University of Alberta

Four Scoring Procedures for High-Stakes and Low-Stakes Tests with Constructed-Response and Selected-Response Item Formats

by ©

Denise Marie Nowicki

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

Doctor of Philosophy

Department of Educational Psychology

Edmonton, Alberta
Spring 2008

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-45575-3
Our file Notre référence
ISBN: 978-0-494-45575-3

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

This study examined the interchangeability of scores yielded by four scoring procedures advanced in the literature (Schaeffer, Henderson-Montero, Julian, & Bene, 2002; Sykes & Hou, 2003) when applied at the group level and student level to low-stakes achievement tests and to high-stakes school leaving examinations containing both selected response (SR) items and constructed response (CR) items. The four scoring procedures include the unweighted procedure in which scores from the set of SR items and the set of CR items/tasks are simply added (UNW); the weighted procedure in which the CR items are given a weight of two while the SR items are weighted one (WCRX2), the weighted procedure in which the CR items are weighted so that they contribute as much to the total scores as the SR items (WN/M), and pattern scores yielded by an Item Response Analysis of the full test.

Descriptive statistics including means, standard deviations of the raw scores, item-test correlations, and reliability for the SR and CR items were calculated on two random samples of 2,000 students from each of the 2002-2003 Alberta English 9 and Mathematics 9 provincial achievement tests and the English 30 and Pure Math 30 provincial school leaving diploma examinations. PARDUX and WINFLUX were used to estimate parameters and place the item parameters on a common score scale. The interchangeability of scores yielded by the four scoring procedures was evaluated at the group and student level using a difference that matters (DTM) and by the magnitude of the standard errors.

The results reveal that 1) the descriptive analyses were stable across samples thus no notable differences were noted between the four scoring procedures at the group level, 2) differences were noted at the student level: pattern scoring generally had the lowest SEs and had the greatest differences at the 10^{th} and 90^{th} percentiles, pattern scoring also resulted with the greatest number of students affected at the four proficiency levels; and the differences in individual student scaled scores were most pronounced when pattern scoring was involved, and 3) results appear to be a function of the raw score weight of the SR and CR items.

It was concluded that, 1) at the group level, the four scoring procedures yielded similar results on all four tests, 2) at the student level, the four scoring procedures did not yield scale score distributions that were sufficiently similar to warrant using the procedures interchangeably, 3) pattern scoring provided the smallest standard errors of the four scoring procedures, particularly at the lower end of the ability distribution, 4) stakes was not a factor affecting the four scoring methods, 5) subject is a factor affecting the scale score distributions, and 6) the four scoring methods can be used for norm referenced without bias. However, the four scoring methods result in different student scale scores and thus would not be appropriate for criterion-referenced testing situations like those used by Alberta Education.

As a result, student scores and ultimately decisions made based on those scores may be affected. This can potentially harm students in that their opportunity for graduation and scholarship may be altered depending on which scoring procedure is used. As such, researchers and government officials should carefully consider the implications of which scoring procedure is chosen for each particular test and examination. Recommendations for further research are provided.

ACKNOWLEDGEMENTS

The completion of this document has been a long and arduous journey. There were many days I thought I would never see this thesis completed. It would not have been, if it were not for some very special people who guided, supported and/or goaded me along the way. I would like to take this space to express my sincere gratitude to those without whom, this thesis would not have been possible.

A very special thank you to my amazing husband Colin Nowicki, who would not let me give up – even when I really, really wanted to. To my incredibly patient children, Zachary and Nikolas Nowicki, who have never known a Mommy who did not have to "work for Uncle Todd". To my ever supportive parents Clarence and Evelyn Charchuk (see Dad, the Charchuk name did appear in the thesis, twice!) who were supportive in so many ways throughout this journey. To my dear friend Antoinette Marais, without who's moral and technical support I would not have been able to complete this thesis.

I would also like to thank Dr. W. Todd Rogers for supporting my completion of this thesis. I am also thankful to Drs. Robert M. Klassen, Jeffrey Bisanz, Janice Causgrove-Dunn, and Ying Cui from the University of Alberta as well as Dr. Delwyn L. Harnisch from the University of Nebraska, Lincoln for their willingness to serve on my thesis committee. Their comments and suggestions contributed greatly to the final product of my work.

Finally, my most sincere thanks to God for the many blessing He has given.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Context

Standardized, large-scale assessment has escalated in the past 10 to 20 years and is prominent in many provinces and all 50 states in the United States (Alberta Education, 2004a; British Columbia Education, 2003; Phelps, 2000; Tindal, 2002). These testing programs are increasingly complex and incorporate a number of subject areas and testing formats, including multiple-choice, constructed-response, and various other performance assessments (Tindal, 2002). Phelps (1998) noted that this increase in assessment has been approved, if not demanded, by the public. In an examination of 70 surveys conducted over the past 30 years regarding public perception of standardized testing, Phelps (1998) noted that "the majorities in favor of more testing, more high stakes testing, or higher stakes in testing have been large, often very large, and fairly consistent over the years, across polls and surveys and even across respondent groups" (p. 14). "In summary, large-scale testing has been on the rise and is supported by the public" (Tindal, 2002, p.1).

Not only is the incidence of large-scale assessment increasing, but so is the importance given to student assessment results (Ercikan, 2002). Assessments have become increasingly complex, designed to reflect an intended curriculum, model good instruction, challenge examinees in solving real-life problems, apply interdisciplinary skills, and report information about examinees' competencies (Ercikan, 2002). One of the ways that test developers have attempted to meet these varied goals is to include both selected-response (SR) and constructed-response (CR) items on large-scale assessments (Ercikan, Schwarz, Julian, Burket, Weber, & Link, 1998; Goldberg & Roswell, 2001; Messick, 1994; Schaeffer, Henderson-Montero, Julian, & Bene, 2002; Sykes & Hou, 2003). It is thought that by including both CR items and SR items, the benefits of both, such as objective scoring, economy, enhanced reliability of tests or subtests composed of SR items, and apparent enhanced validity due to the inclusion of CR items, are being realized (Schaefer et al., 2002; Sykes & Hou, 2003; Wainer & Thissen, 1993).

However, there is some concern about the cost versus the benefit of adding CR items (Wainer & Thissen, 1993; Rudner, 2001). In the same amount of testing time, the reliability of SR subtests or tests is considerably higher than that of CR subtests or tests. This is a result of the "objective" scoring SR items versus CR items, which are typically scored by two or more raters. Although the raters are trained and disparate scores are mediated by a third rater, the lower reliability of the CR items is often attributable to the variability of scoring due to the raters (Wainer & Thissen, 1993). Despite this, "popular notions of authentic and direct assessment have politicized the item-writing profession, particularly in large-scale settings" (Rodriguez, 2003, p. 2). Consequently, if the demand to include CR items remains, the question is "how should CR items and SR items be scored to provide the most equitable results?"

Scoring Procedures

Schaeffer et al. (2002) addressed this question by examining three different ways of scoring low-stakes Grade 9 Biology and English field tests that contained both SR and CR items:

1. *Unweighted raw score procedure*. A student's observed score was equal to the number of points earned from the SR items plus the number of points earned from the CR items.
2. *Weighted raw score procedure*. A weighting scheme designed so that the SR items and CR items contributed the same number of points toward the total score.
3. *Item response theory (IRT) pattern scoring*. Each student's score was based on a maximum-likelihood estimate derived from the student's item-response vector. This procedure used the optimal item weights in terms of item information.

The scoring procedures were compared for total group and subgroups defined in terms of gender and ethnicity. The scaled score distributions, standard errors of measurement and proficiency-level classifications were compared. Schaeffer et al. (2002) reported that the three scoring procedures yielded similar results. The score distributions and correlational patterns for the total group and the gender and ethnic subgroups they considered were comparable. Pattern scoring did provide the smallest standard errors, particularly at the lower end of the scale score distributions. This would help ensure that the scores are more accurate estimates of student ability, especially for students at the lower end of the scale.

