\right]} . \tag{3.7}$$

a_i, b_i, c_i, and D are indices of item discrimination, item difficulty, pseudo-chance-level (guessing parameter), and a scaling constant $D = 1.7$, respectively. Let u_i represent the scored response (0 or 1) of an examinee to item i. The number-correct score R can then be given as

$$R = \sum_{i=1}^{n} u_i , \tag{3.8}$$

and the number-incorrect score W as

$$W = n - R = \sum_{i=1}^{n} \left(1 - u_i\right) . \tag{3.9}$$

Note that $n = R + W$ if no items are omitted. Assuming a common c parameter for all items (i.e., $c_i = c$ for all i), the true-score formula can be expressed as

$$\begin{aligned} T &= \sum_{i=1}^{n} \lambda_i \\ &= \sum_{i=1}^{n} \left[c + (1-c) P_i \right]. \\ &= nc + (1-c) \sum_{i=1}^{n} P_i \end{aligned} \tag{3.10}$$

If the IRT 3PL model fits the item responses well, then T should provide a good approximation of R; that is T can be thought of as $E[R]$. The corrected true score can be represented as

$$C_T = \sum_{i=1}^{n} P_i \,, \tag{3.11}$$

and this defines the probability for an examinee to answer the item correctly based on item difficulty and item discrimination, but not on guessing. Assuming that $n = R + W$, it is straightforward to show

$$\begin{aligned} C_T &= \sum_{i=1}^{n} P_i = \frac{T - nc}{1-c} \\ &= \frac{T - \left[T + (n-T) \right] c}{1-c} \\ &= \frac{(1-c)T - c(n-T)}{1-c} \\ &= E\left[R_C\right] - \frac{c}{1-c} E\left[W\right] \end{aligned} . \tag{3.12}$$

Using the substitution,

$$\frac{c}{1-c} = \frac{1}{k-1}, \tag{3.13}$$

Equation (3.12) is parallel to Equation (3.5), and thus the IRT score C_T appears to bear a strong similarity to the classical formula score C_K. However, as shown in the next section, this impression is incomplete because of the derivation of C_T above does not take into account an examinee's item response pattern.

IRT-Based Methods for Guessing Corrections

In this dissertation, the IRT formula scores, in contrast to the traditional method as the simple analogy in Equation (3.12), took into account the *pattern* of item responses, and resulted in a score that is not a linear function of the number-correct score. Thus, while the traditional method had its greatest impact by preventing guessing, the newly proposed methods had some potential to provide a statistical post-testing correction.

First IRT approach (*formula*). In IRT, the probability of an examinee answering an item correctly depends on the examinee's ability and item discrimination and difficulty (Hambleton et al., 1991). For most MC tests, examinees with very low abilities have probabilities greater than zero of answering even the most difficult items. The 3PL model (Birnbaum, 1968) adds the pseudo-chance parameter (to discrimination and difficulty) to remove the effect of random guessing. Given the IRT framework, ΣP_i, as given in Equation (3.11), represents an examinee's *corrected true score*, C_T, which can be conceptualized as the true score obtained when the effects of guessing are eliminated. The IRT formula score is based on a simplification of a common approach for estimating examinee proficiency. For a *n*-item MC test, the log likelihood of a response pattern for an examinee is given by

$$\begin{aligned} F(c) &= \ln \prod_{i=1}^{n} \lambda_i^{u_i} (1-\lambda_i)^{1-u_i} \\ &= \sum_{i=1}^{n} \left[u_i \ln \lambda_i + (1-u_i) \ln(1-\lambda_i) \right], \end{aligned} \tag{3.14}$$

with u_i =1 or 0 for a correct or incorrect response, respectively. An estimated proficiency is obtained by maximizing this function with respect to θ. To derive the first IRT formula score, the log likelihood function is approximated as a one-term Taylor series at the common guessing parameter c, and maximized with respect to θ. Upon simplification, an estimate of C_K is obtained as well as a broader perspective on the estimated θ.

The standard Taylor one-term power expansion is obtained by

$$H(c) = F(0) + F_{c=0}^{(1)} \cdot c\,. \tag{3.15}$$

Let

$$\begin{aligned} \frac{\partial \lambda_i}{\partial c} &= \frac{\partial}{\partial c}\left[c + (1-c) P_i \right] \\ &= 1 - P_i = Q_i\ . \end{aligned} \tag{3.16}$$

It follows that the first derivative $F_c^{(1)}$ is

$$\begin{aligned} F_c^{(1)} &= \sum_{i=1}^{n} \left[\frac{u_i}{\lambda_i} \frac{\partial \lambda_i}{\partial c} + \frac{(1-u_i)}{(1-\lambda_i)} \frac{\partial (1-\lambda_i)}{\partial c} \right] \\ &= \sum_{i=1}^{n} Q_i \left[\frac{u_i}{\lambda_i} - \frac{(1-u_i)}{(1-\lambda_i)} \right]. \end{aligned} \tag{3.17}$$

At c = 0,

$$\begin{aligned} F_{c=0}^{(1)} &= \sum_{i=1}^{n}\left[\frac{u_i}{P_i}Q_i - \frac{(1-u_i)}{Q_i}Q_i\right] \\ &= \left(\sum_{i=1}^{n} u_i \frac{Q_i}{P_i}\right) - W. \end{aligned} \tag{3.18}$$

Then one-term expansion of $F(c)$ at *c = 0* is given by

$$H(c) = F(0) + \left[\sum_{i=1}^{n}\left(u_i \frac{Q_i}{P_i}\right) - W\right]c \tag{3.19}$$

Next, to maximize $H(c)$ with respect to *θ*, differentiate $F_{c=0}^{(1)}$ with respect to *θ* which yields

$$\begin{aligned} \frac{\partial}{\partial\theta}\left[\sum_{i=1}^{n}\left(u_i \frac{Q_i}{P_i}\right) - W\right] &= \sum_{i=1}^{n} u_i \frac{\partial}{\partial\theta}\left(\frac{Q_i}{P_i}\right) \\ &= -D\sum_{i=1}^{n} a_i u_i \frac{Q_i}{P_i}. \end{aligned} \tag{3.20}$$

Then set the result equal to zero

$$\frac{\partial}{\partial\theta}H(c) = \frac{\partial}{\partial\theta}\left[F(0) + F_{c=0}^{(1)} \cdot c\right] = 0 \ , \tag{3.21}$$

which results in

$$\sum_{i=1}^{n} a_i(u_i - P_i) - c\sum_{i=1}^{n} a_i u_i \frac{Q_i}{P_i} = 0\,. \tag{3.22}$$

Then setting $a_i = 1$ gives the solution for the first IRT formula score C_{T1}

$$C_{T1} = \sum_{i=1}^{n} P_i \approx R - \sum_{i=1}^{n} u_i \eta_i \ , \tag{3.23}$$

where

$$\eta_i = c\frac{Q_i}{P_i}\,. \tag{3.24}$$

Equal *a* parameters was a big assumption, but again, this assumption is also implicit in the classical formula and number correct scoring. To interpret η_i, consider a correct response to an item with 5 options and $c = 0.2$ (a random guess). In this scenario, if the item is very difficult, the probability of answering incorrectly is greater than the probability of providing a correct response. In this case, the potential impact of guessing is higher than it would be for an easier item or a higher ability examinee. To reduce the positive bias introduced by guess, the correct response is adjusted downward by the factor η_i. In intuitive terms, the IRT 3PL model does not "believe" that low-ability examinees should be able to answer difficult items. When such a correct response is encountered, the model treats this as a probable guess and adjusts downward. With regard to examinee proficiency, scores for examinees with lower proficiency levels would be adjusted more when compared to those with higher proficiency levels. So $u_i(1- \eta_i)$ characterizes an item response adjusted downward on the basis of examinee proficiency. This demonstrates the kind of implicit correction employed in IRT 3PL estimates of proficiency. To simplify this result further, assumed that $a_i = 1$. A measure of true score adjusted for c can then be obtained with

$$C_{T1} \approx \sum_{i=1}^{n} u_i \left(1 - \eta_i\right) = R - \sum_{i=1}^{n} u_i \eta_i \,. \tag{3.25}$$

One major goal in this dissertation is to apply the approximation (3.25) with number-correct scores. For this purpose, two different approaches are used to obtain estimates of η_i based on observed-score statistics.

Pseudo-Bayes Probability. The first method of estimating η_i is motivated by Bayes Theorem. To obtain values for these parameters, a random guessing c was assumed, and thus $c = 1/k$. To estimate P_i, that does not require IRT parameter estimates, one option for obtaining a value for P_i is to use the overall sample average p-value for item i, say p_i. However, this is not adequate because the essence of the new method calls for sensitivity to whether a *particular* examinee is expected to answer a question correctly. Likewise, the overall proportion correct for an examinee, say r, is not sensitive to whether an examinee has a higher propensity to answer correctly for some items than others. A solution can be motivated by an analogical application of Bayes rule which combines the estimates of the proportions p_i and r.

Define u_{ji} as the 0-1 response of examinee j on item i, and α_j as the response of a randomly selected item belonging to examinee j. Define

$$\begin{aligned} &P(\alpha_j = 1) = r_j \\ &P(\alpha_j = 0) = 1 - r_j = w_j \ . \end{aligned} \tag{3.26}$$

Now let the expected probability for examinee j to get item i correct given α_j be

$$\begin{aligned} &P\left(u_{ji} = 1 \mid \alpha_j = 1\right) = p_i \\ &P\left(u_{ji} = 1 \mid \alpha_j = 0\right) = q_i \, . \end{aligned} \tag{3.27}$$

The explanation of Equation (3.27) is as follows. Suppose a randomly select response for examinee *j* is $\alpha_j = 1$. Knowing nothing else, it could then be guessed that examinee *j* probably got u_{ji} correct. A reasonable choice for this conditional probability is p_i, which is the *p*-value for *i*. However, if $\alpha_j = 0$, then one would guess a lower probability for a correct response. A reasonable choice for the conditional probability in this case is $q_i = 1 - p_i$. While these choices are informal, they are consistent with intuitive expectations.

The purpose of the randomly-sampled-item idea is to motivate the situation in which there is prior information on an examinee acquired from a set of item responses. This information is then modified by an item's difficulty to produce an updated estimate of the examinee's performance on a test item. The procedure can be accomplished with Bayes Theorem as follows:

$$P\left(\alpha_j = 1 \mid u_{ji} = 1\right) = \frac{P\left(u_{ji} = 1 \mid \alpha_j = 1\right) P\left(\alpha_j = 1\right)}{P\left(u_{ji} = 1 \mid \alpha_j = 1\right) P\left(\alpha_j = 1\right) + P\left(u_{ji} = 1 \mid \alpha_j = 0\right) P\left(\alpha_j = 0\right)}. \qquad (3.28)$$

Substituting Equations (3.26) and (3.27) into (3.28) gives the updated probability for examinee *j* for a correct response:

$$P\left(\alpha_j = 1 \mid u_{ji} = 1\right) = \frac{p_i r_j}{p_i r_j + q_i w_j}. \qquad (3.29)$$

Note that for each item *i*, a different updated probability for a correct response is obtained.

Correcting the probabilities in (3.26) and (3.27) for guessing results in

$$\begin{aligned} r' &= \frac{r-c}{1-c} \quad , \; w' = 1 - r' \\ p_i' &= \frac{p_i - c}{1-c} \; , \; q_i' = 1 - p_i' \; . \end{aligned} \tag{3.30}$$

The *posterior probability* of an examinee's success (the *j* subscript is dropped below for ease of presentation) on an item say $\hat{P}_i$, can then be obtained by combining the examinee's prior information with the probability of success on the item as:

$$\begin{aligned} \hat{P}_i &= \frac{r'p_i'}{r'p_i' + w'q_i'} \\ &= \frac{(r-c)(p_i-c)}{(r-c)(p_i-c) + wq_i} \qquad && r, p_i > c \\ \hat{P}_i &= 0 && \text{otherwise ,} \end{aligned} \tag{3.31}$$

and

$$\hat{Q}_i = 1 - \hat{P}_i . \tag{3.32}$$

These estimates were referred to as pseudo-Bayes item probabilities. The one-term Bayes formula score was then obtained as

$$\begin{aligned} C_{T1B} &= R - \sum_{i=1}^{n} u_i \hat{\eta}_i \\ &= R - c\sum_{i=1}^{n} u_i \frac{wq_i}{(r-c)(p_i-c)} \end{aligned} \tag{3.33}$$

where

$$\hat{\eta}_i = c\frac{\hat{Q}_i}{\hat{P}_i} . \tag{3.34}$$

It should be clear that no correction is applied when $c = 0$.

Conditional Probability. Instead of simply using the overall sample average p-value p_i to estimate $P(u_{ji} = 1)$, in the second approach for estimating η_i, the sample average p-value for item i was conditioned on R. For a n-item test with J examinees, let u_{ji} be defined as the 0-1 response of examinee j on item i; and let u_i be defined as the response of a randomly selected examinee on item i. To estimate the probability of a correct response from examinee j on item i, the expected value of randomly selected with $r_j=R$ can be taken. Assuming a Rasch model, this estimate incorporates all sample information concerning performance on item i (based on the principle of sufficiency). The required probability $P\left(u_{ji} = 1 \mid R\right)$ for an examinee is then obtained as the expected value

$$E_i\left[P\left(u_{ji} = 1 \mid R\right)\right] \approx \frac{\sum\limits_{r_j=R} u_{ji}}{N_{r_j=R}} = \hat{P}_i^*, \tag{3.35}$$

and

$$\hat{Q}_i^* = 1 - \hat{P}_i^*. \tag{3.36}$$

These estimates were referred to as *conditional* item probabilities, the one-term probability formula score is obtained as

$$\begin{aligned} C_{T1P} &= R - c\sum_{i=1}^{n} u_i \frac{\hat{Q}_i^*}{\hat{P}_i^*} \\ &= R - \sum_{i=1}^{n} u_i \hat{\eta}_i^* \,. \end{aligned} \tag{3.37}$$

Another issue was to set the maximum correction factor value η_i. In terms of practical significance, a reasonable maximum amount for guessing correction should be less than 1 point, that is, the amount of credit given for a correct answer. The value of the correction factor was set to be restricted to the interval [0, 1], that is $0 \le \eta_i \le 1$. This implies that $\frac{Q_i}{P_i} \le c^{-1}$, or alternatively, $c \ge \frac{P_i}{Q_i}$.

Comparison between classical formula score and the first IRT formula. The significance of this approach was that an individual's item response pattern is taken into account to provide a score adjustment. For a correct answer, the adjustment requires subtraction of the term η_i from the full point of item credit, and no correction is made when guessing is not present. Although this seems to be very different from the standard logic of classical formula scoring of subtracting partial points on incorrect items, the classical formula shown in Equation (3.5) could be also re-expressed as a sum over attempted items as

$$C_K = \frac{k}{k-1}\sum_{R,W}\left(u_i - \frac{1}{k}\right). \tag{3.38}$$

The classical formula score from this perspective down-weights all correct responses equally, whereas the IRT formula down-weights a correct response proportionally based on the ratio of an examinee’s odds of answering that item incorrectly to the odds of answering the item correctly.

Second IRT approach (*formula*). The approximation above is based on a one-term power expansion of the log likelihood function around a common *c* parameter. An alternative approach is based on factoring the 3PL probability given in Equation (3.6) with η_{i}, therefore, the 3PL IRT probability λ_i can be expressed as

$$\lambda_i = P_i\left(1+\eta_i\right), \tag{3.39}$$

where P_i is the 2PL model and

$$\begin{aligned}\eta_i &= c\frac{Q_i}{P_i} \\ &= c\exp\left[-Da_i\left(\theta-b_i\right)\right]\end{aligned}. \tag{3.40}$$

The log likelihood function can then be written as

$$F\left(c\right)=\sum_{i=1}^{n}u_i\ln\left[P_i\left(1+\eta_i\right)\right]+\left(1-u_i\right)\ln\left[1-P_i\left(1+\eta_i\right)\right]. \tag{3.41}$$

Differentiating *F* with respect to θ, setting the result equal to zero, and simplifying gives

$$\sum_{i=1}^{n}a_iP_i=\sum_{i=1}^{n}a_iu_i\left(1+\eta_i\right)^{-1}. \tag{3.42}$$

Assuming $a_i = 1$, the resulting estimator of or formula for C_T is

$$C_{T2}=\sum_{i=1}^{n}u_i\left(1+\eta_i\right)^{-1}. \tag{3.43}$$

This approach gives a result identical to a *M*-term Taylor expansion of the likelihood function as shown in Appendix A. The M-term Bays formula score, C_{T2B}, using the Bayes' theorem in Equation (3.31) and the assumption $c_i = c$, can then be obtained as

$$\begin{aligned} C_{T2B} &= \sum_{i=1}^{n} u_i \left(1+\hat{\eta}_i\right)^{-1} \\ &= \sum_{i=1}^{n} u_i \left(1+c\frac{wq_i}{(r-c)(p_i-c)}\right)^{-1}. \end{aligned} \tag{3.44}$$

The M-term probability formula score, C_{T2P}, using the conditional probability in Equation (3.35), can be obtained as

$$\begin{aligned} C_{T2P} &= \sum_{i=1}^{n} u_i \left(1+\hat{\eta}_i^{*}\right)^{-1} \\ &= \sum_{i=1}^{n} u_i \left(1+c\frac{\hat{Q}_i^{*}}{\hat{P}_i^{*}}\right)^{-1}. \end{aligned} \tag{3.45}$$

Note that if $c = 0$, then no correction is made and $C_{T2B} = C_{T2P} = R$. Unlike the one-term correction, no bounds are required on η_i with the M-term correction.