Sykes and Hou (2003) also addressed this question. Using a procedure similar to the procedure used by Schaeffer et al. (2002), Sykes and Hou (2003) examined the effect of the scoring procedures on a low-stakes Grade 8 writing examination. In addition to the three scoring procedures used by Schaeffer et al. (2002), they added four additional scoring procedures: a weighted score that deliberately increased the weighting of the CR items by a factor of two; a summed score that involved the sum of the two raters' scores on the CR items; a long form in which 18 SR items and eight CR items were added to the examination; and an all SR item long form in which 20 additional SR items were added and the CR items were removed. Sykes and Hou (2003) found that the pattern scoring provided the smallest standard errors across the ability range of all the forms containing CR items. The solely SR long form was found to have the highest test reliability ($\hat{\alpha} = 0.90$) with the two summed CR form and the CRx2 form having the lowest reliabilities ($\hat{\alpha} = 0.84$ and 0.84, respectively).

Schaeffer et al. (2002) and Sykes and Hou (2003) focused only on low-stakes examinations at the Grade 8 and 9 levels. Low-stakes examinations are examinations in which student grades are not affected and the consequences perceived by the students are low. Students may perceive low-stakes examinations as inconsequential to their personal achievement and as a result they may not be motivated to work as hard as possible to achieve their best as they would on high-stakes examinations (DeMars, 2000; Kiplinger & Linn, 1992; Paris, Lawton, & Turner, 1992; Wolf, Smith, & Birnbaum, 1995; Wolf & Smith, 1995). Conversely, high-stakes examinations are examinations in which the consequences of performance directly affect the achievement of the students writing the examinations. Brown and Walberg (1993), DeMars (2000), Kiplinger and Linn (1993), Wolf and Smith (1995), and Wolf, Smith, and Birnbaum (1995) found that the average scores on high-stakes examinations are generally higher than the average scores on low-stakes examinations.

It has been demonstrated that test-stakes may affect performance on achievement measures with single-formats (Brown & Walberg, 1993; Kiplinger & Linn, 1993; Wolf & Smith, 1995) and multiple-formats (DeMars, 2000; Wolf, Smith & Birnbaum, 1995). However, each of the tests in the above studies was scored using one scoring procedure. It may be possible that different scoring procedures used may have differential results when applied to low-stakes and high-stakes examinations with multiple formats. This has not been addressed in the literature.

Purpose

Consequently, the purpose of this study was to examine the differences at the group level and student level, between the scores yielded by the unweighted raw scores, weighted raw scores where (i) the SR items and CR items are weighted equally and (ii) the CR items are worth twice as much as the SR items, and (iii) pattern scoring procedures when applied to low-stakes examinations and to high-stakes examinations.

To address this purpose, two low-stakes and two high-stakes examinations were used. The use of the two examinations allowed an assessment of the stability of the scores yielded by the four scoring procedures. The low-stakes and high-stakes examinations consisted of the 2003 Provincial Achievement Tests (PATs) and 2003 Diploma Examinations (DIPs) administered in Alberta. The PATs, which are administered at Grades 3, 6, and 9, are considered low-stakes tests as their main purpose is to help the province, school districts, schools, and school planning councils to evaluate how well foundational skills are being taught and to improve student achievement. In contrast, the DIPs are considered to be high-stakes examinations. Administered at the end of each Grade 12 examinable course, the DIPs are school exit examinations used to certify individual student competence. The scores from the examinations are combined with a school awarded mark and the blended marks (50% and 50%, respectively) are used to determine whether each student has passed or not passed the course and to determine scholarship winners (Alberta Education, 2004c).

The examinations in the areas of language arts and mathematics in both the low- and high-stakes levels were used in the present study. Further, the low-stakes tests were at the highest grade level (Grade 9). By doing so, the comparisons made between the tests will not be confounded by different subject matter. However, there were differences in some topics and the level or complexity of the common topics covered between the two grade levels.

Using these tests and examinations, the specific research question addressed was: Are there differences between the scores yielded by the unweighted raw scores, weighted raw scores where (i) the SR items and CR items are weighted equally and (ii) the CR items are worth twice as much as the SR items, and (iii) pattern scoring procedures when applied to low-stakes examinations and to high-stakes examinations?

As presented earlier, research findings suggested that test-stakes may affect performance on achievement measures with single-formats (Brown & Walberg, 1993; Kiplinger & Linn, 1993; Wolf & Smith, 1995) and multiple-formats (DeMars, 2000; Wolf, Smith & Birnbaum, 1995). However, each of the tests in the above studies was scored using one scoring procedure. It was also suggested that scoring procedure affects the reliability of the scores (Schaeffer et al., 2002; Sykes and Hou, 2003). It was therefore hypothesized that the four scoring procedures: unweighted raw scores, weighted raw scores where (i) the SR items and CR items are weighted equally and (ii) the CR items are worth twice as much as the SR items, and pattern scoring, will have differential results when applied to low-stakes and high-stakes examinations with combined SR and CR formats.

Definition of Terms

Constructed Response (CR) Items: Students must formulate their responses in oral or written form (Van den Bergh, 1990). Examples of CR items include short answer, sentence completion, computation, extended response, or essay.

High-Stakes Examinations: High-stakes examinations are those in which the students are motivated to perform well because the test has personal consequences to them (e.g., pass-fail decisions, placement decisions) and because the teacher has emphasized their importance (Wolf & Smith, 1995).

Item Response Theory (IRT): "IRT rests on two basic postulates: (a) The performance of an examinee on a test item can be predicted (or explained) by a set of factors called traits, latent traits, or abilities; and (b) the relationship between examinee's item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic function or item characteristic curve (ICC)" (Hambleton, Swaminathan & Rogers, 1999, p.7).

Low-Stakes Examinations: Low-stakes tests are those in which the test scores are either not provided at the individual student level, or if so, are of little or no consequence to the student or the teacher (Wolf & Smith, 1995).

Rater Bias: Constructed response items require scoring by raters who judge the quality of the examinee's response. The raters, although trained, must interpret and evaluate examinee responses to produce ratings. Their interpretations and evaluations may be affected by a variety of potential response biases that may unfairly affect their judgments regarding the quality of examinee responses (Engelhard, 2002).

Scale Scores: Using IRT to estimate scores using the information in the item responses (Thissen & Orlando, 2001).

Scoring Procedure: In traditional test theory, scoring generally involves adding up the positive responses. In IRT scoring procedures, a person is characterized by degree of ability and the item is characterized by degree of difficulty. This information in the items responses is used for scoring (Thissen & Orlando, 2001).

Selected Response (SR) Items: Students select their response from the one or more alternatives provided (Van den Bergh, 1990). Examples include multiple-choice, matching, and true-false items.

Simultaneous Scaling: "Using any number of designs, data are collected that are sufficient to permit the simultaneous estimation of the parameters of the IRT models to be used for the items comprising all of the forms. Then, scale scores for each examinee are computed using their item responses and the jointly calibrated item parameters in the IRT models; the resulting scores are said to be 'on the same scale'" (Thissen, Nelson, Rosa, & McLeod, 2001, p. 159).

Delimitations of the Study

In an attempt to accommodate the high-stakes versus low-stakes aspect of this study, similar subject areas were chosen to help maintain subject matter consistency across grades. Therefore, for the purposes of this study, the PATs and DIPs are limited to the language arts and mathematics subject areas.

Organization of the Dissertation

Chapter two follows with a review of the literature that addresses the use of SR and CR items, procedures for scoring tests comprised of SR items and CR items, and a review of the two studies in which the procedures have been compared. The chapter concludes with a discussion of the low-stakes/high-stakes literature, including descriptions of the low-stakes and high-stakes tests to be used in the current study.

Chapter three discusses the procedures followed to compute and compare the scores yielded by the unweighted raw score, weighted raw score and IRT pattern score scoring procedures. Chapters four and chapter five follow with the presentation of the analyses and results for the low-stakes tests and high-stakes examinations, respectively.

The dissertation concludes with chapter six which begins with a brief summary of the purpose of the study and the procedures followed to address this purpose. This summary is followed by a discussion of results. The limitations of the study are presented next, followed by the conclusion. The chapter concludes with implications for practice and recommendations for future research.