Evaluation of two proposed corrected scores. The two IRT formulas described above can be used for obtaining sample estimates of corrected true scores, but it is important to ensure both IRT formula scores are unbiased estimate of the corrected true scores. As an estimate of corrected true score C_T, C_{T1} and C_{T2} are unbiased if the expected values of C_{T1} and C_{T2}, $E[C_{T1}] = E[C_{T2}] = C_T = \sum_{i=1}^{n} P_i$. The expected value of C_{T1} is equal to

$$\begin{aligned} E\left[C_{T1}\right] &= E\left[\sum_{i=1}^{n} u_i\left(1-\eta_i\right)\right] = \sum_{i=1}^{n} E\left[u_i\right]\left(1-\eta_i\right) \\ &= \sum_{i=1}^{n}\left(P_i + c\left(1-P_i\right)\right)\left(\frac{P_i - c\left(1-P_i\right)}{P_i}\right) \\ &= \sum_{i=1}^{n} P_i\left(\frac{P_i + cQ_i}{P_i}\right)\left(\frac{P_i - cQ_i}{P_i}\right) \\ &= \sum_{i=1}^{n} P_i\left(1+\eta_i\right)\left(1-\eta_i\right) = \sum_{i=1}^{n} P_i\left(1-\eta_i^{2}\right) \\ &\neq C_T \end{aligned} \quad . \tag{3.46}$$

And the expected value of C_{T2} is

$$\begin{aligned} E\left[C_{T2}\right] &= E\left[\sum_{i=1}^{n} u_i\left(1+\eta_i\right)^{-1}\right] = \sum_{i=1}^{n} E\left[u_i\right]\left(1+\eta_i\right)^{-1} \\ &= \sum_{i=1}^{n}\left(c+\left(1-c\right)P_i\right)\left(1+c\frac{Q_i}{P_i}\right)^{-1} \\ &= \sum_{i=1}^{n}\left(c+\left(1-c\right)P_i\right)\left(1+\frac{c\left(1-P_i\right)}{P_i}\right)^{-1} \\ &= \sum_{i=1}^{n}\left(c+\left(1-c\right)P_i\right)\left(\frac{P_i}{c+\left(1-c\right)P_i}\right) \\ &= \sum_{i=1}^{n} P_i = C_T \end{aligned} \quad . \tag{3.47}$$

Clearly, C_{T1} was not an unbiased estimate because the expected value of C_{T1} was smaller than C_T and negative bias exists. In fact, as shown below, C_{T1} was useful for conceptually understanding the effects of guessing. In addition, C_{T1} equals C_{T2} when C_{T1} is rescaled to C_T by dividing by $(1-\eta_i^2)$. But C_{T2} is an unbiased estimate of the true score and is expected to have more accurate estimation on the true scores.

Baseline correction

The *posterior probability* proposed in Equation (3.31) provides an easy and practical approach to score correction without using IRT response-pattern complexities. Therefore, a simple scoring formula can be obtained as

$$B = \sum_{i=1}^{n} \hat{P}_i \,, \tag{3.48}$$

That is, B is the sum of the posterior probabilities. In this study, index B is used as a baseline criterion for evaluating the other two more elaborate IRT formula scores. That was, for an IRT formula score to be considered useful, it must show less bias and a higher correlation with the corrected true score C_T than the index B.

Study I: Comparisons of Scoring Methods

To evaluate the two IRT formulas (include two one-term formula scores and two M-term formula scores), three simulation studies were designed using the IRT 3PL model to generate data with two sets of item parameters. Examinee abilities θ were generated from the random normal distribution N (0, 1) for all simulations.

Data Generation

Item parameters: Set I. In the first set of item parameters, a 33-item test (labeled SIM hereafter) was generated. In Table 3-1, the item discrimination parameters in (a_i) had three levels (a = 0.5, 1.0, and 1.5) and these three levels were crossed with 11 levels of item difficulty (b_i = -2.5 to 2.5 in steps of 0.5). All guessing parameters (c_i) were fixed at 0.2, consistent with random guessing on MC items having five options.

Table 3-1

Item Parameters: Set I (SIM) All items have c = 0.2

Item	*a*	*b*	**Item**	*a*	*b*	**Item**	*a*	*b*
1	0.5	-2.5	**12**	1.5	-1.0	**23**	1.0	1.0
2	1.0	-2.5	**13**	0.5	-0.5	**24**	1.5	1.0
3	1.5	-2.5	**14**	1.0	-0.5	**25**	0.5	1.5
4	0.5	-2.0	**15**	1.5	-0.5	**26**	1.0	1.5
5	1.0	-2.0	**16**	0.5	0.0	**27**	1.5	1.5
6	1.5	-2.0	**17**	1.0	0.0	**28**	0.5	2.0
7	0.5	-1.5	**18**	1.5	0.0	**29**	1.0	2.0
8	1.0	-1.5	**19**	0.5	0.5	**30**	1.5	2.0
9	1.5	-1.5	**20**	1.0	0.5	**31**	0.5	2.5
10	0.5	-1.0	**21**	1.5	0.5	**32**	1.0	2.5
11	1.0	-1.0	**22**	0.5	1.0	**33**	1.5	2.5

Item parameters: Set II. The second set of item parameter values was obtained from the Abstract Reasoning Test (ART; Embretson, 1998). The test had 30 items and was designed to measure general intelligence. Item parameters were estimated from data from an administration to 787 young adults. Table 3-2 presents the IRT 3PL item parameter estimates. The result from this simulation is used to examine how the IRT formula scores work with data from an existing test.

Table 3-2

Item parameters: Set II (ART)

Item	*a*	*b*	*c*	**Item**	*a*	*b*	*c*
1	1.286	-2.807	0.192	**16**	1.150	-0.882	0.204
2	1.203	0.136	0.162	**17**	0.846	1.303	0.112
3	0.814	-2.033	0.196	**18**	0.986	1.090	0.113
4	0.941	-0.557	0.142	**19**	1.295	0.597	0.115
5	1.083	-1.461	0.153	**20**	1.065	-0.017	0.110
6	0.752	-1.979	0.182	**21**	0.948	0.470	0.095
7	1.363	-1.785	0.146	**22**	1.150	2.609	0.170
8	1.083	-0.776	0.118	**23**	0.928	-0.110	0.155
9	1.149	-0.239	0.214	**24**	0.934	1.957	0.103
10	1.837	-1.247	0.132	**25**	0.728	3.461	0.128
11	1.269	-0.917	0.153	**26**	1.452	1.144	0.107
12	0.783	0.819	0.129	**27**	0.460	-0.799	0.226
13	1.501	-0.963	0.196	**28**	0.609	-1.018	0.192
14	1.417	0.526	0.118	**29**	0.779	1.291	0.142
15	0.949	0.577	0.126	**30**	0.576	1.607	0.178

First Simulation

Using the item parameters in Tables 3-1 and 3-2, a single sample of size N=100000 (separately for each set of parameters) was generated using SAS 9.1 computer software package (SAS Institute, 2003) to study the asymptotic behavior of the various corrections.

The estimates R, C_K, C_{T1B}, C_{T1P}, B, C_{T2B} and C_{T2P} (number-correct score, classical formula score, one-term Bayes formula score, one-term probability formula score, baseline correction, M-term Bayes formula score and M-term probability formula score, respectively) were obtained and compared to the corrected true score C_T. A fixed c (used in score adjustment) was set as the average of the c parameters (0.2 for Set I, and 0.15 for Set II).

Second Simulation

In order to study the new formula scores in moderate-sized samples, another sample n= 5000 for each test was sampled from the data sets generated above with 10 replications. Calibrations of items were conducted with Parscale using the 3PL IRT model (Muraki & Bock, 2003). Examinees' θ*s* were also estimated using a Bayesian expected a posteriori (EAP) method. The IRT estimate of corrected true score C_T (labeled $\hat{C}_T$) was then obtained by substituting sample estimates of item parameters and proficiencies into Equation(3.11). C_T was used as a standard for evaluating corrected scores, although in samples with n=5000, it may be the case that $\hat{C}_T$ provides a better standard because it preserved more information about the true score. However, the issue here is that the estimation error exists in $\hat{C}_T$, and with EAP estimation used to estimate θ, the resulting θ would regress to *zero*. Therefore, $\hat{C}_T$ is not an unbiased estimate of C_T in given neighborhoods of θ. For that reason, the comparison between $\hat{C}_T$ and C_T was obtained to see how well $\hat{C}_T$ explains C_T. The comparison between the corrected scores and C_T is used as a pragmatic criterion to evaluate the reliability of scores, and also how well the corrected score estimates performed. Corrected score estimates were then obtained in two different ways:

1. Corrected scores were obtained by plugging estimated IRT item parameters and estimated theta into Equations (3.25) and (3.43) to get $\hat{C}_{T1}$ and $\hat{C}_{T2}$.
2. Corrected scores were obtained by calculating the *Bayes* formula scores and probability formula scores from the sample observations by using Equation (3.33), (3.44), (3.37), and (3.45) to obtain the formula scores C_{T1B}, C_{T1P}, C_{T2B} and C_{T2P}. For each of these 10 replications, C_{T1B}, C_{T1P}, C_{T2B} and C_{T2P} was computed and compared to their respective values of $\hat{C}_{T1}$ and $\hat{C}_{T2}$.

The purpose here was to evaluate potential information loss due to the Taylor approximation, and the use of pseudo-Bayes estimates and conditional probabilities instead of estimated IRT item probabilities. The quantities $\hat{C}_{T1}$ and $\hat{C}_{T2}$ were the IRT model-based versions of C_{T1} and C_{T2}. They can be thought of as the providing an upper limit to the performance of formula-score estimates of C_{T1B}, C_{T1P}, C_{T2B} and C_{T2P}.

Criteria for Evaluating the Two New IRT Scoring Models

Previous studies have focused on overall comparisons either between examinees' observed scores and formula scores, or between examinees' true scores (based on an IRT model) and formula scores. To find out if the IRT formula scores improved estimates of ability level, examinees were stratified in quartiles based on the known corrected true score, C_T. Analyses in this analysis were carried out separately, by quartile ($Q1 - Q4$). Because corrections made by the formula score, C_K, could result in negative values, all negative values were set to 0.

First simulation. In first simulation, bias and percent of variance accounted for (r^2) were used to evaluate different correction methods for two sets of tests. The bias statistic was computed over examinees, *j*, as

$$Bias = \frac{1}{J}\sum_{j=1}^{J}\left(S_j - C_{Tj}\right), \tag{3.49}$$

where S_j represents the given proficiency estimate (R, C_K, C_{T1B}, C_{T1P}, B, C_{T2B}, or C_{T2P}) for examinee *j*; and C_{Tj} represents the corrected true score for examinee *j*, C_T. The criterion of primary interest was the predictive accuracy of the different scoring model, and this was assessed by obtaining the correlation between the different corrected scores (R, C_K, C_{T1B}, C_{T1P}, B, C_{T2B}, and C_{T2P}) and the corrected true score, C_T. Scoring methods that resulted in lower bias and higher r^2 were considered preferable.

Second simulation. Bias, root mean square error (RMSE), and the correlations were calculated over 10 replications for $\hat{C}_T$, the approximations $\hat{C}_{T1}$ and $\hat{C}_{T2}$ relative to C_T. The RMSE statistics were computed as:

$$RSME = \sqrt{\frac{\sum_{j=1}^{J}\left(\hat{\omega}_j - C_{Tj}\right)^2}{J}}, \tag{3.50}$$

where $\hat{\omega}_j$ equals the IRT estimate of corrected true score $\hat{C}_T$, and IRT estimate of corrected score, $\hat{C}_{T1}$ and $\hat{C}_{T2}$. The corresponding bias statistic and the root mean square errors (RMSE) and the correlation coefficient r of C_{T1B}, C_{T1P}, C_{T2B} and C_{T2P} with C_T were calculated over 10 replications.

Study II: Application to DIF Analyses

The third goal of this dissertation was to demonstrate a potential application of IRT formula scores on DIF analyses. To evaluate how the IRT formula scores performed on a DIF analysis, LR and the MH procedures were applied. This study had two goals: a) to study the effect of different scoring methods on the type 1 error estimation of the DIF procedure, and b) to compare the LR and MH procedures with regard to detection of DIF.

Data Generation

Different factors which are likely to affect the type 1 error of DIF analysis were manipulated, including item parameters, sample size, and ability. In typical DIF there are two groups of examinees (reference group and focal group) and this provides a choice of using either group percent correct for an item (p_R, p_F) or the correct percent across all examinees (p_{R+F}) to capture the *observed p*-value for an item for the purpose of estimating η_i. Based on the pilot work in which (p_R, p_F) was used, large biases in type 1 error rates and LORs were found. Therefore, the correct percentage for an item from the total sample (p_{R+F}) is used to estimate η_i.

Item parameters. Examinee response data were generated using the 3PL IRT model, based on the two sets of item parameters described in the previous study, using SAS 9.1 computer software package (SAS Institute, 2003).

Sample size. Numerous studies indicate that sample sizes of focal and reference groups appear to have an effect on type 1 error (Rogers & Swaminathan, 1993; Roussos & Stout, 1996b; Tian, 1999). In addition, when gender difference is the target, approximately equal focal and reference group sample sizes are reasonable; when the comparison is between majority and minority subjects, unequal sample sizes for both groups are more realistic. Therefore, two different sample size conditions were investigated in this study: 1) equal sample size for focal group and reference group (N_F=N_R=1000), and 2) unequal sample size (N_F=500, N_R=1000).

Ability distribution. A few researchers suggest that large differences in the ability distribution of two groups could result in high type 1 error (Tian, 1999). However, some researchers endorse the opposite conclusion and suggest that ability distribution differences do not significantly affect type 1 error rates unless the ability distribution difference between the two groups is greater than 1 SD (Narayanan & Swaminathan, 1994). Because ability distribution differences between the reference and focal groups usually exist, three conditions are considered in this study:

1. Equal ability distributions: both reference and focal group are N (0, 1).
2. Unequal ability distributions: N (0, 1) for the reference group and N (-0.5, 1) for the focal group.
3. Unequal ability distributions: N (0, 1) for the reference group and N (-1, 1) for the focal group.

Procedure

The MH and LR procedures were studied under various conditions for obtaining matching scores: number-correct score (R), first IRT formulas (C_{T1B} and C_{T1P}), and second IRT formulas (C_{T2B} and C_{T2P}). For each condition, performance over 1000 replications per condition was evaluated. The matching scores were rounded off to integers for the MH procedure. Results for the LR procedure were obtained with SAS Logistic procedure under SAS 9.1. The MH procedure was also performed using SAS 9.1.

Type 1 errors and average log-odds ratio were obtained for both procedures.

Across items, linear regression was used to evaluate the extent to which factors (described below) may have affected the log-odds ratio. Separate linear regression was conducted for each scoring method and for each DIF procedure; the average log-odd ratio across replications was used as the dependent variable for each combination of conditions. The independent variables included item parameters (*a*, *b* and *c*), two different sample size ratio ($N_F/N_R = 1$ and $N_F/N_R = 0.5$), and three different ability distributions (one equal- and two unequal- ability distribution) between reference group and focal group as described above. A standardized regression analysis was conducted as follows:

$$LOR = \beta_a a + \beta_b b + \beta_s s + \beta_d \Delta \tag{3.51}$$

for the SIM test, and because *c* parameters are not constant for the ART test:

$$LOR = \beta_a a + \beta_b b + \beta_c c + \beta_s s + \beta_d \Delta \, , \tag{3.52}$$

where *a*, *b* and *c* represent item parameters, s and Δ represent sample size ratio between reference group and focal group, and different ability distributions, respectively.

Binomial regression was used to evaluate the effect of the independent variables on type 1 error. Separate binomial regressions were conducted for each scoring method and for each DIF procedure; the dependent variable for each analysis was determined by the count of the number of times that the log odds ratio fell outside the 95% confidence interval. The independent variables were same as described above for linear regression for LOR, which include *a*, *b*, *c*, s and Δ. Again, only main effects were tested.

Criteria for Evaluation

The nominal α =.05 level of significance was used for all tests. The empirical type 1 error level is defined as the proportion of times (out of 1000 replication) that the log odds ratio falls outside the 95% confidence interval. The average log-odds ratio was calculated for each item across 1000 replications in order to evaluate bias. Because no DIF was introduced to either test, the true value of the LOR was zero.

CHAPTER IV. RESULTS

In this chapter, a detailed description is given of the results obtained following application of the methodology illustrated in Chapter III. The results of two simulation studies based on IRT 3PL models are reported. In the first study, bias, RMSE statistics, and coefficients of determination r^2 are used to evaluate whether the IRT formulas improve estimation of the corrected true score. In the second study, the logistic regression and Mantel-Haenszel procedure is used to obtain DIF under various conditions for different scoring methods. Type 1 error rates and log odds are used to evaluate the accuracy resulting from conditioning on different formula scores.

Study I: Comparisons of Scoring Methods

The purpose of the first study is to find out if the IRT formulas improved true score estimates and to evaluate potential information loss due to the Taylor approximation, the use of pseudo-Bayes estimates and the use of conditional probabilities estimates. All evaluation statistics are presented first by quartile followed by the full distribution for number-correct scores, corrected true scores, and different formula scores. Descriptive statistics include means, standard deviations, skewness, and kurtosis. Bias statistics are used to determine accuracy and the direction of measurement error (either overestimation or underestimation) of the different scoring methods. The coefficient of determination r^2

is used to provide a measure of how well the true score is predicted by each scoring method.

First Simulation Study

Descriptive statistics results by quartile for different scoring methods based on two sets of item parameters are given in Table 4-1 and Table 4-2. For both test designs, all formula scores resulted in a lower average score than R. Moreover, the standard deviation of corrected true score (C_T) was lower than that for any formula score because the latter include differing amounts of measure error. For every quartile, the classical formula score C_k yielded the highest statistical variability. For all scoring methods, Q_1 and Q_4 showed higher variability with larger score ranges, compared to smaller rages in Q_2 and Q_3.