CHAPTER 2

Literature Review

The literature reviewed is related to the use of both selected response and constructed response items included in the assessment instruments used in large-scale assessment programs and the procedures used to combine the scores obtained from each type of item. First the literature related to the inclusion of constructed response items in assessment instruments that previously included only selected response items is reviewed. This is then followed by a discussion of the psychometric models that underlie the different procedures used to combine the scores from the two types of items. These procedures are then presented in the third section together with the research that has been conducted in which the results of these procedures were compared. The differences between low-stakes and high-stakes assessments are discussed in the fourth section, followed by a description of the two large-scale examination programs considered in the present study: low stakes PATs and high-stakes Grade 12 DIP examinations.

Inclusion of Constructed-Response Items in Large-Scale Assessments

Large-scale assessments are becoming increasingly complex due to the inclusion of both SR and CR items (Ercikan, Schwarz, Julian, Burket, Weber, & Link, 1998; Goldberg & Roswell, 2001; Messick, 1994; Schaeffer, Henderson-Montero, Julian, & Bene, 2002; Sykes & Hou, 2003). It is thought that by including both SR and CR items, the benefits of both, such as objective scoring, economy, enhanced reliability of tests or subtests composed of SR items, and enhanced validity due to the inclusion of CR items, will be realized (Schaefer et al., 2002; Sykes & Hou, 2003; Wainer & Thissen, 1993). If the goal of testing is to make reliable inferences about student achievement, then the score that reflects the quantity of the trait the test is designed to measure must be as truthful and accurate as possible (Suen, 1990).

Combining SR and CR item formats together to form a single total test score has several advantages over reporting the scores separately. Sykes and Yen (2000) suggested four reasons for scaling the two item formats together. First, when the two item types are positively correlated (and optimal item weights are used), the combined SR and CR scores produce a total score that is more reliable than scores reported separately by item type. Second, if the SR and CR items are reported separately, there may not be enough items of either type to ensure good trait definition and stable scaling results. Third, by scaling all the items together it is possible to establish a single standard of performance and set the corresponding cut-point in the score distribution (Lewis, Mitzel, & Green, 1996). Finally, by creating a single scale score using pattern scoring, statistically optimal weights are provided and the need to develop an alternate rationale for establishing a weight is avoided. However, sub-scores included as part of the total score are implicitly weighted by their standard deviations. Consequently, scores with greater standard deviations have more impact on the total score through their greater contribution to the total score variance (Sykes &Yen, 2000). "It is not possible to avoid the weighting issue – even if scores are explicitly unweighted, the scores are implicitly weighted by their standard deviations" (Sykes & Yen, 2000, p. 222).

Dimensionality

A concern that arises when SR items and CR items are combined without taking into account that each item type may be measuring different constructs is that the test form may not be unidimensional. Several researchers have investigated the dimensionality of combined response formats in a variety of subject areas. Bennett, Rock, and Wang (1991) examined the equivalence of SR items and CR items included in the College Board's Advanced Placement Computer Science examination. The SR portion of the test consisted of 50 items. The CR portion of the test was comprised of five items, each requiring the students to write a computer program and analyze the efficiency of certain operations involved in the solution. Two samples of 1,000 students were randomly drawn from a population of 7,372 high school students who wrote the 1998 examination. A confirmatory two-factor model composed of (i) five 10-item parcels of randomly assigned SR items, and (ii) the five CR items was tested. The factors were allowed to be correlated, and the variables marking a given factor were constrained to load only on that factor. Results suggested that a one-factor model provided the best fit.

Using full-information item factor analysis (Bock, Gibbons, & Muraki, 1988), Thissen, Wainer, and Wang (1994) examined the dimensionality of the SR items and CR items included in the computer science and chemistry tests of the College Board's Advanced Placement Program. They used the same sample used by Bennett et al. (1991) for the computer science test and a sample of 2,686 students for the advanced chemistry test. The advanced chemistry test was composed of 75 SR questions and nine CR items. The SR items were randomly divided into 15 item parcels of five items. The students had choice among which of the CR items they would respond to. Consequently, the CR items were divided into groups based on the student choices. Clear evidence was found that the CR items predominately measured the same construct as the SR items. However, Thissen et al. also noted a small amount of local dependence among the CR items that resulted in a small amount of multidimensionality. However, the factor loadings of the CR items on the second dimension were small, indicating that the CR items did not measure something different very well.

Bridgeman and Rock (1993) examined the dimensionality of a computer-delivered version of the Graduate Record Examination (GRE) General Test that included both SR and CR items. A sample of 349 students took the GRE General Test in October 1989 and the CR item/task computer-version four months later in February 1990. The relationship among the SR items and CR items were explored using exploratory and confirmatory factor analysis. In order to better approximate the linear factor model assumption of multivariate normality, item parcels of at least four SR items were formed and analyzed. The Tucker-Lewis index, Chi-Square/degrees of freedom, and mean off-diagonal standardized residuals were used to evaluate goodness-of-fit. The factors representing the SR items and CR items were correlated 0.93, suggesting that the two item types were measuring the same construct.

By simultaneously scaling scores from SR items and CR items, Ercikan et al. (1998) examined whether the two item types measured the same construct. Approximately 800 students in each of Grades 3, 5, and 8 were administered SR items and CR items in reading, language, mathematics and science. The SR items were calibrated using the three-parameter logistic (3PL) model (Lord, 1980) and the CR items were calibrated using a two-parameter partial credit (2PPC) model, a special case of Bock's (1972) nominal model which is equivalent to Muraki's (1992) generalized partial credit model. Two sets of analyses were conducted to examine any loss of information due to simultaneous calibration. The results of the first analysis were used to examine whether the simultaneous calibration of the two item types lead to a loss of information, and the results of the second analysis were used to ascertain whether simultaneous calibration lead to scores that were different than those obtained with separate calibrations. The results indicated that the SR items and CR item assessed constructs that were sufficiently similar to allow the creation of a common scale and provide a single set of scores for responses to both item types (Ercikan et al., 1998).

In summary, research demonstrates that SR items and CR items currently used on large-scale assessments are measuring the same constructs and therefore can be simultaneously calibrated. However, as noted by Sykes and Yen (2000), sub-scores included as part of the total score are implicitly weighted by their standard deviations; scores with greater standard deviations have more impact on the total score through their greater contribution to the total score variance. Therefore, it is not possible to avoid the weighting issue. This is further complicated in that it is often required that the SR items and CR items be weighted according to some psychometric or political agenda.

Psychometric Models

Total test scores may be computed using classical test score theory procedures and item response theory procedures. Each of the procedures is discussed below, beginning with the classical test score procedures. First, the model underlying the scoring procedures is presented. Then the scoring procedures based on the model are presented and discussed.

Classical Test Score Theory

In 1904, Charles Spearman laid the foundation for classical test score theory (CTST). The essence of Spearman's theory was that any test score is comprised of two hypothetical components, a true score and an error score, given as:

$$X_{jf} = \tau_j + \varepsilon_{jf}, \qquad (1)$$

where X_{jf} is the observed score for student *j* on form *f* of test X,

τ_j is the true score for student *j*,

ε_{jf} is the error score for student j that arises because X_{jf} does not necessarily equal τ_j.

Spearman defined the true score, which, as shown in equation 1, is constant for student *j*, as

$$\xi_f(X_{jf}) = \tau_j, \qquad (2)$$

where the $f \to \infty$ forms are parallel or fully interchangeable.

He further assumed that:

$$X_{jf} \sim NID(\tau_j, \sigma^2_{\varepsilon_j}) \text{, from which it follows:} \quad (3)$$

$$\varepsilon_{jf} \sim NID(0, \sigma^2_{\varepsilon_j}) , \quad (4)$$

where $\sigma^2_{\varepsilon_j}$ is the variance error of measurement for student *j*, and

$$\xi_f(X_{jf}) - \tau_j = 0 . \quad (5)$$

The problem with this intra-student model is that there is no unique solution for the two unknowns - τ_j and ε_{jf} - in equation 1. Further, it is both not possible to construct and administer an infinite number of forms.

Spearman (1904) proposed using an inter-student model to obtain estimates of $\sigma^2_{\varepsilon_j}$ to address this situation. In addition the assumptions made above he added the following assumptions:

1. $\xi_s(\varepsilon_{jf}) = 0$
2. $\rho_{\varepsilon_j \varepsilon_{j'}} = 0$
3. $\rho_{\varepsilon_j \tau_j} = 0$ (6)

The first assumption implies the mean of the error scores for a population of students is zero. The second assumption implies that the error scores across students are independent. Lastly the third assumption states that error and true scores are independent. Given these assumptions, it follows that the observed score variance for a population of students can be composed into two parts, the variance among true scores and the variance among error scores:

$$\sigma^2{}_X = \sigma^2{}_{\tau+\varepsilon} = \sigma^2{}_\tau + \sigma^2{}_\varepsilon, \tag{7}$$

where σ_X^2 is the variance of the observed scores of the students in the population,

σ_τ^2 is the variance of the true scores of the students in the population, and

σ_ε^2 is the variance of the error scores of the students in the population.

It can be shown that:

$$\sigma_\varepsilon^2 = \frac{\sum_{j=1}^{N_{pop}} \sigma_j^2}{N_{pop}}. \tag{8}$$

Now if σ_ε^2 is zero, then it must be that $\sigma_j^2 = 0$ and $X_j = \tau_j$ for all students in the population. If σ_ε^2 is small, then if follows σ_j^2 were small and X_j were close to τ_j. Consequently, Spearman paid attention in how to estimate σ_ε^2.

Equation 7 can be rewritten as:

$$\begin{aligned} \sigma_\varepsilon^2 &= \sigma_X^2 \left(1 - \frac{\sigma_\tau^2}{\sigma_X^2}\right) \\ &= \sigma_X^2 (1 - \rho_{XX}), \end{aligned} \tag{9}$$

where $\rho_{XX} = \frac{\sigma_\tau^2}{\sigma_X^2}$ is the reliability of test X as defined by Spearman (1904). Procedures for estimating σ_X^2 were known. Spearman provided a procedure for estimating ρ_{XX} which approximated the parallel forms. He showed that the correlation between two parallel forms equaled the reliability. Two forms are parallel if they satisfy the following three conditions:

a. the items in each form are relevant to and representative of the construct being measured;

b. $\mu_{X_{f1}} = \mu_{X_{f2}}$; and

c. $\sigma^2_{X_{f1}} = \sigma^2_{X_{f2}}$.

Reliability and standard error. The reliability coefficient is the ratio of true-score variance to observed-score variance. A test is reliable if its observed scores are highly correlated with its true scores. For a perfectly reliable test,

$$\rho^2_{X\tau} = 1 = \frac{\sigma^2_{\tau}}{\sigma^2_{X}}, \tag{10}$$

and all of the observed variance reflects true-score variance rather than error variance (Allen & Yen, 1979). When reliability increases, error-score variance decreases. When error variance is small, a student's observed score is close to his or her true score.

Standard error and confidence intervals. In order to construct a confidence interval, an estimate of σ_E is required. If it is assumed that the σ_E is the same for all students in the sample (homoscedasticity), then

$$\hat{\sigma}_E = s_E = s_X\sqrt{1 - r_{XX'}}, \tag{11}$$

where s_E is the estimated standard error of measurement,

s_X is the standard deviation of the observed score,

$r_{XX'}$ is the estimated reliability of x (Allen & Yen, 1979).

A confidence interval can be constructed if the following assumptions are made: a) the true-score theory holds as discussed in the previous section, b) errors of measurement are normally distributed, and c) the assumption of homoscedasticity. When these assumptions are met, a confidence interval for a student's true score can be constructed

$$X \pm z_c s_E , \qquad (12)$$

where X is the observed score for the student,

z_c is the critical value for the chosen confidence interval (Allen & Yen, 1979).

Calculation of the observed score X_j. To calculate a student's observed score in CTST on a SR test, the items answered correctly are assigned a value of 1 and those answered incorrectly are assigned a value of 0. The observed test score for the student is the number of correct responses. To calculate the student's observed score on a CR test each of the items is assigned a maximum possible value. The student's score on each item runs from 0 to the maximum value for the item. In contrast to selection items, two or more trained raters assign the scores for CR items. The final score for the item is the mean if the two scores if the two scores are close enough; otherwise the student's response is marked by a third scorer. The score awarded in this situation is that which is given by the third scorer. The test score for the student is the sum of the scores awarded to the CR items.

Both unweighted and weighted procedures are used to add the item scores when both SR and CR items are included on a test. In the *unweighted raw score procedure*, a student's observed score is equal to the number of points earned from the SR items plus the number of points earned from the CR items. In the weighted raw score procedure, the SR and/or CR items scores are purposefully weighted so they count a prescribed amount toward the total test score (Schaeffer et al., 2002). Sometimes the weighting is statistical, such as ensuring that the variances of both the SR and CR scores are equal before adding the two scores. Other times, the "standardized" weights are adjusted so that the two scores reflect a desired set of weights. To ensure that the desired weights are achieved, it may be necessary to first use statistical weighting to achieve equal variance for both scores.

Although CTST has a strong foundation in psychometric practice, it has a few shortcomings. Perhaps the greatest concern is that student characteristics and test characteristics cannot be separated. A student's ability as defined above can only be defined in terms of the particular test that was written. If the test was "hard," then a student of middle ability may appear to have low ability; in contrast, if the test was "easy" the same student may appear to have higher ability. The difficulty of the test item is defined by the number of students in the group of interest, who answer the item correctly. If the students are of high ability, the items will be seen as easy; if the students are of lower ability, the items will be seen as difficult. Test and item characteristics change as the students change, and the students change as the test context changes. Therefore, it is very difficult to compare students who take different tests and compare items written by different groups of students (Hambleton, Swaminathan, & Rogers, 1991).

Another issue concerns the standard error of measurement. The standard error of measurement is assumed to be the same for all of the students writing the test. However scores on any test are unequally precise measures for students of different abilities. Keeping in mind that, as reliability increases, error decreases, it follows then that X approaches τ_j. These differences then are only of concern when reliability decreases. Therefore, the assumption of equal standard error of measurement is questionable when reliability is low (Hambleton et al., 1991).

CTST is a weak true score theory since the true scores and error scores are unobservable theoretical constructs and, therefore, equations (1) through (10) cannot be proved or disproved. There are no assumptions about the frequency distribution of the scores and there are no formal statistical tests that can be used to examine how well the model fit the data (Gierl, 2001). Therefore, CTST is by its nature a unidimensional theory. If a resulting score is not consistent with the theory, the discrepancies are likely attributed to sampling fluctuations or that the subtests are really not parallel (Lord, 1980). Multidimensionality is not considered. Item Response Theory (IRT) is known as a strong theory because strong assumptions must be met (Embretson & Reise, 2000). IRT is discussed is the next section.

Item Response Theory

Lord (1952), considered the limitations of CTST and proposed an alternative theory that he called Item Response Theory (IRT). The desirable features of IRT included the possibility of obtaining item characteristics that are not group-dependent and scores that describe student ability that are not item dependent. Further, it is important to note that in IRT, the standard error of estimation ($SE(\hat{\theta})$), which serves the same purpose as standard error of measurement in CTST, varies with ability level (Hambleton & Swaminathan, 1985; Hambleton et al., 1991; Lord, 1980).

Item response theory rests on two basic postulates: (a) the performance of a student on a test can be predicted by a set of factors called latent traits, or abilities;

and (b) the relationship between the student's item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic curve (ICC) if the test is unidimensional and an item characteristic function (ICFF) if the test is multidimensional. For simplicity, IRT is discussed here for the unidimensional case.

IRT models involve two key assumptions: (a) the ICCs have a specific form, and (b) local independence is obtained. Each is discussed below.

The form of an ICC specifies the relationship between the student's latent trait or ability measured by the test and the probability of a correct response to the item (Hambleton et al., 1991; Lord 1980). For dichotomous items, where a student's response is considered correct or incorrect, the ICC regresses the probability of item success on trait level or ability. For polytomous items, such as open ended questions, the ICC regresses the probability of responses in each category on trait level or ability (Embretson & Reise, 2000). Figure 2 illustrates ICCs for four dichotomous items from an IRT model where the relative ordering of item difficulty is constant across score levels.