Table 4-1

Descriptive Statistics with N=25000 in Each Quartile: SIM

Q	Statistic	R	C_T	C_K	C_{T1B}	C_{T1P}	B	C_{T2B}	C_{T2P}
Q_1	Mean	14.234	9.553	9.558	9.387	10.356	11.728	10.683	11.773
	SD	3.143	2.425	3.889	3.360	3.548	3.170	3.367	3.243
	Skewness	-0.202	-0.894	-0.134	-0.395	-0.108	-1.017	-0.454	-0.112
	Kurtosis	-0.008	0.199	-0.198	0.168	-0.193	1.769	0.399	-0.134
Q_2	Mean	18.335	14.646	14.669	14.104	15.451	15.328	15.284	16.318
	SD	2.501	1.084	3.127	2.464	2.826	1.972	2.452	2.636
	Skewness	-0.008	-0.071	-0.008	0.052	0.032	-0.122	0.025	0.043
	Kurtosis	-0.098	-1.175	-0.098	0.021	-0.073	0.192	-0.024	-0.080
Q_3	Mean	21.265	18.310	18.331	17.337	19.021	17.617	18.415	19.590
	SD	2.353	1.083	2.941	2.428	2.708	1.845	2.364	2.542
	Skewness	-0.093	0.054	-0.093	0.026	-0.060	-0.004	-0.006	-0.046
	Kurtosis	-0.071	-1.189	-0.071	-0.015	-0.054	0.127	-0.026	-0.065
Q_4	Mean	25.286	23.370	23.358	21.819	23.821	21.062	22.740	24.101
	SD	2.763	2.418	3.454	3.216	3.250	2.665	3.048	3.096
	Skewness	0.112	0.891	0.112	0.447	0.146	0.769	0.372	0.176
	Kurtosis	-0.231	0.189	-0.231	0.211	-0.228	1.151	0.081	-0.207

Note. R: Number-correct score; C_T*: Corrected true score;*

C_K*: Classical formula score;* C_{T1B}*: One-Term Bayes formula score;*

C_{T1P}*: One-Term probability formula score; B: Baseline score;*

C_{T2B}*: M-Term Bayes formula score;* C_{T2P}*: M-Term probability formula score.*

Table 4-2

Descriptive Statistics with N=25000 in Each Quartile: ART

Q	Statistic	R	C_T	C_K	C_{T1B}	C_{T1P}	B	C_{T2B}	C_{T2P}
Q_1	Mean	10.814	7.622	7.454	7.704	8.057	9.544	8.548	9.059
	SD	3.055	2.489	3.533	3.299	3.245	3.233	3.249	3.062
	Skewness	-0.115	-0.568	-0.017	-0.219	0.008	-0.768	-0.282	-0.009
	Kurtosis	-0.217	-0.594	-0.429	-0.252	-0.379	0.564	-0.112	-0.339
Q_2	Mean	15.674	13.364	13.145	13.110	13.610	13.940	13.768	14.216
	SD	2.418	1.244	2.845	2.469	2.646	1.918	2.389	2.505
	Skewness	-0.013	-0.074	-0.013	0.020	0.035	-0.155	0.009	0.035
	Kurtosis	-0.112	-1.169	-0.112	-0.053	-0.100	0.188	-0.036	-0.090
Q_3	Mean	19.252	17.563	17.355	16.940	17.692	16.742	17.453	18.061
	SD	2.340	1.241	2.753	2.455	2.620	1.845	2.365	2.493
	Skewness	-0.059	0.067	-0.059	0.014	-0.032	0.051	0.013	-0.009
	Kurtosis	-0.099	-1.182	-0.099	-0.078	-0.120	0.105	-0.081	-0.112
Q_4	Mean	23.882	22.927	22.803	22.065	23.003	20.769	22.386	23.149
	SD	2.573	2.169	3.027	2.941	2.930	2.553	2.818	2.840
	Skewness	-0.068	0.532	-0.068	0.094	-0.054	0.510	0.085	-0.029
	Kurtosis	-0.403	-0.598	-0.403	-0.355	-0.401	0.363	-0.345	-0.420

Table 4-3 and Table 4-4 summarize descriptive statistics for the full distribution. Predictably, all formula scores had a lower average score than *R*. However, unlike the results by quartile, all IRT formula scores averages were close to the corrected true score averages for both tests. Among the four IRT formula scores, C_{T1B} had descriptive statistics that closely tracked those of the corrected true score for full distribution.

Although all four IRT formula scores were better estimates of the corrected true score for the full distribution, none of them closely tracked the corrected true score in any quartile (see Table 4-1 and Table 4-2). The criteria of bias and r^2 provided more sensitive information, in this context, for comparing the different formula scores than simple descriptive statistics.

Table 4-3

Descriptive Statistics for Full Distribution: SIM

Statistic	R	C_T	C_K	C_{T1B}	C_{T1P}	B	C_{T2B}	C_{T2P}
Mean	19.780	16.470	16.479	15.662	17.162	16.434	16.779	17.946
SD	4.865	5.391	6.070	5.388	5.820	4.202	5.244	5.358
Skewness	-0.107	-0.005	-0.094	-0.051	-0.083	-0.228	-0.111	-0.033
Kurtosis	-0.287	-0.362	-0.327	-0.060	-0.368	0.942	-0.033	-0.347

Table 4-4

Descriptive Statistics for Full Distribution: ART

Statistic	R	C_T	C_K	C_{T1B}	C_{T1P}	B	C_{T2B}	C_{T2P}
Mean	17.405	15.369	15.189	14.955	15.590	15.249	15.539	16.121
SD	5.456	5.915	6.403	5.961	6.185	4.770	5.752	5.843
Skewness	-0.115	-0.085	-0.010	-0.089	-0.069	-0.241	-0.109	-0.042
Kurtosis	-0.580	-0.668	-0.621	-0.484	-0.653	0.240	-0.449	-0.633

Bias

In Table 4-5 and Table 4-6, bias estimates are given for all scores by quartile for both tests. Other than R, the baseline index B had the highest bias in Q_1 and Q_4. This index appears to be the least useful in Q_1 where guessing is likely to have the greatest impact. C_{T2P} also showed high bias in the first quartile. Overall, the classical formula score C_K had the smallest bias in every quartile for the SIM test. However, for the ART test, C_{T1B}, C_{T1P}, and C_{T2B} each had the lowest bias in Q_1, Q_4 and Q_3, respectively. A trend was apparent for the new corrected scores: for C_{T1B} and C_{T2B}, bias trended positive to negative from Q_1 to Q_4. In absolute value, bias increased from Q_1 to Q_4 for C_{T1B}, but decreased from Q_1 to Q_3 and then increased in Q_4 for C_{T2B}. C_{T1P} and C_{T2P} had similar trends in bias. Both had positive bias in every quartile and had a trend to decrease from Q_1 to Q_4. IRT formula scores always resulted in less bias than R.

Table 4-5

Bias by Quartile: SIM

Quartile	R	C_K	C_{T1B}	C_{T1P}	B	C_{T2B}	C_{T2P}
Q_1	4.681	0.004	-0.166	0.803	2.175	1.130	2.220
Q_2	3.689	0.023	-0.541	0.805	0.682	0.638	1.673
Q_3	2.955	0.021	-0.973	0.711	-0.693	0.105	1.280
Q_4	1.916	-0.012	-1.551	0.451	-2.307	-0.630	0.731

Table 4-6

Bias by Quartile: ART

Quartile	R	C_K	C_{T1B}	C_{T1P}	B	C_{T2B}	C_{T2P}
Q_1	3.192	-0.168	0.083	0.435	1.923	0.927	1.437
Q_2	2.309	-0.219	-0.254	0.246	0.575	0.404	0.852
Q_3	1.689	-0.208	-0.623	0.129	-0.821	-0.110	0.498
Q_4	0.956	-0.124	-0.861	0.076	-2.158	-0.541	0.222

Table 4-7 summarizes bias estimates for the full distribution. The second IRT formula scores were derived from the 3PL model, and therefore C_{T2B} and C_{T2P} were expected to provide a better approximation throughout the quartiles. However, C_K still had the smaller bias compared to new scoring methods with only exception that for the ART test, C_{T2B} resulted in the smallest bias.

Table 4-7

Bias for Full Distribution

Test	R	C_K	C_{T1B}	C_{T1P}	B	C_{T2B}	C_{T2P}
SIM	3.310	0.009	-0.807	0.692	-0.036	0.309	1.476
ART	2.037	-0.180	-0.414	0.221	-0.120	0.170	0.752

Plots comparing bias for the various scores are given in Figure 4-1 to 4-10. In these scatter plots, true score categories were created by rounding fractional true scores, C_T, to the nearest integer and then averaging corrected scores within these categories.

Figure 4-1. SIM test: Comparison of bias for *R* and C_K.

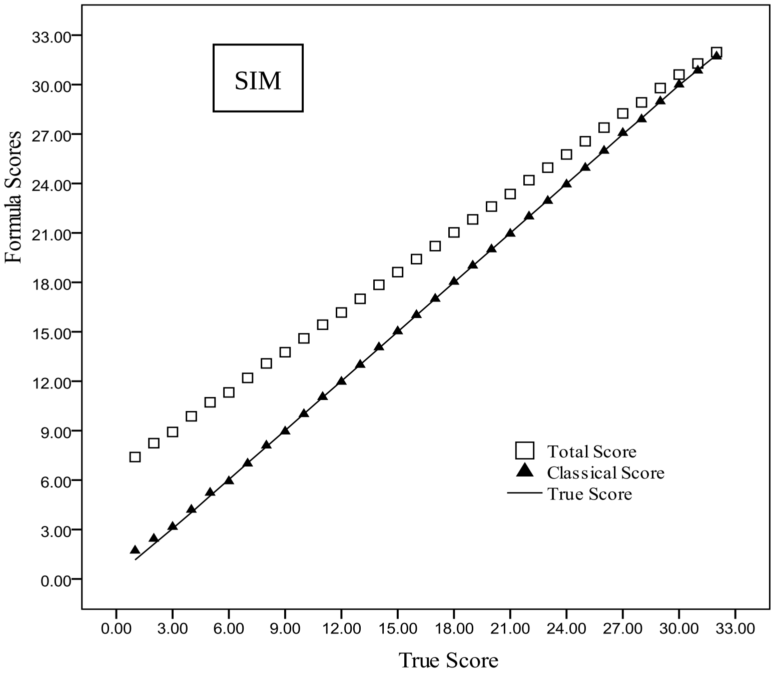
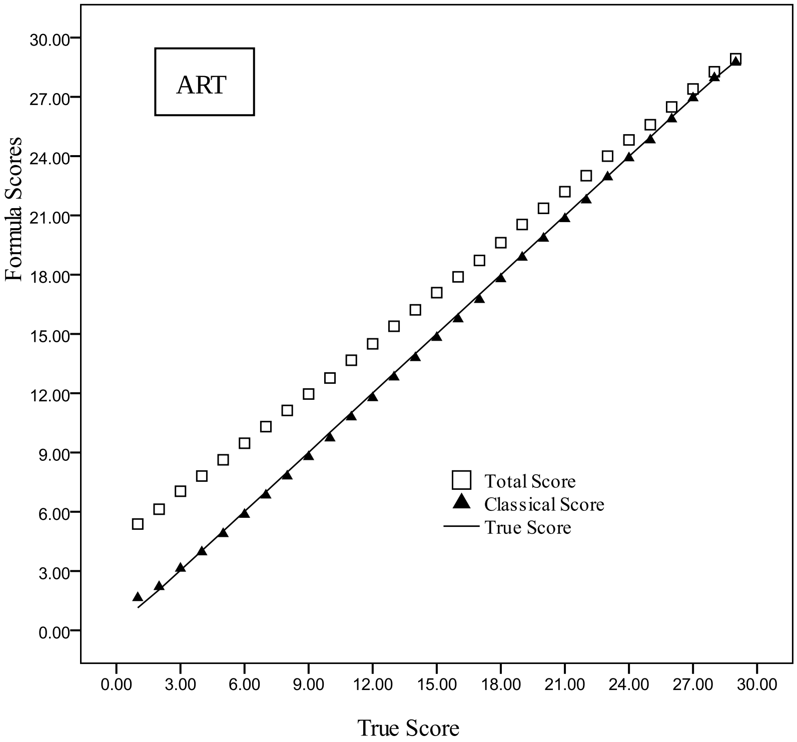


Figure 4-2. ART test: Comparison of bias for *R* and C_K.



For both sets of tests, as it can be seen in Figure 4-1 and 4-2, the classical formula score C_K provided a nearly unbiased estimate of C_T while the number-correct score R initially showed a positive bias and then diminished to zero at the upper range of the true score.

Figure 4-3 and 4-4 demonstrate comparisons among two Bayes formula scores (C_{T1B} and C_{T2B}) and the baseline score (B). In both figures, C_{T1B} and C_{T2B} were compared to the rival score B, and both were at least as good as B over the range. It is evident that C_{T1B} had good estimation in the lower range of true score but exhibited a negative bias at the high end. C_{T2B}, on the other hand, had a positive bias in the lower range of true score and provided a better approximation at the higher end than C_{T1B}.

Figure 4-5 and 4-6 exhibit comparisons among two probability formula scores (C_{T1P} and C_{T2P}) and the baseline score (B). Similar results to the Bayes formula scores were found. C_{T1P} still revealed the least bias compared to C_{T2P} and B. Again, the figures show neither C_{T2B} nor C_{T2P} provided a better approximation of the true score throughout the range as its expectation. Yet they both provided at least as good estimation as B over the range and as good as C_{T1B} and C_{T1P} at the higher end.

Figure 4-3. SIM test: Comparison of bias for C_{T1B}, C_{T2B} and baseline score *B*.

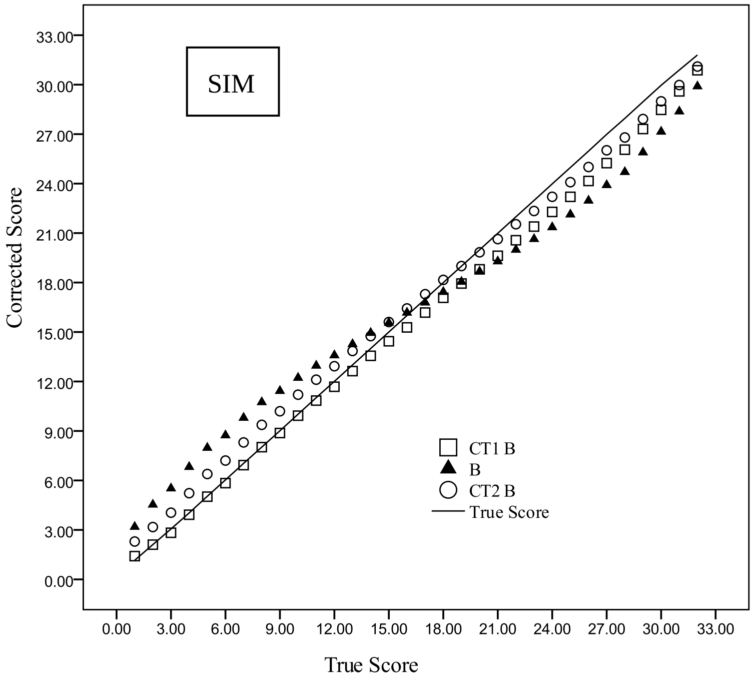


Figure 4-4. ART test: Comparison of bias for C_{T1B}, C_{T2B} and baseline score *B*.

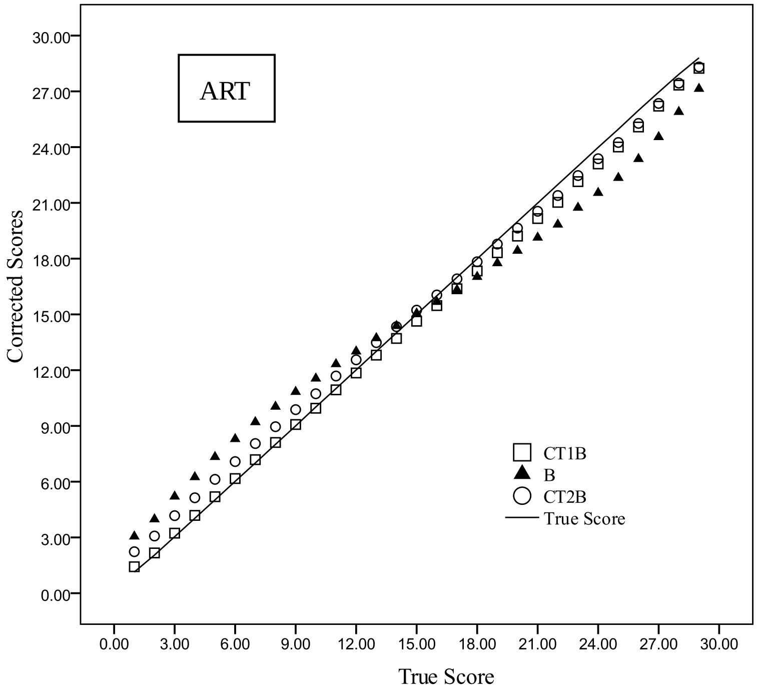


Figure 4-5. SIM test: Comparison of bias for C_{T1P}, C_{T2P} and baseline score *B*.