Several notes may be made about the ICCs. First, each ICC is S shaped, which plots the probability of a correct response as a monotonic and increasing function of ability. In the middle of each curve, small changes in ability imply large changes in item solving probabilities. At the extremes of the curves, large changes in ability result in small changes in probabilities. Second, although all four ICCs shown in Figure 1 have the same general shape, they differ in where they are located. The location of each ICC reflects the item's difficulty. The location represents the extent to which each item differs in probability across the ability levels (Embretson & Reise, 2000). For example, the ability level associated with a probability of 0.5 is much lower for Item 1 and Item 2 than for Item 3 and Item 4. Thus, Item 1 and Item 2 are easier.

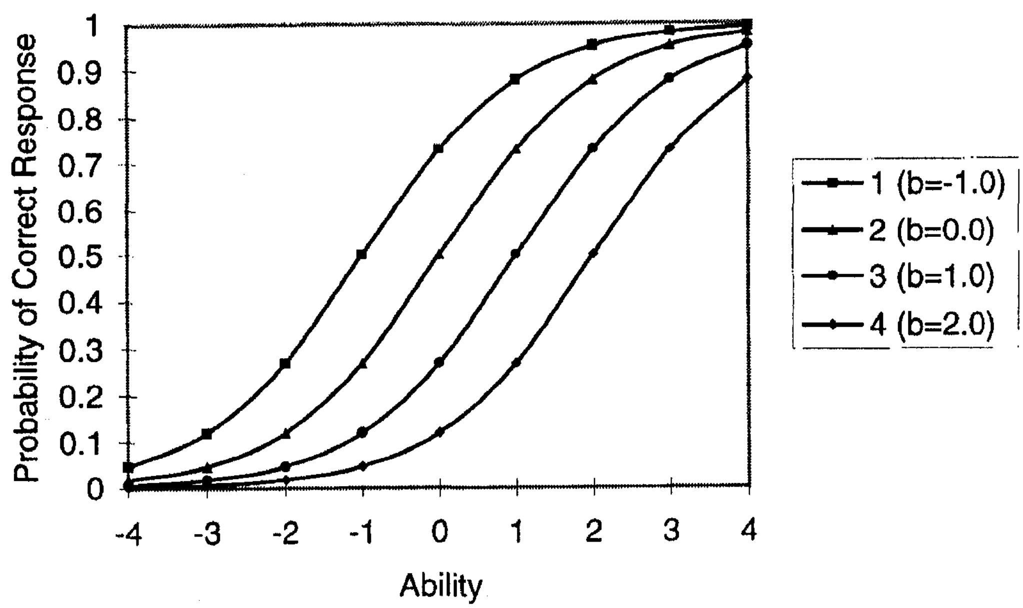

Figure 1. Item Characteristic Curves for the One-Parameter Model (Hambleton et al., 1991, p.14).

As implied above, an assumption of a unidimensional model is that the test of interest is unidimensional. Unidimensionality assumes that only one "dominant" factor or ability is measured by the items that make up the test. The second assumption of local independence is related to unidimensionality. Local independence means that when the abilities influencing test performance are held constant, the student's responses to any pair of items are statistically independent (Hambleton et al., 1991; Lord 1980). When the assumption of unidimensionality is met, the assumption of local independence is obtained (Hambleton et al., 1991; Lord, 1980).

A popular distinction between the IRT models is the number of item parameters used to describe an item. The three most popular unidimensional models are the one-, two- and three-parameter logistic models, each named due to the number of item parameters involved (Hambleton et al., 1991). These models are appropriate for dichotomous item response data. A fourth model, the two-parameter partial credit model, is a unidimensional model appropriate for polytomous response data that is unidimensional. The one-, two- and three-parameter logistic models and the two-parameter partial credit model are discussed below.

One-parameter logistic model. The one-parameter logistic (1PL) model (see Figure 1) specifies that the probability of a correct response to an item is a function of a student's ability and one item parameter: item difficulty (b_i). The item difficulty is a location parameter that indicates the position of the ICC in relation to the ability scale. It corresponds to the point on the ability scale where the probability of a correct answer is 0.50. The b_i is a location parameter, reflecting the position of the ICC in relation to the ability scale. As the b_i parameter increases, greater ability is required for student to have a 50% chance of getting the item right; hence the harder the item. If the abilities are transformed with a mean of 0 and standard deviation of 1, the values of b_i vary from about -2.0 to +2.0. Items with b_i near -2.0 are very easy, and item with b_i near 2.0 are very difficult for that group of students (Hambleton et al., 1991).

The one-parameter ICCs for four different items are displayed in Figure 1. These ICCs were determined from the following equation:

$$P_i\left(\theta_j\right) = \frac{e^{\left(\theta_j - b_i\right)}}{1 + e^{\left(\theta_j - b_i\right)}}, \qquad i = 1, 2, \ldots, n, \qquad (13)$$

where $P_i\left(\theta_j\right)$ is the probability that a randomly chosen student j with ability θ_j answers item i correctly, or $P_{ij}(\theta)$ is the probability that a randomly chosen student j with ability θ answers item i correctly

b_i is the item i difficulty parameter,

n is the number of items in the test, and

e is a transcendental number whose value is 2.718 (correct to three decimals).(Hambleton et al., 1991; Lord, 1980; Sykes & Hou, 2003).

In the 1PL model, it is assumed that item difficulty is the only item characteristic that affects student performance. This is demonstrated in Figure 1 where Item 1 is the easiest ($b_1 = -1.0$), followed in turn by item 2 ($b_2 = 0.0$), item 3 ($b_3 = 1.00$), and Item 4 ($b_4 = 2.00$), which is the most difficult item. The four ICCs are parallel because the model assumes that all the items discriminate equally. It also assumes that the students of very low ability have no chance of answering the item correctly, thus no allowance is made for guessing. This is also demonstrated in Figure 2 where the lower asymptotes of the four ICCs are zero (Hambleton et al., 1991).

Two-parameter logistic model. The two-parameter logistic model (2PL: Lord 1980) specifies the probability of a correct response to an item as a function of a student's ability and two item parameters: difficulty (b_i) and discrimination (a_i). This model was introduced to account for the lack of equality of item discrimination assumed for the one-parameter model. The item difficulty, b_i, is the same as in the 1PL model. It marks the point on the ability scale where the probability of correctly answering an item is 0.50. The item discrimination parameter, a_i, is proportional to the slope of the ICC at point b_i on the ability scale. Items with steeper slopes are more useful for distinguishing higher ability students than items with less steep slopes. The scale of a_i is theoretically from -∞ to +∞. Negatively discriminating items are removed as they suggest the probability of correctly answering the item decreases as ability increases. It is also unusual to have a_i values larger than 2. Therefore, the usual range of a_i is from zero to two (Hambleton et al., 1991). The probability that student *j* with ability θ answers item *i* correctly for the two-parameter model is given by:

$$P_i(\theta_j) = \frac{e^{[Da_i(\theta - b_i)]}}{1 + e^{[Da_i(\theta - b_i)]}}, \qquad (14)$$

where a_i is the discrimination parameter (slope parameter for item *i*), and D is the scaling factor 1.7 introduced to make the logistic function as close as possible to the normal ogive function. (Hambleton et al., 1991; Lord, 1980; Sykes & Hou, 2003).

Figure 2 shows ICCs for the 2PL model for four different items. The difficulties of these four items are the same as the corresponding items in Figure 2.

However, the slopes of the ICCs are no longer parallel as they were in the 1PL model (cf., Figure 2) due to the different item discrimination parameters. Notice that the ICC for Item 3 crosses the ICCs for Item 1, Item 2 and Item 4. Item 1 is the most discriminating ($a_1 = 1.5$). The least discriminating item is Item 3 ($a_3 = 0.5$). Inspection of the ICCs for these two items reveals that the ICC for Item 1 rises much more sharply than the ICC for Item 3. As with the 1PL model, the item difficulty in the 2PL model still corresponds to the point on the ability scale where the probability of a correct answer is 0.50. The lower asymptotes are still zero as the 2PL model, like the 1PL model, does not take guessing into account.

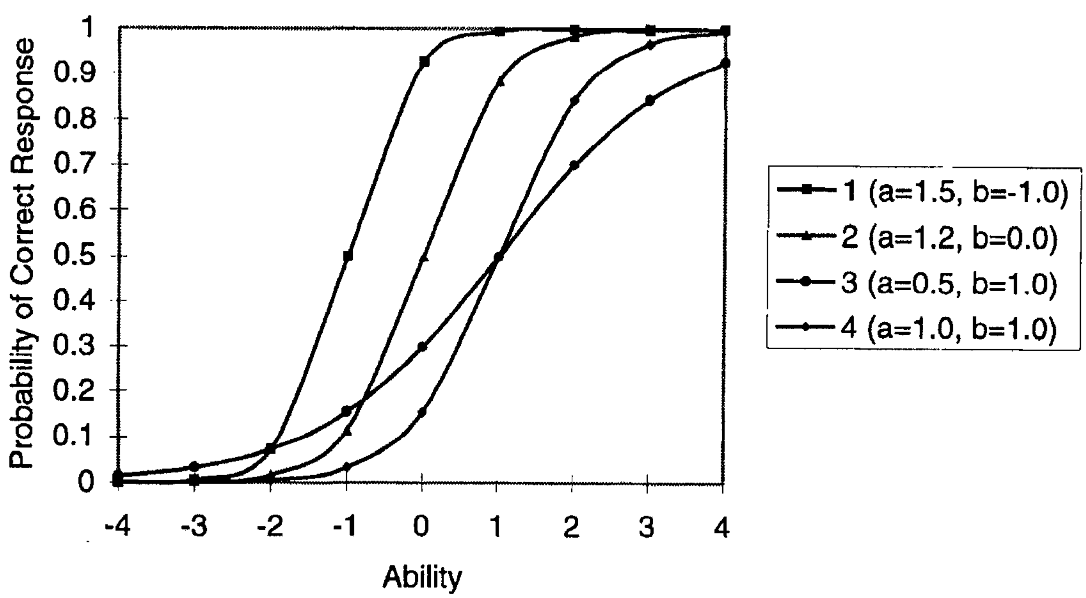

Figure 2. Item Characteristic Curves for the Two-Parameter Model (Hambleton et al., 1991, p.16).

Three-parameter logistic model. The three-parameter logistic model (3PL: Lord 1980) specifies the probability of a correct response to an item as a function of a student's ability and three item parameters: difficulty (b_i), discrimination (a_i), and pseudo-guessing (c_i). The item difficulty, b_i, is defined differently for the 3PL model due to the presence of the pseudo-guessing parameter. In this case, b_i is located at the point on the ability scale for which $p_i = \frac{1+c_i}{2}$. The item parameter, a_i, is still proportional to the slope of the ICC at b_i on the ability scale. The pseudo-guessing parameter, c_i, represents the probability of a student with low ability answering the item correctly. This parameter provides a possible nonzero lower asymptote for the ICC. Typically, the guessing parameter assumes values that are smaller than the value that would result if the student guessed randomly on the item (Hambleton et al., 1991). The probability that student j with ability θ answers item i correctly for the three-parameter model is given by the equation:

$$P_i(\theta_j) = c_i + (1 - c_i)\frac{e^{[Da_i(\theta - b_i)]}}{1 + e^{[Da_i(\theta - b_i)]}}, \tag{15}$$

where c_i is the pseudo-guessing parameter for item i. (Hambleton et al., 1991; Lord, 1980; Sykes & Hou, 2003).

Figure 3 shows four typical ICCs for the 3PL model. The four curves differ in their location on the ability scale (b_i). A comparison between Item 4 and Item 3 to Item 2 and Item 1 (and especially Item 4 and Item 1) reflects the effect of item difficulty b_i, on the location of the ICCs. The more difficult items (Items 3 and 4) are found at the higher end of the ability scale, whereas the easier items are found at the lower end of the ability scale. The steepness of each ICCs is influenced by the item discrimination parameters (a_i), especially the more discriminating Item 1 and much less discriminating Item 3. Finally, unlike the 1PL and 2PL models, the values of the lower asymptotes in the 3PL model are affected by the pseudo-guessing parameter (c_i). This is best exemplified by the difference between the lower asymptote of Item 4, and that of Item 1 which is considerably lower and therefore less susceptible to guessing than Item 4. The 3PL model was used in the present study with the dichotomously scored selected response items. This decision was based on the use of the 3 PL model in the previous research upon with the present study is based on and the intent to compare the results of the present study with the results obtained in the previous studies.

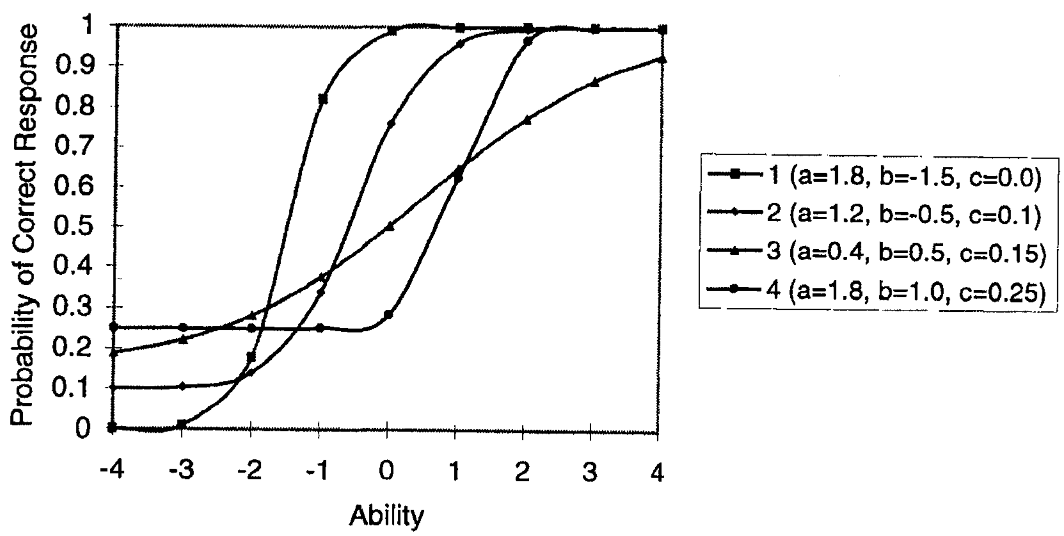

Figure 3. Item Characteristic Curves for the Three-Parameter Model (Hambleton et al., 1991, p.18).

Two-parameter partial credit model. In some cases, researchers use testing formats that cannot be scored as right versus wrong as with the 1PL, 2PL and 3PL models. In these multiple-category types of item responses, polytomous IRT models are required to represent the nonlinear relationship between student ability and the probability of responding in a particular category (Embretson & Reise, 2000). Several polytomous models are available including the Graded Response Model (Samejima, 1969), the Partial Credit Model (Masters, 1982), and the Generalized Partial Credit Model (Muraki, 1992). Since Muraki's (1992) generalized partial credit model (GPCM) was used in the previous studies, it was selected for the present study. Consequently, the GPCM is presented below.

The difference between the two-parameter logistic model and the two-parameter partial credit model is not the number of parameters, but rather, the difference in terms of the presence of an ICC for each scoring category in the scoring process. Samejima (1972) referred to the set of curves and the equation that produces them as the operating characteristic function (OCF) of the item. The OCF relates "how the probability of a specific categorical response is formulated according to the laws of probability as well as psychological assumptions about item response behaviour" (Muraki, 1992, p.160).