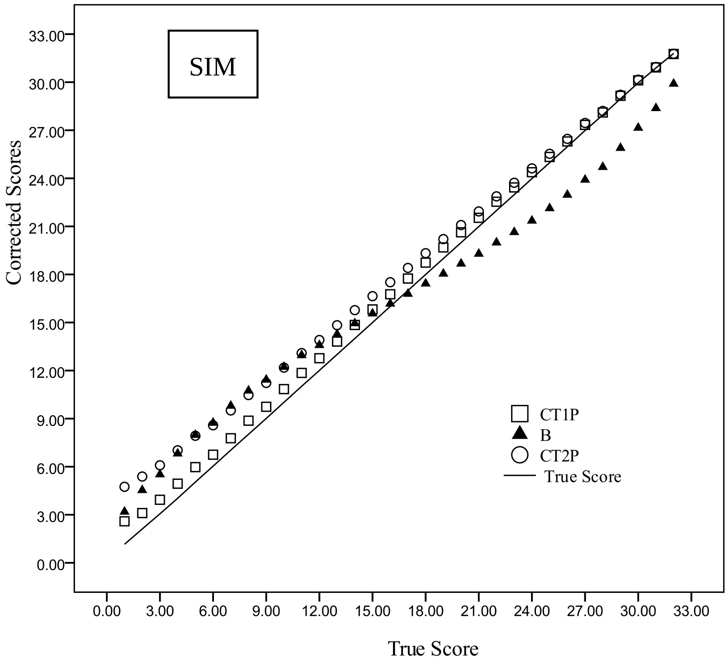


Figure 4-6. ART test: Comparison of bias for C_{T1P}, C_{T2P} and baseline score *B*.

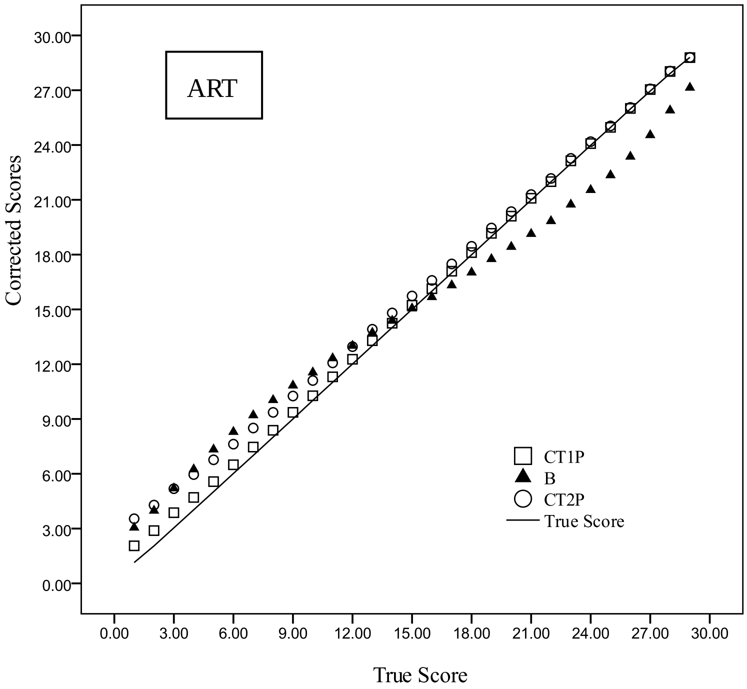


Figure 4-7 to 4-10 showed comparisons among the two different approaches (C_{T1B} vs. C_{T1P}; C_{T2B} vs. C_{T2P}) used to obtain IRT formula scores and the classical formula score C_K. Figure 4-7 and 4-8 revealed a stable pattern that in both tests, one-term Bayes formula score C_{T1B} performed almost as good as C_K in the lower range, where one-term probability formula score C_{T1P} had a positive bias. In contrast, C_{T1P} estimation was almost the same as C_K at the high end, and had better estimation compared to C_{T1B}, which had a negative bias. Figure 4-9 and 4-10 exhibit comparisons among C_K, C_{T2B} and C_{T2P}. C_K revealed the least bias throughout the range. And again, M-term Bayes formula score C_{T2B} showed better approximation to the true score in the lower end while C_{T2P} had better estimation at the higher end of score.

Figure 4-7. SIM test: Comparison of bias for C_{T1B}, C_{T1P} and C_K.

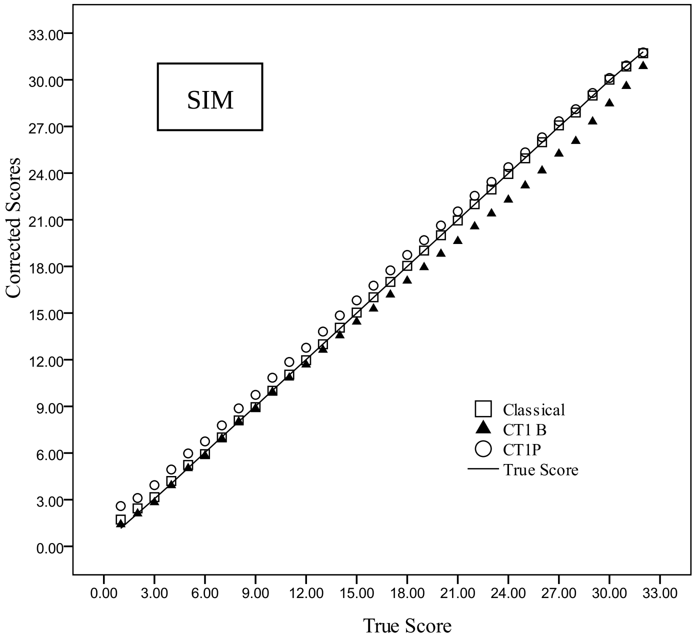


Figure 4-8. ART test: Comparison of bias for C_{T1B}, C_{T1P} and C_K.

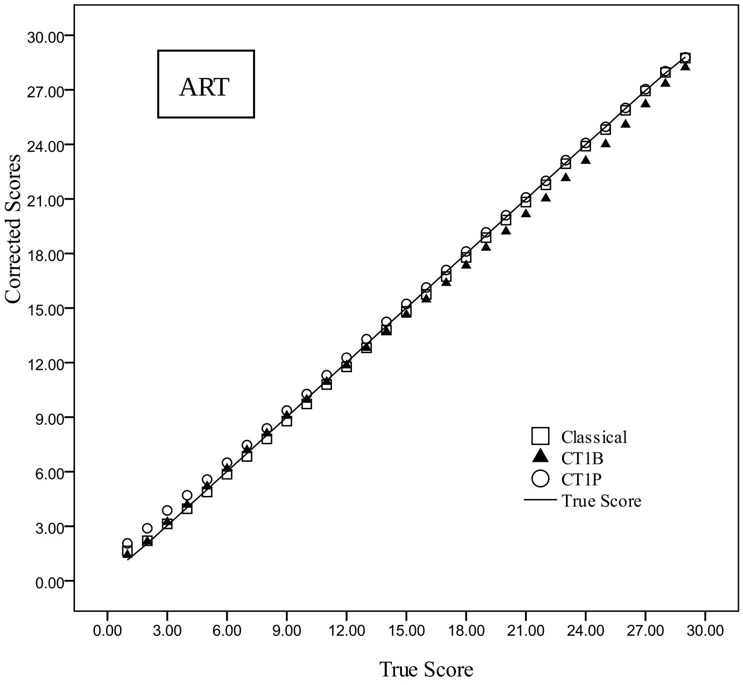


Figure 4-9. SIM test: Comparison of bias for C_{T2B}, C_{T2P} and C_K.

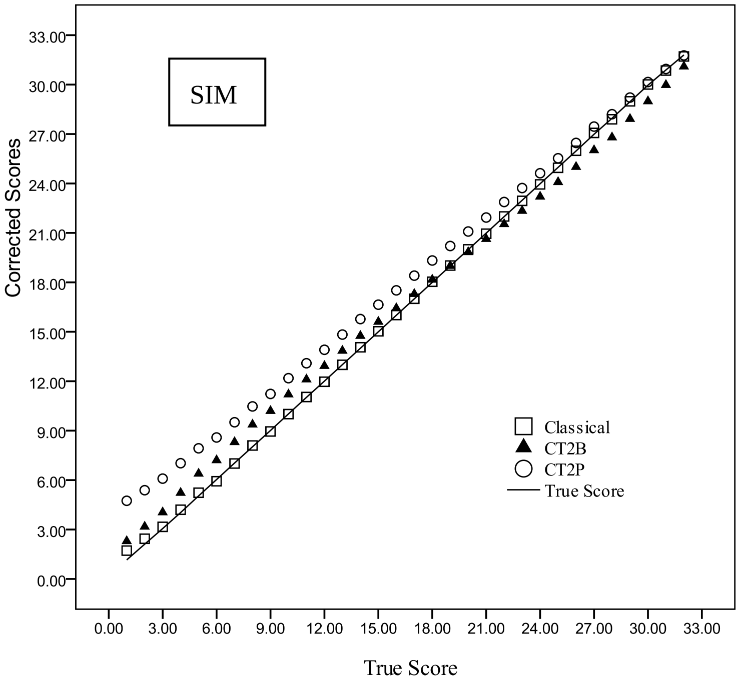
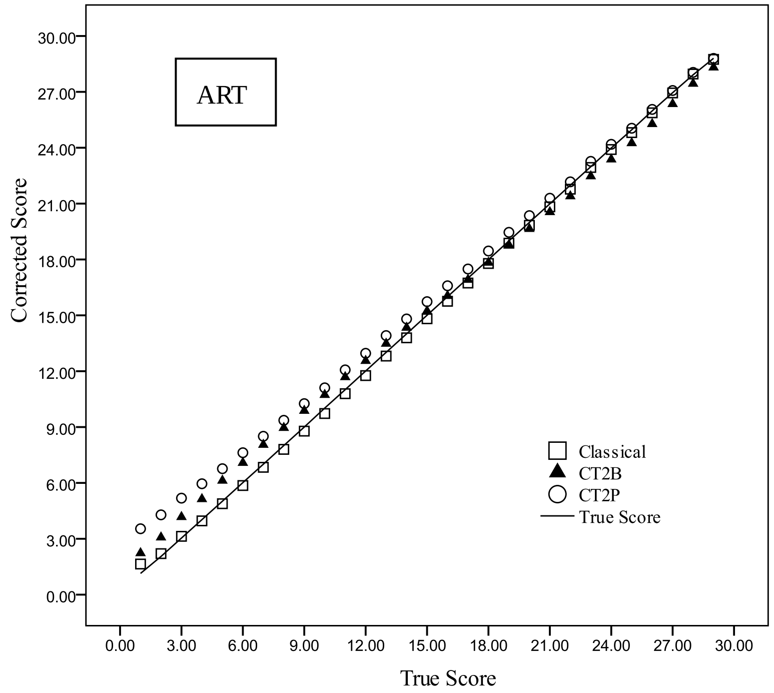


Figure 4-10. ART test: Comparison of bias for C_{T2B}, C_{T2P} and C_K.



Coefficient of Determination r^2

In Table 4-8 and 4-9, r^2 for the different correction scores are given by quartile for two sets of tests. The estimates of r^2 for R and C_K were identical except in the first quartile (due to rounding up of negative values to 0), because they were related by a linear transformation. The r^2 estimation results were similar for both tests. The baseline score B accounted for more variance than the classical formula score C_K in Q_1 and Q_4, but about the same in Q_2 and Q_3 (where variability is lower). The IRT formula scores C_{T1B}, C_{T2B}, C_{T1P}, and C_{T2P}, in contrast, always had a higher r^2 than C_K, and accounted for more variance in Q_1 and Q_4 than in Q_2 and Q_3. In Q_1, where guessing had the largest effect, compared to C_K, the advantage was about 11.3%, 6.3%, 5.5%, and 4.8% of variance for

C_{T1B}, C_{T2B}, C_{T1P}, and C_{T2P}, respectively. The advantage of the IRT-based corrections diminished to 1-3% in the remaining quartiles. Table 4-10 summarizes the r^2 statistics for the full distribution. In contrast to the r^2 between R and C_T, the IRT formula scores had higher, though similar, r^2 for both tests.

Table 4-8

r^2 by quartile: SIM

Quartile	R	C_K	C_{T1B}	C_{T1P}	B	C_{T2B}	C_{T2P}
Q_1	0.382	0.379	0.492	0.462	0.403	0.473	0.429
Q_2	0.123	0.123	0.152	0.144	0.123	0.146	0.137
Q_3	0.138	0.138	0.154	0.151	0.138	0.153	0.147
Q_4	0.492	0.492	0.497	0.498	0.520	0.504	0.500

Table 4-9

r^2 by Quartile: ART

Quartile	R	C_K	C_{T1B}	C_{T1P}	B	C_{T2B}	C_{T2P}
Q_1	0.461	0.458	0.532	0.517	0.473	0.521	0.496
Q_2	0.196	0.196	0.217	0.214	0.196	0.214	0.209
Q_3	0.215	0.215	0.226	0.225	0.215	0.225	0.223
Q_4	0.532	0.532	0.546	0.541	0.553	0.548	0.541

Table 4-10

r^2 for Full Distribution

Test	R	C_K	C_{T1B}	C_{T1P}	B	C_{T2B}	C_{T2P}
SIM	0.786	0.786	0.818	0.814	0.774	0.813	0.806
ART	0.856	0.857	0.870	0.870	0.842	0.868	0.866

The classical formula score C_K provided the least bias among all corrected scores, but the IRT formula scores had higher r^2 values to the corrected true score C_T than the number-correct score R (or C_K) – especially in the first quartile. Comparing the two different approaches to obtain IRT based corrected scores, the two formula scores obtained with the Bayes method (C_{T1B} and C_{T2B}) were more accurate than those obtained with the conditional probability method (C_{T1P} and C_{T2P}) in every studied aspect.

To minimize bias in C_{T1B} and C_{T2B} and keep the higher r^2, a linear transformation was applied in which C_{T1B} and C_{T2B} were scaled to C_K. Because C_K can always be computed directly from the data, this scaling requires no additional information; however, the usefulness of the scaling does depend on the accuracy of the classical formula score. The bias statistics differences between C_{T1B} and C_K decreased at higher proficiency levels (as in Figure 4-7 and 4-8). Moreover, C_{T2B} and C_K were both better estimations at mid-range of C_T and further off at extreme ranges (see Figure 4-9 and 4-10). Both C_{T1B}

and C_{T2B} were better represented as quadratic transformations comparing to linear and cubic transformation. For SIM test, the scaled C_{T1B} and C_{T2B} were obtained as C_{S1B} and C_{S2B}, with the regression

$$\begin{aligned} C_{S1B} &= \text{-}0.72477 + 1.06107\,\mathrm{C}_{T1B} + 0.00017669\,\mathrm{C}_{T1B}^{2} \\ C_{S2B} &= \text{-}1.88151 + 1.08464\,\mathrm{C}_{T2B} + 0.00078921\,\mathrm{C}_{T2B}^{2}. \end{aligned} \tag{4-1}$$

And for ART test, C_{S1B} and C_{S2B} were obtained with the regression

$$\begin{aligned} C_{S1B} &= \text{-}0.72477 + 1.06107\,\mathrm{C}_{T1B} + 0.00017669\,\mathrm{C}_{T1B}^{2} \\ C_{S2B} &= \text{-}1.88151 + 1.08464\,\mathrm{C}_{T2B} + 0.00078921\,\mathrm{C}_{T2B}^{2}. \end{aligned} \tag{4-2}$$

Updated r^2 and bias statistics for two tests are shown in Table 4-11 to 4-14. Since they were related by a linear transformation, the estimates of r^2 for C_{S1B} and C_{S2B} were almost identical to C_{T1B} and C_{T2B} in each quartile and for the full range. Bias-wise, when the analyses carried out by quartile, C_{S1B} and C_{S2B} resulted in smaller bias (in absolute value) compared to C_{T1B} and C_{T2B}, but the result still had a slightly larger bias than C_K (see Table 4-11 and Table 4-12). However, when the analyses focused on overall comparison, C_{S1B} and C_{S2B} had the same bias as C_K (see Table 4-13 and Table 4-14). Figure 4-11 and Figure 4-12 give comparisons among C_{S1B}, C_{S2B}, and C_K. For both sets of tests, as it is shown in the figures, C_{S1B} and C_{S2B} performed comparable to C_K, and all provided nearly unbiased estimate of C_T. In contrast to the untransformed results C_{T1B} and C_{T2B} (Figure 4-7 and Figure 4-10), C_{S1B} and C_{S2B} improved significantly on overall bias reduction (Figure

4-11). Much smaller bias was found on lower and upper end of score after scaling.

Table 4-11

r^2 and Bias by Quartile: SIM

Quartile	Statistic	C_K	C_{T1B}	C_{S1B}	C_{T2B}	C_{S2B}
Q_1	r^2	0.379	0.492	0.492	0.473	0.472
	Bias	0.004	-0.166	-0.052	1.130	-0.057
Q_2	r^2	0.123	0.152	0.152	0.146	0.146
	Bias	0.023	-0.541	0.120	0.638	0.086
Q_3	r^2	0.138	0.154	0.154	0.153	0.153
	Bias	0.021	-0.973	0.042	0.105	0.021
Q_4	r^2	0.492	0.497	0.496	0.504	0.505
	Bias	-0.012	-1.551	-0.074	-0.630	-0.014

Note.C_{T1B}*: One-Term Bayes formula score;* C_{T2B}*: M-Term Bayes formula score;*

C_{S1B}*: Scaled One-Term Bayes formula score;* C_{S2B}*: Scaled M-Term Bayes formula score*

Table 4-12

r^2 and Bias by Quartile: ART

Q	Statistic	C_K	C_{T1B}	C_{S1B}	C_{T2B}	C_{S2B}
Q_1	r^2	0.458	0.532	0.532	0.521	0.520
	Bias	-0.168	0.083	-0.159	0.927	-0.165
Q_2	r^2	0.196	0.217	0.217	0.214	0.214
	Bias	-0.219	-0.254	-0.147	0.404	-0.159
Q_3	r^2	0.215	0.226	0.226	0.225	0.225
	Bias	-0.208	-0.623	-0.261	-0.110	-0.269
Q_4	r^2	0.532	0.546	0.546	0.548	0.548
	Bias	-0.124	-0.861	-0.151	-0.541	-0.125

Table 4-13

r^2 and Bias for Full Distribution: SIM

Statistic	C_K	C_{T1B}	C_{S1B}	C_{T2B}	C_{S2B}
r^2	0.786	0.818	0.818	0.813	0.814
Bias	0.009	-0.807	0.009	0.309	0.009

Table 4-14

r^2 and Bias All Quartiles: ART

Statistic	C_K	C_{T1B}	C_{S1B}	C_{T2B}	C_{S2B}
r^2	0.857	0.870	0.870	0.868	0.868
Bias	-0.180	-0.414	-0.180	0.170	-0.180

Figure 4-11. SIM test: Comparison of bias for C_{S1B}, C_{S2B} and C_K.