The GPCM was developed based on the assumption that the probability of selecting the *k*th category over the *k* minus first (*k*-1) category is governed by the dichotomous response model (Muraki, 1991). In the GPCM for a constructed response item denote, $P_{ik}(\theta_j)$ is the specific probability of selecting *k*th category from

m_i possible categories of item i. "For each of the adjacent categories, the probability of the specific categorical response k over k-1 is given by the conditional probability" (Muraki, 1992, p. 160):

$$C_{ik} = P_{ik|k-1,k}\left(\theta_j\right) = \frac{P_{ik}(\theta_j)}{P_{i,k-1}(\theta_j) + P_{ik}(\theta_j)},$$

where $k = 2,3,\ldots, m_i$. Then,

$$P_{ik}\left(\theta_j\right) = \frac{c_{ik}}{1-c_{ik}} P_{i,k-1}\left(\theta_j\right). \qquad (16)$$

Note that $\frac{c_{jk}}{1-c_{jk}}$ is the ratio of the two conditional probabilities that may be expressed as $P_{ik}(\theta_j)\, \mathrm{e}^{\left(a_j\left(\theta - b_{jv}\right)\right)}$. Equation 17 may be referred to as the operating characteristic function for the GPCM. $P_{ik}(\theta)$ is given by (Muraki, 1992, p. 161):

$$P_{ik}\left(\theta_j\right) = \frac{\mathrm{e}^{\left(\sum_{v=1}^{k} a_i\left(\theta - b_{iv}\right)\right)}}{\sum_{c=1}^{m_i} \mathrm{e}^{\left(\sum_{v=1}^{c} a_i\left(\theta - b_{iv}\right)\right)}}, k = 1,\ldots, m_i, \qquad (17)$$

where b_{ik} are the item step parameters (Masters, 1982) that are located at the points on the θ scale where the plots of $P_{i,k-1}\left(\theta_j\right)$ and $P_{i,k}\left(\theta_j\right)$ intersect. These two curves, which can be referred to as the item category response functions (ICRFs), intersect only once, and the intersection can occur anywhere along the θ scale (Muraki, 1992), and

a_i is the item discrimination.

In the GPCM, b_{i1} is arbitrarily set to 0. This is not a location factor. It could be any value as the term including this parameter is removed from the numerator and denominator of the model:

$$P_{ik}\left(\theta_j\right)=\frac{\mathrm{e}[Z_{i1}(\theta)]\mathrm{e}\left[\sum_{v=2}^{k}Z_{iv}(\theta)\right]}{\mathrm{e}[Z_{i1}(\theta)]+\sum_{c=2}^{m_i}\mathrm{e}\left[Z_{i1}(\theta)+\sum_{v=2}^{c}Z_{iv}(\theta)\right]}=\frac{\mathrm{e}\left[\sum_{v=2}^{k}Z_{iv}(\theta)\right]}{1+\sum_{c=2}^{m_i}\mathrm{e}\left[\sum_{v=2}^{c}Z_{iv}(\theta)\right]}, \qquad (18)$$

where $Z_{ik(\theta)=a_i}(\theta-b_{ik})$.

The item discrimination parameter reflects the degree to which categorical responses vary among items as θ changes (Muraki, 1992).

Figure 4 shows the ICRFs for the GPCM for three items with four categorical responses. Figure 5a shows the ICRFs for an item with a_i = 1.0, b_{i2} = -2.0, b_{i3} = 0.0 and b_{i4}= 2.0. If b_{i2} is changed to -0.5, then the probability of responding to the second category decreases, as shown in Figure 4b. In other words, the range of the θ values of persons who are more likely to respond to the second category than to the other categories decreases. If the slope parameter is changed from 1.0 to 0.7 as shown in Figure 4c, the curves become flatter, but the intersection points on all

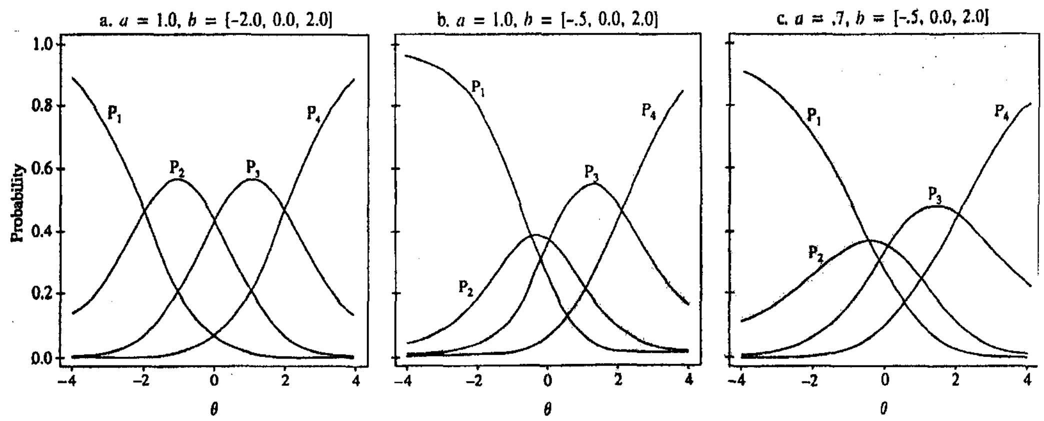

Figure 4. Item Category Response Functions for a Four Category Item (Muraki, 1992, p. 163).

ICRFs are left unchanged. The discriminating power of the ICRFs decreases for all categorical responses (Muraki, 1992).

True score CTST vs. IRT. In IRT, every student at ability θ has the same number-right true score (Lord, 1980). Since each $p_i(\theta j)$ increases as θ increases, number-right true score is an increasing function of ability. This is the same as τ discussed in the above section on Classical Test Score Theory. True score τ and ability θ are the same but expressed on different scales of measurement. The major difference is that the classical measurement scale for τ depends on the items in the test whereas the IRT measurement scale for θ is independent of the items in the test. This makes θ more useful than τ when comparing different tests for students of the same ability (Lord, 1980).

Parameter estimation. Item response theory is based on the assumption that the probabilities of a response from a student on an item can be estimated from knowledge of the student's ability and the item parameters. Therefore knowledge of the values of ability and item parameters are required to obtain the item response function that can then be used to estimate the probability of a response for student on a particular item (Hambleton et al., 1991; Swaminathan, 1983). This can be done by using the item responses of a random sample of students from the population of interest who wrote a test. Once these item responses are obtained, the ability parameters and item parameters can be estimated (Hambleton et al., 1991; Swaminathan, 1983).

Maximum likelihood, marginal maximum likelihood, and Bayesian estimation are the most widely used estimation procedures. Since maximum likelihood estimation procedures were used in the previous studies, it was selected for the present study and is described below. Since the dichotomous item response model can be thought of as a special case of the polytomous item response model in which the number of categories is two, the discussion below can be applied to polytomously scored items when each score category is treated as a "binary item" (Hsu, Ackerman & Fan, 1999; Muraki, 1992).

Maximum likelihood estimation is a search process based on finding the value of θ that maximizes the likelihood of a student's item response pattern (Embretson & Reise, 2000). Maximum likelihood estimators are (a) consistent (i.e., the estimations approach the true parameter being estimated as the sample size and number of items increase); (b) sufficient (i.e., functions of sufficient statistics when they exist); (c) efficient (i.e. asymptotically, maximum likelihood estimators have the least amount of variance); and (d) asymptotically normally distributed (Swaminathan, 1983). When estimating ability, the assumption of local independence must hold true and thus, the probability of observing the response pattern is the product of the probabilities of observing each item response.

The conditional likelihood of a response pattern can be computed by:

$$L(u_1, u_2, \ldots, u_n, \left| \theta_j \right.) = \prod_{i=1}^{n} P_i^{u_i} Q_i^{1-u_i}, \qquad (19)$$

where u_i is the observed response to item *i*,

$P_i = P(U_i | \theta_j)$ is the probability of answering the item correctly,

$Q_i = 1 - P(U_i | \theta_j)$ is the probability of answering the item incorrectly.

Since P_i and Q_i are functions of θ_j and the item parameters, the likelihood function is also a function of those parameters (Hambleton et al., 1991). However, since the likelihood function is a product of quantities, each bounded between 0 and 1, the resulting likelihood function would be very small. Considering the properties of logarithms (Hambleton et al., 1991):

$$\ln xy = \ln x + \ln y$$

and

$$\ln x^a = a \ln x,$$

the expression for the log-likelihood function is (Hambleton et al., 1991)

$$\ln L(u | \theta_j) = \sum_{i=1}^{n} \left[u_i \ln P_i + (1 - u_i) \ln(1 - P_i) \right], \qquad (20)$$

where u is the vector of item responses a student, and

ln is the natural logarithm.