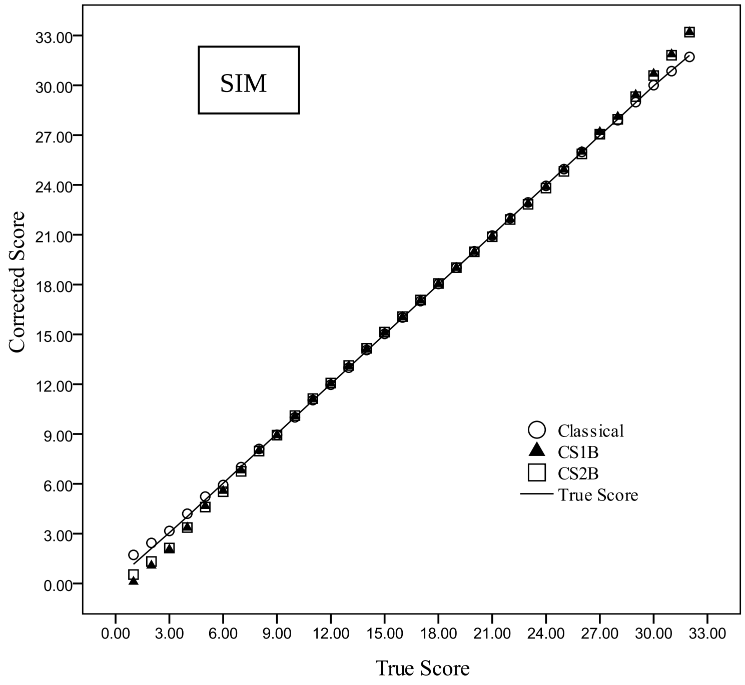
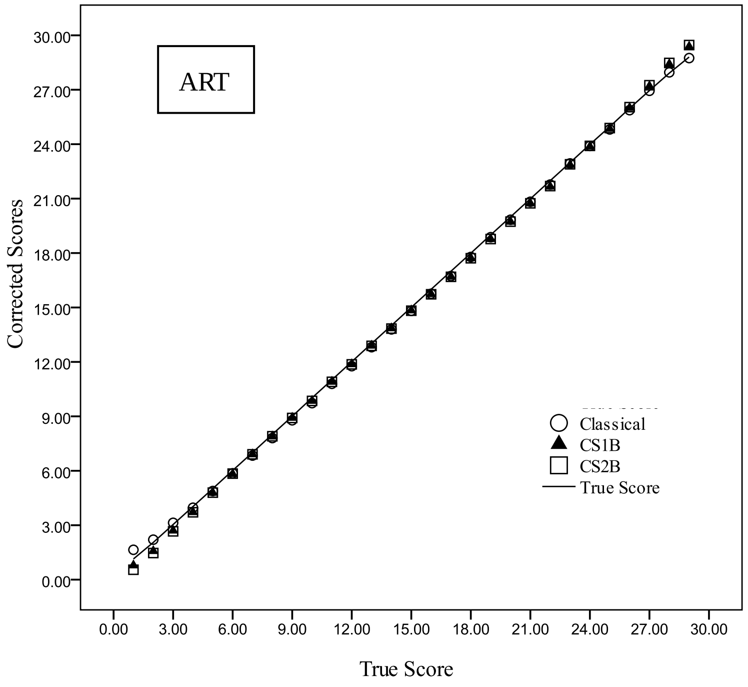


Figure 4-12. ART test: Comparison of bias for C_{S1B}, C_{S2B} and C_K.



Summary of First Simulation Study

The classical formula score provided the least bias of formula score methods, but the IRT formula scores had higher correlations with the corrected true score than the number-correct score—especially in the first quartile. If one is interested only in comparing aggregate test scores to some criterion, this would argue in favor of the classical correction. However, if the goal is to remove the effects of unreliability due to guessing while substantially reducing bias, the IRT formulas have better measurement properties.

Second Simulation Study

The purpose of the second simulation study is to determine the practical utility of using the IRT formula scores in moderately large samples. Accordingly, a set of $n= 5000$ sample for each test was sampled from the data sets generated in the first simulation study with 10 replications. To evaluate the two new formula scores, several benchmarks were created. Recall that the score $\hat{C}_T$ was obtained by substituting sample estimates of item parameters and proficiencies into Equation(3.11). The comparison between $\hat{C}_T$ and C_T is then obtained to establish the maximum level of predictability based on IRT estimates. Second, the scores $\hat{C}_{T1}$ and $\hat{C}_{T2}$ were determined with estimated IRT item parameters and θ using Equations (3.25) and(3.43). These can be used as benchmarks for determining how much information was lost in calculating C_{T1B}, C_{S1B}, C_{T1P}, C_{T2B}, C_{S2B} and C_{T2P} with the observed score methods (Bayes formula scores, scaled-Bayes formula scores and probability formula scores). Note also that $\hat{C}_{T1}$ and $\hat{C}_{T2}$ contained measurement error as well as sampling error in IRT parameters. The corresponding bias statistics, RMSE and r^2 of $\hat{C}_T$, $\hat{C}_{T1}$, $\hat{C}_{T2}$, C_{T1B}, C_{S1B}, C_{T1P}, C_{T2B}, C_{S2B} and C_{T2P} associated with C_T are calculated over 10 replications. Results are presented first by quartile followed by all range.

In Table 4-15 and 4-16, bias, RMSE and r^2 relative to C_T are first given by the first quartile (based on C_T) then calculated for the full range for two tests. Results from two tests were similar. The average biases and RMSE of $\hat{C}_T$ in the first quartile were 0.944 (SIM) and 2.189(ART), and 0.891 (SIM) and 2.185 (ART), and $\hat{C}_T$ explained about 53.2% (SIM) and 54.5% (ART) of the variance of C_T in the first quartile (Table 4-15). However, $\hat{C}_T$ was a better predictor of C_T for SIM and ART in the full distribution: not only was its average bias very small (-0.009 and 0.048, respectively), but the average RMSE was also smaller (2.112 and 2.030) for two tests (Table 4-16). For the full distribution, $\hat{C}_T$ explained about 84.8% (SIM) and 88.2% (ART) of the variance of C_T.

Information Loss due to the Taylor Approximation

To evaluate potential information loss due to Taylor approximation in obtaining, $\hat{C}_{T1}$ and $\hat{C}_{T2}$ were compared with C_T. It can be seen in Table 4-15 that the corresponding absolute values of bias in the first quartile were smaller than $\hat{C}_T$. The RMSEs for $\hat{C}_{T1}$ and $\hat{C}_{T2}$ were slightly higher than $\hat{C}_T$ for SIM and were slightly lower for ART. The amounts of C_T variance explained for SIM and ART by $\hat{C}_{T1}$ (50.6% and 53%) and $\hat{C}_{T2}$ (51% and 53.6%) were only slightly lower than for $\hat{C}_T$ (53.2% and 54.9%). Thus, $\hat{C}_T$ accounted about 2% more variance than $\hat{C}_{T1}$ and about 1.5% more than $\hat{C}_{T2}$ in the first quartile.

Table 4-15

First Quartile Results for 10 Replications of N=1250

		Bias			RMSE			r^2		
Test	Statistic	$\hat{C}_T$	$\hat{C}_{T1}$	$\hat{C}_{T2}$	$\hat{C}_T$	$\hat{C}_{T1}$	$\hat{C}_{T2}$	$\hat{C}_T$	$\hat{C}_{T1}$	$\hat{C}_{T2}$
SIM	Mean	0.944	-0.649	0.421	2.189	2.302	2.257	0.532	0.506	0.510
	SD*	0.116	0.099	0.090	0.057	0.067	0.055	0.014	0.014	0.012
ART	Mean	0.891	-0.355	0.500	2.185	2.173	2.183	0.545	0.530	0.536
	SD*	0.151	0.130	0.106	0.068	0.062	0.055	0.019	0.019	0.018

Note. $\hat{C}_T$ *: IRT estimate corrected true score;*

$\hat{C}_{T1}$ *: IRT estimate one-term formula score;*

$\hat{C}_{T2}$ *: IRT estimate M-term formula score;*

** The standard deviation (SD) measures the stability of the bias result across 10 replications.*

For the full range (Table 4-16), compared to $\hat{C}_T$, $\hat{C}_{T1}$ and $\hat{C}_{T2}$ had larger bias, RMSE, although the differences were not large. The proportion of C_T variance explained for SIM and ART by $\hat{C}_{T1}$ (83.5% and 87.6%) and $\hat{C}_{T2}$ (83.7% and 87.7%), both were only slightly less than by $\hat{C}_T$ (84.8% and 88.2%). Consequently, there appears to be very little information lost due to Taylor approximation.

Table 4-16

All Quartiles Results for 10 Replications of N=5000

		Bias			RMSE			r^2		
Test	Statistic	$\hat{C}_T$	$\hat{C}_{T1}$	$\hat{C}_{T2}$	$\hat{C}_T$	$\hat{C}_{T1}$	$\hat{C}_{T2}$	$\hat{C}_T$	$\hat{C}_{T1}$	$\hat{C}_{T2}$
SIM	Mean	-0.009	-0.935	0.060	2.112	2.503	2.270	0.848	0.835	0.837
	SD*	0.084	0.082	0.073	0.023	0.037	0.028	0.003	0.003	0.003
ART	Mean	0.048	-0.491	0.150	2.030	2.264	2.132	0.882	0.876	0.877
	SD*	0.078	0.098	0.078	0.018	0.042	0.020	0.003	0.003	0.003

*Note. * The standard deviation (SD) measures the stability of the bias result across 10 replications.*

Information Loss of Pseudo-Bayes and Conditional probability Estimates

It would be expected on theoretical grounds that the IRT estimates, $\hat{C}_{T1}$ and $\hat{C}_{T2}$ would lead to better estimation on C_T, compared to either Bayes or probability formula socre, and this indeed was the case for both cases of the first quartile and the full range. Both $\hat{C}_{T1}$ and $\hat{C}_{T2}$ had smaller biases (Table 4-17 and Table 4-18), RMSEs (Table 4-19 and Table 4-20), and higher r^2 (Table 4-21 and Table 4-22). There were two exceptions to this general finding for both sets of tests: the scaled-Bayes formula scores C_{S1B} and C_{S2B} always had smaller bias compared to $\hat{C}_{T1}$ and $\hat{C}_{T2}$; and in the first quartile, C_{T1B} had a smaller bias than $\hat{C}_{T1}$. The latter result was possible an artifact of overfit because sample statistics rather than population estimates were used to construct the formula scores. It is also important to recognize that the effectiveness of the scaling depends on the accuracy of the classical formula score.

Table 4-17

Average Bias: First Quartile Results for 10 Replications of N=1250

		IRT		*First Formula*			*Second Formula*		
Test	Statistic	$\hat{C}_{T1}$	$\hat{C}_{T2}$	C_{T1B}	C_{T1P}	C_{S1B}	C_{T2B}	C_{T2P}	C_{S2B}
SIM	Mean	-0.649	0.421	-0.228	0.902	-0.082	1.078	2.256	-0.090
	SD	0.099	0.090	0.071	0.061	0.080	0.072	0.057	0.075
ART	Mean	-0.355	0.500	0.087	0.502	-0.160	0.933	1.459	-0.166
	SD	0.130	0.106	0.083	0.074	0.080	0.079	0.071	0.083

Note. $\hat{C}_{T1}$*: IRT estimate one-term formula score;* $\hat{C}_{T2}$*: IRT estimate M-term formula score;*

C_{T1B}*: One-Term Bayes formula score;* C_{T1P}*: One-Term probability formula score;*

C_{T2B}*: M-Term Bayes formula score;* C_{T2P}*: M-Term probability formula score;*

C_{S1B}*: Scaled One-Term Bayes formula score;* C_{S2B}*: Scaled M-Term Bayes formula score*

Table 4-18

Average Bias: Full Distribution Results for 10 Replications of N=5000

		IRT		*First Formula*			*Second Formula*		
Test	Statistic	$\hat{C}_{T1}$	$\hat{C}_{T2}$	C_{T1B}	C_{T1P}	C_{S1B}	C_{T2B}	C_{T2P}	C_{S2B}
SIM	Mean	-0.935	0.060	-0.842	0.729	0.002	0.285	1.499	0.002
	SD	0.082	0.073	0.040	0.027	0.026	0.027	0.027	0.026
ART	Mean	-0.491	0.150	-0.388	0.255	-0.165	0.193	0.772	-0.165
	SD	0.098	0.078	0.046	0.038	0.042	0.042	0.038	0.042

Table 4-19

Average RMSE: First Quartile Results for 10 Replications of N=1250

		IRT		*First Formula*			*Second Formula*		
Test	Statistic	$\hat{C}_{T1}$	$\hat{C}_{T2}$	C_{T1B}	C_{T1P}	C_{S1B}	C_{T2B}	C_{T2P}	C_{S2B}
SIM	Mean	2.302	2.257	2.406	2.696	2.679	2.673	3.326	2.769
	SD	0.067	0.055	0.049	0.055	0.056	0.038	0.048	0.055
ART	Mean	2.173	2.183	2.260	2.285	2.403	2.434	2.623	2.472
	SD	0.062	0.055	0.055	0.041	0.066	0.043	0.037	0.066

Table 4-20

Average RMSE: Full Distribution Results for 10 Replications of N=5000

		IRT		*First Formula*			*Second Formula*		
Test	Statistic	$\hat{C}_{T1}$	$\hat{C}_{T2}$	C_{T1B}	C_{T1P}	C_{S1B}	C_{T2B}	C_{T2P}	C_{S2B}
SIM	Mean	2.503	2.270	2.496	2.594	2.544	2.373	2.851	2.595
	SD	0.037	0.028	0.028	0.027	0.029	0.025	0.021	0.030
ART	Mean	2.264	2.132	2.202	2.229	2.287	2.160	2.307	2.313
	SD	0.042	0.020	0.026	0.019	0.025	0.019	0.016	0.026

It appears that in moderately large samples, much of information in C_T was retained by Bayes and probability formula scores, as indicated by the high correlations with C_T, especially in the full range (Table 4-22). The correlations between C_T and C_{T1B} (note that the scaled formulas have the same correlational properties as the original ones) were 0.91 and 0.93 for the SIM and ART tests, respectively; which were about the same as $\hat{C}_{T1}$. Even though the correlations were smaller in the first quartile compared to correlations of the full range (see Table 4-21), all formula scores had correlations in range of 0.65 -0.73 and were only slightly smaller than compared to IRT estimate scores $\hat{C}_{T1}$ and $\hat{C}_{T2}$ (0.71 and 0.73 for SIM and ART, respectively). In Table 4-21, it can be seen that in the first quartile, $\hat{C}_{T1}$ explained about 1.4% and 5% more of C_T variance than C_{T1B} and C_{T1P} for the SIM test. For ART, the difference was even smaller. There was no average r^2 difference between $\hat{C}_{T1}$ and C_{T1B}, and only 1.9% difference between $\hat{C}_{T1}$ and C_{T1P}. Comparable results were found between $\hat{C}_{T2}$ and C_{T2B}, C_{T2P} and also in the full distribution.

Table 4-21

Average r^2: First Quartile Results for 10 Replications, N=1250

		IRT		*First Formula*		*Second Formula*	
Test	Statistic	$\hat{C}_{T1}$	$\hat{C}_{T2}$	C_{T1B}	C_{T1P}	C_{T2B}	C_{T2P}
SIM	Mean	0.506	0.510	0.492	0.456	0.473	0.429
	SD	0.014	0.012	0.013	0.017	0.013	0.017
ART	Mean	0.530	0.536	0.530	0.511	0.519	0.495
	SD	0.019	0.018	0.018	0.018	0.018	0.018

Table 4-22

Average r^2: Full Distribution Results for 10 Replications, N=5000

		IRT		*First Formula*		*Second Formula*	
Test	Statistic	$\hat{C}_{T1}$	$\hat{C}_{T2}$	C_{T1B}	C_{T1P}	C_{T2B}	C_{T2P}
SIM	Mean	0.835	0.837	0.820	0.815	0.815	0.807
	SD	0.003	0.003	0.004	0.004	0.004	0.004
ART	Mean	0.876	0.877	0.871	0.871	0.868	0.867
	SD	0.003	0.003	0.003	0.003	0.003	0.003

Comparison between Bayes Formula Scores

Both SIM and ART tests revealed similar results. Bias was higher for C_{T2B} than C_{T1B} in the first quartile, and the direction and the magnitude of the bias were consistent with the expectations from the first simulation study. However, C_{T2B} provided a better approximation throughout the quartiles, which was also consistent with the result from the first simulation. The RMSE was also higher for C_{T2B} than C_{T1B} in the first quartile, yet C_{T2B} had smaller RMSE in the full score range. The average squared correlation for C_{T1B} and C_{T2B} in the first quartile were r^2=0.492, r^2=0.473 (SIM) and r^2=0.530, r^2=0.519 (ART), respectively. These were either the same or slightly lower than the large-sample squared correlations given in the first simulation (see Table 4-8 and Table 4-9). The one-term Bayes formula score C_{T1B} consistently had a higher r^2 than C_{T2B} throughout quartiles. Similar to the results in the first simulation study, the scaled-Bayes formula scores C_{S1B} and C_{S2B} improved significantly on bias estimation from C_{T1B} and C_{T2B} while keeping r^2 identical to that of C_{T1B} and C_{T2B}.

Comparison between Probability Formula Scores

Results for bias, RMSE, and r^2 showed comparable trends with the Bayes formula scores, with the exception of a greater bias was found in full score range of C_{T2P}. Again, this result was consistent with the finding from the first simulation study.

Comparison between Bayes and Probability Formula Score

Similar with the results in the first simulation study, the Bayes formulas scores (C_{T1B} and C_{T2B}) retained more true score information and had smaller bias in the first quartile than the probability formula scores (C_{T1P} and C_{T2P}). Overall, C_{T1B} performed best among these four alternatives.

Summary of Second Simulation Study

Relative to a pragmatic criterion created through IRT calibration, the IRT-based corrections $\hat{C}_{T1}$ and $\hat{C}_{T2}$ tracked the corrected true score C_T closely. Moreover, there was not much information loss due to Taylor approximation. The use of Bayes and probability formula scores also resulted in little information loss for the two tests studies with moderately large sample sizes. Finally, the moderate-sized samples resulted in similar result with large-sized samples.

Study II: Applications to DIF Analyses

The purpose of study II is to demonstrate a potential application of IRT formula scoring methods to DIF. The MH and the LR procedures were used to evaluate how the IRT formula scores performed as conditioning scores for DIF analysis, compared to number-correct score. Average type 1 errors and average log-odds ratios were obtained for both procedures, under the condition of no DIF (e.g., the null hypothesis is true). The average log-odds ratio was calculated for each item across 1000 replications in order to evaluate bias. Linear regression was then used to evaluate which factors affect differences in the average log-odds ratio and type 1 error (dependent variables) across items. Separate linear regression was conducted for each scoring method and for each DIF procedure. Independent variables including item parameters, ability distributions, focal group sample size were tested. The nominal α =.05 level of significance was used for all tests.

Type 1 Error

The Type 1 error rates for each DIF identification procedure, by all combinations of the factors included in this study, are summarized in Table 4-23 and Table 4-24 for the two tests. The results showed that a similar pattern of performance on type 1 error rates for all different scoring methods. The average type 1 error rate was close to or less than 0.05 for equal means in the θ distributions (Δ=0). As predicted, type 1 error rates increased as the separation between the ability distributions of the two groups increased. This effect was more pronounced when the focal group had n=1000 cases versus n=500 cases.

Table 4-23

Mean Type 1 error Proportions at $\alpha = 0.05$ *for SIM*

				Type 1 Error Rate				
Procedure	Δ	n_R	n_F	R	C_{T1B}	C_{T1P}	C_{T2B}	C_{T2P}
MH	0	1000	500	0.047	0.048	0.048	0.048	0.048
		1000	1000	0.049	0.048	0.049	0.048	0.049
	0.5	1000	500	0.070	0.056	0.059	0.057	0.063
		1000	1000	0.079	0.058	0.063	0.060	0.068
	1	1000	500	0.143	0.074	0.088	0.079	0.106
		1000	1000	0.177	0.089	0.106	0.094	0.128
LR	0	1000	500	0.048	0.048	0.047	0.048	0.047
		1000	1000	0.049	0.048	0.049	0.049	0.049
	0.5	1000	500	0.073	0.057	0.060	0.059	0.064
		1000	1000	0.080	0.059	0.062	0.060	0.067
	1	1000	500	0.165	0.091	0.106	0.100	0.125
		1000	1000	0.190	0.094	0.112	0.103	0.135

Results for the SIM test are shown in Table 4-23. When comparisons were made within the same scoring method (R, C_{T1B}, C_{T2B}, C_{T1P}, or C_{T2P}) under the same settings (Δ, n_R, n_F), the MH procedure had a lower probability of incurring type 1 errors than the LR procedure in almost all cases. Similarly, results from ART also showed that MH had lower type 1 errors at higher delta settings (see Table 4-24). Findings from both tests were consistent with previous studies that have found the LR procedure to have slightly higher type 1 error rates than the MH procedure (Swaminathan & Rogers, 1990; Narayanan & Swaminathan, 1996; Huang, 1998).

The results showed that type 1 error rates varied across different scoring methods. Type 1 errors associated with IRT formula scores were consistently lower in every condition. Type 1 error differences between conditioning on IRT formula scores versus R increased when Δ and focal group size increased. For the MH procedure based on SIM with$\Delta=0$, average type 1 error rate differences between R and IRT-based scores were about 0.001 for $n_F=500$ and $n_F=1000$. However, when Δ increased to 0.5, the average differences increased to 0.010 for $n_F=500$ (range = 0.007 to 0.014) and 0.017 for $n_F=1000$ (range = 0.011 to 0.021). When $\Delta=1$, the differences increased to 0.056 (range = 0.037 to 0.069) and 0.072 (range = 0.049 to 0.088). Similar results were found for the LR procedure and the ART test.

To compare type 1 errors between two IRT formula scores (C_{T1B} vs. C_{T2B} and C_{T1P} vs. C_{T2P}), it appeared that the first IRT formula scores (C_{T1B} and C_{T1P}) had lower type 1 errors than the second IRT formula scores (C_{T2B} and C_{T2P}). Within the same IRT formula, Bayes formula scores resulted in lower type 1 error rates, compared to probability formula scores (C_{T1B} vs. C_{T1P}, and C_{T2B} vs. C_{T2P}). Overall, C_{T1B} had the lowest average type 1 error in every setting. The same trends were found for both tests and both DIF procedures.

Table 4-24

Mean Type 1 error Proportions at $\alpha = 0.05$ *for ART*

				Type 1 Error Rate				
Procedure	Δ	n_R	n_F	R	C_{T1B}	C_{T1P}	C_{T2B}	C_{T2P}
MH	0	1000	500	0.050	0.051	0.051	0.051	0.051
		1000	1000	0.050	0.050	0.049	0.050	0.049
	0.5	1000	500	0.059	0.053	0.055	0.053	0.056
		1000	1000	0.062	0.054	0.057	0.056	0.057
	1	1000	500	0.093	0.062	0.067	0.065	0.073
		1000	1000	0.109	0.068	0.072	0.070	0.082
LR	0	1000	500	0.050	0.050	0.050	0.050	0.050
		1000	1000	0.050	0.049	0.049	0.050	0.049
	0.5	1000	500	0.060	0.053	0.054	0.054	0.057
		1000	1000	0.062	0.054	0.055	0.054	0.057
	1	1000	500	0.114	0.073	0.080	0.081	0.091
		1000	1000	0.116	0.069	0.075	0.074	0.088

Log-odds Ratio

Because no DIF was simulated, the value of the LOR was expected to be near zero; thus, LOR simultaneously represented the indicator of DIF effect size and bias. If LOR is greater than 0, an item favors the reference group. On the contrary, if LOR is less than 0, the item favors the focal group. Because positive and negative DIF tend to cancel across items within a test, the average LOR across items is not an appropriate evaluation statistic. For this reason, average root mean squared log-odds ratios (RMS) across items were used for the two tests as shown in Table 4-25 and Table 4-26.

Table 4-25

Average Root Mean Squared Log-Odds Ratio for SIM

				RMS-LOR				
Procedure	Δ	n_R	n_F	R	C_{T1B}	C_{T1P}	C_{T2B}	C_{T2P}
MH	0	1000	500	0.010	0.009	0.010	0.010	0.009
		1000	1000	0.007	0.007	0.006	0.007	0.007
	0.5	1000	500	0.081	0.046	0.052	0.056	0.064
		1000	1000	0.084	0.047	0.053	0.057	0.066
	1	1000	500	0.168	0.092	0.103	0.112	0.132
		1000	1000	0.170	0.094	0.104	0.113	0.133
LR	0	1000	500	0.011	0.010	0.011	0.010	0.010
		1000	1000	0.007	0.007	0.007	0.007	0.007
	0.5	1000	500	0.083	0.046	0.053	0.056	0.065
		1000	1000	0.084	0.047	0.054	0.055	0.065
	1	1000	500	0.178	0.102	0.116	0.121	0.141
		1000	1000	0.175	0.098	0.111	0.113	0.135

The average RMS resulted from both tests and both procedures clearly indicated a pattern of increasing effect size with increasing Δ for each scoring method. Visually comparing effect sizes between two procedures the average RMS of the MH procedure was consistently smaller than that of the LR procedure.

The RMS pattern associated with focal group sample size appeared inconsistent. Regardless of scoring methods, the RMSs of the MH procedure were always smaller with 500 individuals in the focal group compared to 1000 individuals, but no difference in the group abilities (Δ=0). On the other hand, for the LR procedure, the RMSs were larger when focal group size was 500 in almost all scoring methods. However, the RMS differences between two different sample sizes were generally small: the greatest difference was 0.018 when using the LR procedure with Δ=1 with C_{T2B} on the ART test.

These results are consistent with those on type 1 error rates and illustrates that RMS can vary across different scoring formulas with the exception of the Δ=0 condition. When Δ=0, RMSs were almost identical among all five scoring methods studied here. When Δ increased, RMSs associated with IRT formula scores were consistently lower than those for R in every setting and this advantage increased when Δ increased. When Δ≠0, the RMSs were slightly higher for the second IRT formula scores (C_{T2B} and C_{T2P}) compared to the first IRT formula scores (C_{T1B} and C_{T1P}) in each setting. Within the same

IRT approach, Bayes formula scores gave lower RMSs, compared to probability formula scores (C_{T1B} vs. C_{T1P} and C_{T2B} vs. C_{T2P}). Overall, for both sets of tests and both DIF procedures, C_{T1B} had the lowest RMSs in every setting.

Table 4-26

Average Root Mean Squared Log-Odds Ratio for ART

				RMS-LOR				
Procedure	Δ	n_R	n_F	R	C_{T1B}	C_{T1P}	C_{T2B}	C_{T2P}
MH	0	1000	500	0.012	0.012	0.013	0.013	0.012
		1000	1000	0.007	0.007	0.007	0.008	0.007
	0.5	1000	500	0.053	0.029	0.033	0.034	0.039
		1000	1000	0.055	0.032	0.035	0.036	0.042
	1	1000	500	0.118	0.064	0.073	0.074	0.089
		1000	1000	0.119	0.066	0.074	0.075	0.090
LR	0	1000	500	0.012	0.012	0.012	0.011	0.011
		1000	1000	0.007	0.007	0.007	0.007	0.007
	0.5	1000	500	0.054	0.030	0.034	0.035	0.042
		1000	1000	0.055	0.031	0.035	0.035	0.041
	1	1000	500	0.133	0.079	0.091	0.089	0.106
		1000	1000	0.116	0.069	0.075	0.074	0.088

Factors influencing the LOR

For the five different scoring methods studied in this thesis, linear regression was used to evaluate which independent variables affect the log-odds ratio of the two DIF procedures.

Among four different IRT formula scores, C_{T1B} resulted in the greatest improvement on reducing LOR.

Table 4-27 and Table 4-28 display the regression results for C_{T1B} and *R*. Values in the tables represent the regression coefficients for both tests and both DIF procedures.

Table 4-27

Summary of Regression Parameter Estimates for SIM

		LOR		Type 1 Error	
Scoring Methods	Main Effects	MH	LR	MH	LR
R	*a*	-0.084^{**}	-0.082^{**}	0.780^{**}	0.833^{**}
	b	0.040^{**}	0.042^{**}	-0.024	0.006
	Δ	-0.039^{*}	-0.030^{*}	1.428^{**}	1.583^{**}
	R/F Ratio	0.003	0.010	-0.184^{**}	-0.132^{**}
	r^2	0.478	0.462	0.549	0.577
C_{T1B}	*a*	-0.075^{**}	-0.071^{**}	0.175^{*}	0.174^{*}
	b	0.013^{**}	0.016^{**}	-0.059^{**}	-0.031
	Δ	-0.015	0.003	0.582^{**}	0.749^{**}
	R/F Ratio	0.002	0.011	-0.098^{**}	-0.038^{*}
	r^2	0.387	0.367	0.448	0.536

Note. $^{**}p<.01$; $^{*}p<.05$

As shown in Tables 4-27 and 4-28 for both sets of tests and both DIF procedures, the regression analyses revealed that item discrimination (*a*) and item difficulty (*b*) were significantly related to the LOR for different scoring methods. Item discrimination had a negative correlation with the LOR for all five scoring methods; as item discrimination increased, the LOR decreased. Conversely, item difficulty showed a positive relationship; LOR was higher when the item was more difficult. For the ART test, in addition to the effects addressed above, the guessing parameter (*c*) also had a positive relationship with LOR, but only on C_{T1B} and C_{T1P} under the MH procedure.

Table 4-28

Summary of Regression Parameter Estimates for ART

		LOR		Type 1 Error	
Scoring Methods	Main Effects	MH	LR	MH	LR
R	*a*	-0.078^{**}	-0.076^{**}	0.651^{**}	0.616^{**}
	b	0.030^{**}	0.032^{**}	0.062	0.098
	c	0.088	0.052	2.293	2.043
	Δ	-0.020^{*}	-0.004	0.803^{**}	0.974^{**}
	R/F Ratio	0.002	0.010	-0.097^{**}	-0.020
	r^2	0.495	0.473	0.379	0.429
C_{T1B}	*a*	-0.068^{**}	-0.063^{**}	0.159	0.136
	b	0.008^{**}	0.012^{**}	-0.013	0.020
	c	0.195^{*}	0.142	0.902	1.613
	Δ	-0.014^{*}	0.011	0.270^{**}	0.385^{**}
	R/F Ratio	0.002	0.010	-0.042^{*}	0.023
	r^2	0.418	0.382	0.144	0.210

Note. $^{**}p<.01$; $^{*}p<.05$

In both tests and both DIF procedures, the associations between main effects and LOR under IRT formula scores (C_{T1B}, C_{T1P}, C_{T2B}, and C_{T2P}) were not as strong as that observed using the number-correct score (R). When the number-correct scores were compared in both tests, almost all main effects under the IRT formula scores had less impact on the LOR. The only exception was the guessing parameter c which had greater impact. Results suggested that the IRT formula scores reduced the confounding of bias with the item discrimination, item difficulty, group ability difference, and different group size ratio but did not reduce the relationship with the guessing. This is because the IRT formula scores more effectively conditioned out residual effects related to proficiency and item parameters. The r^2 estimates obtained by conditioning on IRT formula scores were expected to the smaller. Indeed, the r^2, LORs of IRT formula scores were lower as shown in Tables 4-27 and 4-28.

Factors Influencing Type 1 Errors

Binomial regression was used to obtain to determine which independent variables had effects on type 1 error rates of the two DIF procedures. Similar to LOR, in both tests and both DIF procedures, the associations between main effects and type 1 error under the C_{T1B}, C_{T1P}, C_{T2B}, and C_{T2P} were generally not as strong as that observed using R. In both DIF procedures, relationships between item difficulty and type 1 error were stronger for the IRT formula scores than for the number-correct score for the SIM test. For the ART test, the relationships between group size ratio and type 1 error were stronger for the IRT formula scores than for the number-correct score on the LR procedure.

Ideally, the performance of DIF detection procedures should be unaltered across different tests. However, unlike the LOR, the two DIF procedures revealed different trends for type 1 errors on the SIM test. For both DIF procedures, item discrimination and group ability difference had significant and positive relationship with type 1 error for all five different scores; as item discrimination or group ability difference increased, so did the type 1 error rate. Using the MH procedure, all main effects had significant relationships with type 1 error for all five different scores with the exception of item difficulty. Item difficulty of C_{T1B} and C_{T2B} had a significantly negative effect on the type 1 error rate. The LR procedure, on the other hand, yielded different results. Item difficulty did not affect the type 1 error for any of the scoring methods. Group size ratio did not show significant relationship with type 1 error for C_{T2B} and C_{T2P}.

On ART test, more comparable results were found between two DIF procedures. For both DIF procedures, group ability difference also had a significantly positive relationship with type 1 error rate for all five different scores; item discrimination did not have an effect on type 1 errors for the IRT formula scores except C_{T2B} and C_{T2P}, yet significant effects were found for *R*. Only C_{T2P} of the LR procedure showed an association between item difficulty and type 1 error rate. In general, as it was found in LOR results, IRT formula scores decreased the effects on type 1 error of DIF detection.

Summary of Study II

When applied to MH and LR DIF analyses, the IRT formula scores resulted lower bias in both the LOR and lower type 1 error rates compared to the number-corrected score. Highly similar patterns were found for the other IRT formulas studied in this thesis. Overall, the new formula scores decreased bias in the LOR by about 5.6% and in the type 1 error rate by about 9.6%.

CHAPTER V. DISCUSSION

Multiple-choice items are often favored in standardized achievement tests because of relatively easier scoring. The goal of cognitive measurement is to get the optimum performance of examinees relative to a target construct. Yet in order to get the best possible result in the exam, examinees use various strategies, not all of which are construct relevant. Guessing is one of them. Different methods have been applied to remove guessing effects from test scores, which include penalties for wrong answers and partial credit for omitting responses. The most common formula scoring methods adjust for guessing equally for every item and every examinee. However, the IRT formula scores were functions of both the proficiency of an examinee as well as the examinee's pattern of responses across items.

Correction within the Framework of IRT 3-PL Model

The first purpose of this dissertation was to investigate conceptually how "correction for guessing" works within the framework of a 3PL IRT model, and in turn whether IRT formula scores are able to produce more reliable and accurate estimates of true scores that would be obtained without guessing. Unlike the classical formula scores in which points are subtracted from the number-correct scores based on the number of incorrect responses, the IRT formula scores adjusted proficiency estimates based on correctly answered items only. The same logic is evident in a maximum likelihood estimation of proficiency, which views with varying degrees of suspicion correct answers to questions that are difficult relative to an examinee’s proficiency.

Comparison of Different Scores

Two IRT formula scores were obtained by two different methods (pseudo-Bayes and conditional probability) which that use observed scores only. Results from the first simulation study using two different sets of item parameters did not favor a particular IRT formula score relative to item bias when compared to the classical formula scores. However, the IRT formula scores showed notable improvement when compared to the number-correct score.

Although the classical formula score performed better in terms of bias statistics, it was shown that the IRT formula scores had higher correlations with the corrected true score than the number-correct or the classical formula scores. In terms of r^2, both IRT formula scores provided practical improvement over the classical formula score. Overall, the first IRT formula scores seemed to work best in the first quartile. The second IRT formula score appeared to have a slight advantage in the fourth quartile. The advantage of the IRT formula scores was about 10% in the first quartile and diminished to 1-2% in the remaining quartiles. The IRT formula scores improved the accuracy of test scores in the lower tail of a test-score distribution, an area of much interest in current testing programs. What constitutes improvement in test score accuracy is somewhat dependent on the application. However, the IRT formula scores provided an increase in reliability in the neighborhood of lower proficiency. This is precisely the score range where many assessment programs are struggling with the question of how to improve measurement.

To evaluate potential information loss due to the Taylor approximation, the IRT estimated true score $\hat{C}_T$ was predicted to be a better estimate of the corrected true score C_T, compared to the IRT estimate of formula scores $\hat{C}_{T1}$ and $\hat{C}_{T2}$. Results, however, from bias statistics in the first quartile did not match the expectation, possibly due to the use of EAP estimates of theta which have some degree of regression to the mean. On the other hand, results from inspection of RMSE and r^2 were as expected: $\hat{C}_T$ was the best estimator of C_T. When the comparisons were made for the full distribution, $\hat{C}_T$ tracked C_T closely in every aspect. Both $\hat{C}_{T1}$ and $\hat{C}_{T2}$ were only slightly less efficient than $\hat{C}_T$. Therefore, it was concluded that not much information was lost due to the Taylor approximation.

The use of the pseudo-Bayes and conditional probability procedures resulted in little information loss. In the second simulation study, a moderate-sized sample was randomly selected from the first simulation study. It was expected that the IRT formula scores based on estimated IRT parameters would perform better in bias, RMSE, and r^2 statistics than the IRT formula scores obtained with the pseudo-Bayes and conditional probability methods. The latter however had smaller bias on both test sets. This result is possibly an artifact of overfit because sample statistics rather than population estimates were used to construct the IRT formula scores. With EAP estimation, the resulting $\hat{\theta}$ regresses partially to zero. Conversely, the IRT-based formula scores were calculated with observed item responses, which preserved more sample information.

Given the goal of increasing reliability in light of guessing while substantially reducing bias, the IRT formula scores appear to provide a potentially useful tool. If a 3PL framework is accepted, the order of preference for score type would be: pattern-scored θ, sample-based IRT formula scores, and number-correct score.

Application to DIF Analyses

The third goal of this dissertation was to demonstrate the potential application of IRT formula scores to DIF analyses. Because IRT formula scores can be obtained without reference to IRT parameter estimates, they have a potential use in large-scale programs that use number-correct scores for secondary analyses such as DIF. In fact, with applied and conditioning scores for the MH and the LR DIF analyses, the IRT formula scores decreased bias in both the average LOR and the average type 1 error compared to the number-corrected score by about 5.6% and 9.6%, respectively. Both the LOR and the type 1 error rates for the MH were slightly lower than those for the LR procedure. The one-term Bayes formula score C_{T1B} showed the most improvement in reducing bias under different conditions. It was shown that item discrimination and group separation still influenced both LOR and type 1 error, but less so than for the number-correct score. These results are consistent with the finding from several previous studies (Tian, 1999; Zwick et al. 1997).

Educational Importance of the Study

The number-correct score remains an operational aspect of many assessment programs. One reason is because of its communicative value to students, parents, and teachers. However, it can be a misleading measure of examinee proficiency level because it does not account for guessing. The IRT formula scores adjust for unexpectedly correct item responses. One obstacle for IRT scoring is that a more proficient examinee will receive more credit for a correct answer to a particular item than a less proficient examinee. This equity issue regarding 3PL scoring may draw diverse reactions from test users. In fact, it is the complexity of the IRT score interpretation that limits its value in testing programs.

Index G. Nevertheless, the IRT formula scores present a potentially useful tool in psychometric research. In the first IRT formula score, η_i represents a correction factor for each item; and $\eta_i\left(1-\eta_i\right)^{-1}$ represents correction factor for the second IRT formula score.

At the student level, for each examinee, index G is defined as the sum of correction factors across correct responses and can be considered as overall guessing effect for the examinee. These indices can distinguish that certain examinees likely benefitted more from guessing. For example, Table 5-1 demonstrates item responses, number-correct score, and index G level for three examinees.

Table 5-1

Difference among Index G at Student Level

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	R	G*
E1	1	1	1	1	0	0	4	S
E2	1	1	1	0	1	0	4	M
E3	1	1	0	0	1	1	4	L

Note. * *S: Small; M: Medium; L: Large.*

Assume item difficulty level for item 1 to item 6 ranges from easy to difficult. The three students received same number-correct score; however, their answering patterns were very different. There examinees would receive same score under with the number-correct score, classical formula score, or 1PL latent score. Yet, different 3PL IRT formula scores were found for these examinees. Comparing their item responses, examinee 3 answered the easiest and the hardest item correctly but was wrong on the moderate difficult items; examinee 1 was right on easy to moderate difficult items but not the difficult items;

whereas examinee 2 had mixed item responses. Computation of index G (assuming the availability of a *c* estimate) shows that examinees who answered very difficult items would have larger index G. Based on the new IRT formula scores, if the item is very difficult, the probability of answering incorrectly is greater than the probability of providing a correct response. In this case, the potential impact of guessing is higher than it would be for an easier item. Therefore, index G provides the researchers additional information about the potential guessing behavior of a student. Index G would also provide teachers with more information about students' proficiency so that they may better distinguish between students apparently having the same number-correct score.

On the item level, index G can be viewed as measure of overall guessing intensity for the item and is calculated as the sum of correction factor across correct responses of students with the same number-correct score. For instance, in Table 5-2, suppose item1 and item 2 had the same item difficulty, it is expected that the proportions of correct response for these item are the same for students with same number-correct scores (in this dissertation, item discrimination and guessing parameter were both set constant). However, from Table 5-2, the item responses were different, and item 2 resulted in a higher value on index G then the item 1. This may be an indicator that examinees tended to guess on item 2 more than to guess on item 1, and this may be due to the design of the item. Therefore, potentially, index G could be used in classical item analysis packages as a quality indicator for test items.

Table 5-2

Difference among Index G at Item Level

	Item 1	Item 2	Item …	R
E1	1	0	…	4
E2	0	1	…	4
E3	0	1	…	4
G*	S	L		

Note. In this study, item discrimination parameter and guessing parameter are set constant for all items.

** S: Small; M: Medium; L: Large.*

Summing up, as for psychometric value, these indices G could help researchers to refine studies examining the characteristics of guessers as well as to flag lower-quality items in test development.

Limitations and Future Research

Guessing imparts a type of unreliability to test scores that is different from random measurement error. This can result in statistical bias in analyses using number-correct scores. Although the IRT formula scores more closely estimated true scores in first quartile across different combinations of item parameters, and decreased both LOR and Type 1 error in DIF analyses, they should still be applied with some caution until their properties can be empirically validated across a wider domain of measurement data including size and variability of the guessing parameter *c*, sample size, and combinations of item parameters (*a*, *b*, and *c*).

Future research may also help clarify how the choice of a common *c* value can be made, and this choice is necessary for a number of statistical procedures (e.g., SIBTEST). In this study, a common random guessing parameter *c* was necessary for creating the SIM test, but a reasonable guess, which might be different from the actual *c* values for a particular test. It is not yet known how close the guess needs to be for bias reduction to occur.

The variability of the *a* parameter in the correction equation was ignored in devising IRT formula scores, though there is no theoretical barrier to including them in the equations. This assumption was made primarily to obtain practical estimators. The assumption is also implicit in the classical formula score. In fact, the *a* parameters are likely to have an effect on the accuracy of the IRT formula scores, and could possibly be incorporated in large-scale programs in which the *a* parameters are available.

In this dissertation, only main effects but not interaction effects were included in regression analyses used to predict bias in LORs and type 1 error rates. In a study by Uttaro and Millsap (1994) in which no-DIF conditions were evaluated, a significant interaction was found between item discrimination and average group ability distribution, and an interaction was also found between guessing parameter and average group ability.

Additionally, to make two test sets comparable, test length, average item parameters of two tests were chosen similarly in the present study. However, according to Uttaro and Millsap (1994), type 1 error rate decreases when the test length increases. Therefore, more study would be useful for evaluating test length and interaction effects.

Many DIF studies not only reported type 1 error, but also reported results from power analyses (Finch, 2005; Jodoin and Gierl, 2001; Kristjansson et.al., 2005). In this study, only the no-DIF was examined. Future DIF research can be designed to investigate both DIF and no-DIF conditions by using the IRT formula scores, with corresponding examination of both type 1 error and power. From the examinee's point of view, the presence of type 2 errors seems to be a more serious problem. Therefore, it would be interesting to examine whether the application of the new IRT formula scores in DIF analysis improves power.

Another limitation of this study was the use of structured item parameters to simulate test data. A set of item parameters from the existing ART test was obtained in this study to compare the IRT formulas across the actual and structured parameters. Though the results did not show much different between the two sets of item parameters in the simulation, simulated test data probably does not reflect the unique examinee quality of real test data, and additional applications to a number of real test data sets would be useful for understanding how these IRT formula scores work in an operational context.

This dissertation intended to clarify proficiency estimation under the IRT 3PL models, and then to derive a new scoring approach to correct for guessing. Two new IRT formula scores were developed that can be used use with observed-score data. Although restrictions and limitations exist as addressed above, the results included in this investigation may provide a new perspective as well as new tools for evaluating test scores. It is hoped that future research will overcome these limitations to improve this method and provide a more accurate true score estimation.

Appendix A

First, express the likelihood as a function of the common c parameter:

$$\begin{aligned} F(c) &= \ln \prod_{i=1}^{n} \lambda_i^{u_i} (1-\lambda_i)^{1-u_i} \\ &= \sum_{i=1}^{n} u_i \ln \lambda_i + (1-u_i) \ln(1-\lambda_i) \ . \end{aligned} \tag{1A}$$

where

$$\lambda_i \left(u_i = 1 \middle| \theta, a_i, b_i, c_i \right) = c + (1-c) P_i \tag{2A}$$

and

$$P_i = \frac{\exp\left[D a_i \left(\theta - b_i \right) \right]}{1 + \exp\left[D a_i \left(\theta - b_i \right) \right]} . \tag{3A}$$

Then the standard Taylor M-term power expansion is then obtained by

$$H(c) = F(0) + F_{c=0}^{(1)} \cdot c + \frac{1}{2!} F_{c=0}^{(2)} \cdot c^2 + \frac{1}{3!} F_{c=0}^{(3)} \cdot c^3 + \ldots + \frac{1}{m!} F_{c=0}^{(m)} \cdot c^m \ . \tag{4A}$$

Let

$$\begin{aligned} \frac{\partial \lambda_i}{\partial c} &= \frac{\partial}{\partial c} \left[c + (1-c) P_i \right] \\ &= 1 - P_i = Q_i \ . \end{aligned} \tag{5A}$$

It follows that the first derivative $F_c^{(1)}$ is

$$\begin{aligned} F_c^{(1)} &= \sum_{i=1}^{n} \left[\frac{u_i}{\lambda_i} \frac{\partial \lambda_i}{\partial c} + \frac{\left(1-u_i\right)}{(1-\lambda_i)} \frac{\partial (1-\lambda_i)}{\partial c} \right] \\ &= \sum_{i=1}^{n} Q_i \left[\frac{u_i}{\lambda_i} - \frac{\left(1-u_i\right)}{(1-\lambda_i)} \right] . \end{aligned} \tag{6A}$$

At $c = 0$,

$$\begin{aligned} F_{c=0}^{(1)} &= \sum_{i=1}^{n}\left[\frac{u_i}{P_i}Q_i - \frac{\left(1-u_i\right)}{Q_i}Q_i\right] \\ &= \left(\sum_{i=1}^{n} u_i \frac{Q_i}{P_i}\right) - W . \end{aligned} \tag{7A}$$

The second derivative $F_c^{(2)}$ is

$$\begin{aligned} F_c^{(2)} &= -\sum_{i=1}^{n} Q_i \left[\frac{u_i}{\lambda_i^2}\frac{\partial \lambda_i}{\partial c} - \frac{\left(1-u_i\right)}{(1-\lambda_i)^2}\frac{\partial (1-\lambda_i)}{\partial c}\right] \\ &= -\sum_{i=1}^{n} Q_i^2 \left[\frac{u_i}{\lambda_i^2} - \frac{\left(1-u_i\right)}{(1-\lambda_i)^2}\right] \end{aligned} \tag{8A}$$

at c=0,

$$\begin{aligned} F_{c=0}^{(2)} &= -\sum_{i=1}^{n} Q_i^2 \left[\frac{u_i}{\lambda_i^2} - \frac{\left(1-u_i\right)}{(1-\lambda_i)^2}\right] \\ &= -\left[\sum_{i=1}^{n}\left(u_i \frac{Q_i^2}{P_i^2}\right) - W\right]. \end{aligned} \tag{9A}$$

The third derivative $F_c^{(3)}$ is

$$\begin{aligned} F_c^{(3)} &= \sum_{i=1}^{n} Q_i^2 \left[\frac{u_i}{\lambda_i^3}\frac{\partial \lambda_i}{\partial c} - \frac{\left(1-u_i\right)}{(1-\lambda_i)^3}\frac{\partial (1-\lambda_i)}{\partial c}\right] \\ &= \sum_{i=1}^{n} Q_i^3 \left[\frac{2u_i}{\lambda_i^3} - \frac{2\left(1-u_i\right)}{(1-\lambda_i)^3}\right] \end{aligned} \tag{10A}$$

at c=0,

$$\begin{aligned} F_{c=0}^{(3)} &= \sum_{i=1}^{n} Q_i^3 \left[\frac{2u_i}{\lambda_i^3} - \frac{2\left(1-u_i\right)}{(1-\lambda_i)^3}\right] \\ &= 2\left[\sum_{i=1}^{n}\left(u_i \frac{Q_i^3}{P_i^3}\right) - W\right]. \end{aligned} \tag{11A}$$

The forth derivative $F_c^{(4)}$ is

$$\begin{aligned} F_c^{(4)} &= -\sum_{i=1}^{n} Q_i^3 \left[\frac{u_i}{\lambda_i^4} \frac{\partial \lambda_i}{\partial c} - \frac{(1-u_i)}{(1-\lambda_i)^4} \frac{\partial (1-\lambda_i)}{\partial c} \right] \\ &= -\sum_{i=1}^{n} Q_i^4 \left[\frac{6u_i}{\lambda_i^4} - \frac{6(1-u_i)}{(1-\lambda_i)^4} \right] \end{aligned} \qquad (12A)$$

at c=0,

$$\begin{aligned} F_{c=0}^{(4)} &= -\sum_{i=1}^{n} Q_i^4 \left[\frac{6u_i}{\lambda_i^4} - \frac{6(1-u_i)}{(1-\lambda_i)^4} \right] \\ &= -6\left[\sum_{i=1}^{n} \left(u_i \frac{Q_i^4}{P_i^4} \right) - W \right]. \end{aligned} \qquad (13A)$$

According to this mathematical pattern, the M-term expansion of $F(c)$ at $c = 0$ can be shown as

$$\begin{aligned} H(c) = F(0) &+ \left[\sum_{i=1}^{n} \left(u_i \frac{Q_i}{P_i} \right) - W \right] c - \frac{1}{2}\left[\sum_{i=1}^{n} \left(u_i \frac{Q_i^2}{P_i^2} \right) - W \right] c^2 \\ &+ \frac{1}{3}\left[\sum_{i=1}^{n} \left(u_i \frac{Q_i^3}{P_i^3} \right) - W \right] c^3 - \frac{1}{4}\left[\sum_{i=1}^{n} \left(u_i \frac{Q_i^4}{P_i^4} \right) - W \right] c^4 \\ &+ \ldots - \frac{1}{m}\left[\sum_{i=1}^{n} \left(u_i \frac{Q_i^m}{P_i^m} \right) - W \right] c^m. \end{aligned} \qquad (14A)$$

Next, maximizing $H(c)$ with respect to θ refers to differentiate $F_{c=0}^{(1)}$, $F_{c=0}^{(2)}$ … with respect to θ yields

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\left[\left(\sum_{i=1}^{n}u_i\right)\frac{Q_i}{P_i}-W\right]=\sum_{i=1}^{n}u_i\,\frac{\partial}{\partial\theta}\left(\frac{Q_i}{P_i}\right)=-D\sum_{i=1}^{n}a_iu_i\,\frac{Q_i}{P_i},\\
&\frac{\partial}{\partial\theta}\left[\sum_{i=1}^{n}\left(u_i\,\frac{Q_i^{\,2}}{P_i^2}\right)-W\right]=\sum_{i=1}^{n}u_i\,\frac{\partial}{\partial\theta}\left(\frac{Q_i^{\,2}}{P_i^2}\right)=-2D\sum_{i=1}^{n}a_iu_i\left(\frac{Q_i}{P_i}\right)^2,\\
&\frac{\partial}{\partial\theta}\left[\sum_{i=1}^{n}\left(u_i\,\frac{Q_i^{\,3}}{P_i^3}\right)-W\right]=\sum_{i=1}^{n}u_i\,\frac{\partial}{\partial\theta}\left(\frac{Q_i^{\,3}}{P_i^3}\right)=-3D\sum_{i=1}^{n}a_iu_i\left(\frac{Q_i}{P_i}\right)^3,\\
&\mathit{and}\\
&\frac{\partial}{\partial\theta}\left[\sum_{i=1}^{n}\left(u_i\,\frac{Q_i^{\,4}}{P_i^4}\right)-W\right]=\sum_{i=1}^{n}u_i\,\frac{\partial}{\partial\theta}\left(\frac{Q_i^{\,4}}{P_i^4}\right)=-4D\sum_{i=1}^{n}a_iu_i\left(\frac{Q_i}{P_i}\right)^4.
\end{aligned}\tag{15A}$$

To solve $\sum_{i=1}^{n}P_i$, set the result equal to zero

$$\frac{\partial}{\partial\theta}H(c)=\frac{\partial}{\partial\theta}\left\{\begin{array}{l}F(0)+F_{c=0}^{(1)}\cdot c+\dfrac{1}{2!}F_{c=0}^{(2)}\cdot c^2\\[2ex] +\dfrac{1}{3!}F_{c=0}^{(3)}\cdot c^3-\dfrac{1}{4!}F_{c=0}^{(4)}\cdot c^4+\ldots-\dfrac{1}{m!}F_{c=0}^{(m)}\cdot c^m\end{array}\right\}=0.\tag{16A}$$

This results in

$$\begin{aligned}
&\sum_{i=1}^{n}a_i(u_i-P_i)-c\sum_{i=1}^{n}a_iu_i\,\frac{Q_i}{P_i}+c^2\sum_{i=1}^{n}a_iu_i\left(\frac{Q_i}{P_i}\right)^2\\
&-c^3\sum_{i=1}^{n}a_iu_i\left(\frac{Q_i}{P_i}\right)^3+c^4\sum_{i=1}^{n}a_iu_i\left(\frac{Q_i}{P_i}\right)^4-\ldots+c^m\sum_{i=1}^{n}a_iu_i\left(\frac{Q_i}{P_i}\right)^m=0
\end{aligned}\tag{17A}$$

Assume $a_i=1\quad\forall i$, then

$$\begin{aligned}
&\sum_{i=1}^{n}(u_i-P_i)-c\sum_{i=1}^{n}u_i\,\frac{Q_i}{P_i}+c^2\sum_{i=1}^{n}u_i\left(\frac{Q_i}{P_i}\right)^2\\
&-c^3\sum_{i=1}^{n}u_i\left(\frac{Q_i}{P_i}\right)^3+c^4\sum_{i=1}^{n}u_i\left(\frac{Q_i}{P_i}\right)^4-\ldots+c^m\sum_{i=1}^{n}u_i\left(\frac{Q_i}{P_i}\right)^m=0
\end{aligned}\tag{18A}$$

Because $u_i=u_i^2=\ldots=u_i^n$, and let $\eta_i=c\left(\frac{Q_i}{P_i}\right)$, the equation can be simplified as

$$\sum_{i=1}^{n}(u_i-P_i)-\sum_{i=1}^{n}u_i\eta_i+\sum_{i=1}^{n}u_i\eta_i^2-\sum_{i=1}^{n}u_i\eta_i^3+\sum_{i=1}^{n}u_i\eta_i^4-\ldots+\sum_{i=1}^{n}u_i\eta_i^m=0\tag{19A}$$

and

$$\begin{aligned}\sum_{i=1}^{n} P_i &= \sum_{i=1}^{n} u_i - \sum_{i=1}^{n} u_i \eta_i + \sum_{i=1}^{n} u_i^2 \eta_i^2 - \sum_{i=1}^{n} u_i^3 \eta_i^3 + \sum_{i=1}^{n} u_i^4 \eta_i^4 - \ldots + \sum_{i=1}^{n} u_i^m \eta_i^m \\ &= \sum_{i=1}^{n} u_i \left(1 - \eta_i + \eta_i^2 - \eta_i^3 + \eta_i^4 - \ldots + \eta_i^m\right).\end{aligned} \tag{20A}$$

When $m \to \infty$,

$$\begin{aligned}\lim_{m\to\infty}\left(1 - \eta_i + \eta_i^2 - \eta_i^3 + \eta_i^4 - \ldots + \eta_i^m\right) &= \lim_{m\to\infty}\left(1 + \sum_{m=1}^{\infty}\left(-\eta_i\right)^m\right) \\ &= \frac{1}{1+\eta_i}.\end{aligned} \tag{21A}$$

Therefore,

$$\sum_{i=1}^{n} P_i = \sum_{i=1}^{n} u_i \left(1+\eta_i\right)^{-1}. \tag{22A}$$

REFERENCES

Abu-Sayf, F. K. (1979). The scoring of multiple-choice tests: A closer look. *Educational Technology, 19,* 5-15.

Albanese, M. A. (1988). The projected impact of the correction for guessing on individual scores. *Journal of Educational Measurement 25(2),* 149-157.

Alnabhan, M. (2002). An empirical investigation of the effects of three methods of handling guessing and risk taking on the psychometric properties of a test. *Social Behavior and Personality, 30,* 645-652.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96-116). Baltimore: Johns Hopkins University Press.

Angoff, W. H. (1989). Does guessing really help? *Journal of Educational Measurement 26(4),* 323-336.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp.3-30). Hillsdale, NJ: Lawrence Erlbaum.

Angoff, W. H., & Schrader, W. B. (1984). A study of hypotheses basic to the use of rights and formula scores. *Journal of Educational Measurement, 21(1),* 1-17.

Avila, C., & Torrubia, R. (2004). Personality, expectations, and response strategies in multiple-choice question examinations in university students: A test of Gray's hypothesis. *European Journal of Personality, 18,* 45-59.

Ben-Shakhar, G. & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement, 28(1),* 23-35.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores (chapters 17-20). Reading, MA: Addison-Wesley.

Bliss, L. B. (1980). A test of Lord's assumption regarding examinee guessing behavior on multiple choice tests using elementary school children. *Journal of Educational Measurement, 17,* 147-153.

Bridgeman, B., & Schmitt, A. (1997). Fairness issue in test development and administration. In W. W. Willingham, & N. S. Cole (Eds.), *Gender and fair assessment* (pp. 185-226). Mahwah, NJ.: Lawrence Erlbaum.

Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: a decision-theoretic view of formula scoring. *Journal of Educational Measurement, 30 (4),* 277 - 291.

Burton, R. F. (2002). Misinformation partial knowledge and guessing in true/false tests. *Medical Education, 36,* 805-811.

Burton, R. F. (2005). Multiple-choice and true/false tests: Myths and apprehensions. *Assessment and Evaluation in Higher Education, 30,* 65-72.

Burton, R. F., & Miller, D. J. (1999). Statistical Modelling of Multiple-Choice and True/False Tests: Ways of Considering, and of Reducing, the Uncertainties Attributable To Guessing. *Assessment and Evaluation in Higher Education, 24 (4),* 399-411.

California Department of Education (2006). California standards tests: Technical report spring 2005 administration. Retrieved July 7, 2006 from http://www.cde.ca.gov/ta/tg/sr/documents/startechrpt05.pdf.

Camilli, G. (2006) Test fairness. In R. L. Bernnan (Ed.), *Educational measurement (4th ed.)* (pp. 221-256). Westport, CT.:Praeger.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based on Mantel-Haenszel statistics. *Journal of Educational Measurement, 34(2),* 123-139.

Camilli, G. & Shepard, L. A. (1994) *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Chevalier, S. A. (1998). *A review of scoring algorithms for ability and aptitude tests.* Paper presented at the annual meeting of the Southwestern Psychological Association, New Orleans, LA, April.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201-219). New York: Macmillan.

Cronbach, L.J. (1984). *Essentials of psychological testing (4th edn),* New York: Harper Row.

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

Davis, F. B. (1967). A note on the correction for chance success. *Journal of Experimented Education, 3,* 43-47

Diamond, J., & Evans, W. (1973). The correction for guessing. *Review of Educational Research 43,* 181-191.

Dimitrov, D.M. (2004). *Ability re-estimation in the Rasch model.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, April.

Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement, 23(4),* 283-298.

Dodden, H. (2004). The Relationship Between Item Parameters and Item Fit. *Journal of Educational Measurement, 41(3),* 261-270.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In: P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp.35-66). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N. J., & Kulick, E. M. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement 23,* 355-368.

Ebel, R. L. (1972). *Essentials of educational measurement.* New York: Prentice-Hall, (p. 252).

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380-396.

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

Ferrara, S., & DeMauro, G. E. (2006). Standardized assessment of individual achievement in K-12. In R. L. Bernnan (Ed.), *Educational measurement (4th ed.)* (pp. 579-621). Westport, CT.:Praeger.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29* (4), 278-295.

Frary, R. B. (1988). *Formula scoring of multiple-choice tests (Correction for guessing)* NCME Instructional Topics in Educational Measurement.

Grandy, J. (1987). *Characteristics of examinees who leave questions unanswered on the GRE General Test under rights-only scoring (ETS RR-87-38)*. Princeton, NJ: Educational Testing Service.

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

Hauser, C., & Kingsbury, G. (2004). Differential item functioning and differential test functioning in the Idaho Standards Achievement Tests for spring 2003. Retrieved August 16, 2008 from http://www.nwea.org/research/getreport.asp?reportID=3.

Higham, P. A. (2007). No special K! A signal detection framework for the strategic regulation of memory accuracy. *Journal of Experimental Psychology: General, 136*, 1-22.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenzel procedure. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Holzinger, K. J. (1924). On scoring multiple-response tests. *Journal of Educational Psychology, 15,* 445-447.

Hopkins, K., & Stanley, J. (1981). *Educational and psychological measurement and evaluation. (6th ed.)*. Englewood, NJ.: Prentice Hall.

Huang, C. Y. (1998). *Factors influencing the reliability of DIF detection methods.* Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998.

Hunter, J. E. (1975). *A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items.* Paper presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type 1 error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329-349.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65 (6),* 935-953.

Kurz, T. B. (1999). *A review of scoring algorithms for multiple-choice tests.* Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX.

Li, H.-H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61,* 647-677.

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75 (2),* 164-174.

Lord, F. M. (1963). Formula scoring and validity. *Educational and Psychological Measurement, 23,* 233-239.

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement, 12,* 7-12.

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

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Mazor, K. M., Kanjee, A., & Clauser B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Education Measurement, 32,* 131-144.

Meijer, R. R. & Nering, M. L. (1999). Computerized adaptive testing: overview and introduction. *Applied Psychological Measurement, 23 (3),* 187-194.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105-118.

Muijtjens, A.M.M., Mameren, H., Hoogenboom, R.J.I., Evers, J.H.L. & van der Vleuten, C.P.M. (1999). The effect of a 'don't know option' on test scores: Number-right and formula scoring compared. *Medical Education, 33,* 267-275.

Muraki, E. & Bock, R. D. (2003). PARSCALE version 4.1 [Computer Program]. Lincolnwood, IL: Scientific Software International, INC.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and Simultaneous Item Bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18,* 315-328.

Narayanan P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20,* 257-274.

New York State Education Department (2005). NYS testing program mathematics Grade 8 technical report. Retrieved August 16, 2008 from http://www.emsc.nysed.gov/osa/pub/gr8math05report.pdf.

Penfield, R. D., & Camilli, G., (2007). Differential item functioning and item bias. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics 26* (pp. 125-167). Amsterdam, The Netherlands: Elsevier.

Puhan, G., & Gierl, M.J. (2006). Evaluating the effectiveness of two-stage testing on English and French versions of a science achievement test. *Journal of Cross-Cultural Psychology, 37*, 136-154.

Roberts, Dennis (1995). Let's talk about the "correction for guessing" formula. Retrieved August 13, 2008, from http://www.personal.psu.edu/users/d/m/dmr/papers/CORR4GUS.pdf

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17,* 105-116.

Roussos, L. A., Schnipke, D. L., & Pashley, P. J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning. *Journal of Educational and Behavioral Statistics, 24(3),* 293-322.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355-371.

Rowley, G., & Traub, R. E. (1977). Formula scoring, number-right Scoring, and test-taking strategy. *Journal of Educational Measurement, 14(1),* 15-22.

San Martín, E., del Pino, G., & De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement, 30,* 183-203.

SAS Institute Inc. (2003). SAS/BASE software: Release 9.1 manuals. Cary, NC: author.

Scharf, E. M., & Baldwin, L. P. (2007). Assessing multiple choice question (MCQ) tests – a mathematical perspective. *Active Learning in Higher Education, 8(1),* 31-47.

Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias. In: P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp.197-239). Hillsdale, NJ: Lawrence Erlbaum.

Shepard, L. A. (1981). Identifying bias in test items. In B. F. Green (Ed.), *New direction in testing and measurement: Issues in testing-Coaching, disclosure and test bias,* No. 11 (pp. 79-104). San Francisco: Jossey-Bass.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Tian, F. (1999). *Detecting differential item functioning in polytomous items.* Unpublished doctoral dissertation, Faculty of Education, University of Ottawa.

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure the detection of differential item functioning. *Applied Psychological Measurement, 18,* 15-25.

Van Den Wittenboer, G., Hox, J. J., & De Leeuw, E. D. (2000). Latent class analysis of respondent scalability. *Quality & Quantity, 34,* 177-191.

Veerkamp, W. J. J. (2000). Taylor approximations to logistic IRT models and their use in adaptive testing. *Journal of Educational and Behavioral Statistics, 25(3),* 307-343.

Wainer, H., & Wright, B.D. (1980). Robust estimation of ability in the Rasch model. *Psychometrika, 45,* 373-391.

Waller, M.I. (1989). Modeling guessing behavior: A comparison of two IRT models. *Applied Psychological Measurement, 13,* 233-243

Willingham, W. W., & Cole, N. S. (1997). Introduction. In W. W. Willingham, & N. S. Cole (Eds.), *Gender and fair assessment* (pp. 1-15). Mahwah, NJ.: Lawrence Erlbaum.

Wise, S. L., Bhola, D. S., & Yang, S. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice, 25(2),* 21-30.

Wise, S. L., & DeMars, C.E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43,* 19-38.

Xie, Y. (2002). An application of a special two-class item response model using Markov chain Monte Carlo method. Graduate School of Education: University of California at Berkeley. Downloaded [6/4/2006] from http://bear.berkeley.edu/Publications/PP2--Yiyu.pdf.

Yamamoto, K. (1989). *A hybrid model of IRT and latent class models (ETS RR-89-41).* Princeton, NJ: Educational Testing Service.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Bernnan (Ed.), *Educational measurement (4th ed.)* (pp. 111-154). Westport, CT.:Praeger.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In: P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp.337-347). Hillsdale, NJ: Lawrence Erlbaum.

Zwick, R., Thayer, D., & Mazzeo, J. (1997). Descriptive and Inferential Procedures for Assessing Differential Item Functioning in Polytomous Items. *Applied Measurement in Education, 10*(4), 321-344.

Curriculum Vita

TING-WEI CHIU

Education

Rutgers University, New Brunswick, NJ

Ph.D. Education - Concentration in Statistics & Measurement	*May, 2010*
M.Ed. Educational Statistics, Measurement, and Evaluation	*May, 2002*
M.S. Statistics	*January, 2002*

Fu-Jen Catholic University, Taipei, Taiwan

B.S. Applied Psychology with minor in Business Administration	*June, 1998*

Professional Experience

Rutgers University, New Brunswick, NJ

Graduate Assistant – Office of Institutional Research and Academic Planning	*09/08-*
Graduate Assistant – Department of Sociology	*09/07-06/08*
Graduate Assistant – The Center for Educational Policy Analysis	*09/04-06/06*
Graduate Assistant – Teaching Excellent Center	*09/03-06/04*

Rutgers University, Newark, NJ

Statistic Consultant – Rutgers Learning Center	*02/07-08/07*

Ministry of Education, Taipei, Taiwan

Research Associate	*02/03-08/03*

Fu-Jen Catholic University, Taipei, Taiwan

Research Assistant – Department of Applied Psychology	*08/98-07/99*

Taipei City Psychiatric Center, Taipei, Taiwan

Clinical Psychologist, Intern/Practicum	*06/97-08/97*

Research and Publications

Chiu, T-W (2010). Correction for Guessing in the Framework of the 3PL Item Response Theory Model. *Doctoral dissertation.*

Chiu, T-W., & Camilli, G. (2010). A New IRT 3PL-based Correction for Guessing Method. The annual meeting of National Council of Measurement in Education, Denver, CO, April, 2010.

Camilli, G., Prowker, A., Dossey, J., Lindquist, M., **Chiu, T-W**., Vargas, S., & de la Torre, J. (2008). Summarizing Item Difficulty Variation with Parcel Scores. *Journal of Educational Measurement, 45(4),* 363-389.

Chiu, T-W. & Camilli, G. (2008). New IRT-based Corrections for Guessing and Applications in DIF Analyses. Poster was presented at the annual meeting of National Council of Measurement in Education, New York, March, 2008.

Camilli, G., Prowker, A., & **Chiu, T-W**. (2006). Value-Added Information Derived From Multilevel Rasch Models. Paper was presented at the annual meeting of American Educational Research Association, San Francisco, April, 2006.