Both the likelihood function and the natural log of the likelihood function are monotonically related, thus the value of θ that maximizes $L(u|\theta_j)$ will also maximize $\ln L(u|\theta_j)$ (Gierl, 2001).

The value of θ that makes the log-likelihood function for a student a maximum is defined as the maximum likelihood estimate of θ for that student (Hambleton et al., 1991). For short tests, it may be possible to add the log-likelihood functions for each item response together and then get a rough estimate of the student's ability level. However, researchers are often dealing with large tests where thousands of students respond to 50 items. In these cases, iterative computerized statistical search procedures are required to pinpoint exactly where the maximum of the log-likelihood function is given a particular pattern of item responses. One of the most frequently used procedures to find the maximum of the log-likelihood function is the iterative Newton-Raphson procedure (Hambleton et al., 1991; Embretson & Reise, 2000).

The first step in the Newton-Raphson scoring procedure is to specify a starting value for θ. This θ is a guess at what a student's ability level may be. Using this value for θ, the first and second derivatives of the log-likelihood function are computed. The ratio for those values is then computed (the first derivative divided by the second derivative). A new updated ability level estimate is created by taking the old estimate minus the ratio. Using this updated ability level estimate, the iterative procedure is repeated until the ratio is less than a predetermined small value (e.g. 0.001) (Embretson & Reise, 2000).

In describing the procedures for estimating θ, an assumption was made that the item parameters were known. In typical IRT applications, both the item parameters and the trait levels are unknown and must be estimated from the same data. Marginal maximum likelihood estimation (MMLE) is a popular procedure for estimation with unknown ability levels and item parameters (Hambleton et al., 1991; Embretson & Reise, 2000).

In MMLE, an a priori distribution of ability based on the assumption that the students are selected randomly from the population, is used to estimate the item parameters. The distribution must approximate the distribution of ability therefore a large sample size is necessary. In the resulting marginal maximum likelihood estimates, the item parameters are consistent as the number of students increase. The expectation/maximization (EM) algorithm was developed by Bock and Aiken in 1981. The EM algorithm is an iterative procedure where the expected frequencies for a correct response and ability level are successively improved with each iteration (Embretson & Reise, 2000).

Item information. Item information functions, $I_i\left(\theta_j\right)$, provide the contribution items make to ability estimation at points along the ability continuum. They provide a procedure of describing items in IRT. The item information function for the 3PL model is given by:

$$I_i\left(\theta_j\right)=\frac{2.89a_i^2\left(1-c_i\right)}{\left[c_i+e^{1.7a_i(\theta-b_i)}\right]\left[1+e^{-1.7a_i(\theta-b_i)}\right]^2}, \qquad (21)$$

where $I_i\left(\theta_j\right)$ is the information provided by the item i at θ for student j (Hambleton et al., 1991).

As shown in Figure 5, information is generally higher when the ability parameter (*b)* is closer to θ than when it is far from θ, item discrimination (*a*) is high, and guessing (*c*) approaches zero (Hambleton et al., 1991). For the 3PL model, an item provides the maximum information at $\theta_{\max}$ where (Hambleton et al., 1991):

$$\theta_{\max} = b_i + \frac{1}{Da_i} 1n\left[0.5\left(1+\sqrt{1+8c_i}\right)\right]. \qquad (22)$$

If guessing is minimal, then $\theta_{\max} = b_i$. However, when the guessing parameter is greater than zero, the item will provide its maximum information at an ability level slightly higher than its difficulty (Hambleton et al., 1991).

In the polytomous model, each option or category provides information about the student's ability. The information function for an item's individual response category (item-category information function) is (De Ayala, 1993):

$$I_{ik}\left(\theta_j\right) = \frac{P_{ik}^{'}(\theta_j)^2}{P_{ik}(\theta_j)}, \qquad (23)$$

and for the entire item is:

$$I_{ik}\left(\theta_j\right) = \sum_{k=0}^{m} \frac{P_{ik}^{'}(\theta_j)^2}{P_{ik}(\theta_j)} \qquad (24)$$

where $I_{ik}\left(\theta_j\right)$ is the information provided by category *k* of item *i* at θ for student *j*,

$P_{ik}(\theta_j)$ is the probability of the student responding in category score k or higher on item i, and

$P^{'}_{ik}(\theta_j)$ is the first derivative of $P_{ik}(\theta_j)$.

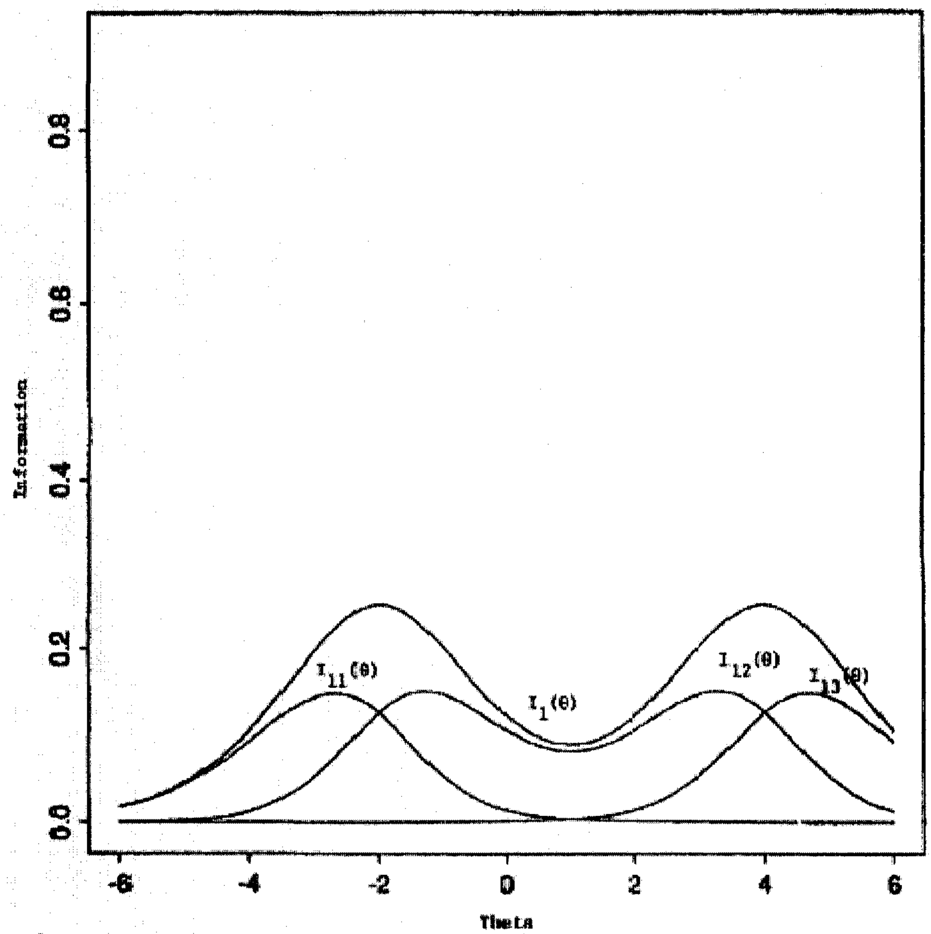

Figure 5. Item Information Function and Item-Category Information Function for Item 1 (Muraki, 1993, p.18).

The item-category information functions and item information functions for Item 1 are shown in Figure 6. Unlike dichotomous items, the item information functions for polytomous items are not necessarily unimodal. In Figure 6, the distance between the two adjacent item-category parameters category $b_{1,2}$ and $b_{1,3}$ is large thus the information becomes lower at the middle range of the θ scale (Muraki, 1993). This loss of information over the middle range becomes less pronounced as the distance between the parameters decreases.

The information function for Item 3 in Figure 6 looks relatively unimodal where the distance between $b_{3,2}$ and $b_{3,3}$ is 2.0. This type of item is preferred if the sample of students is assumed to be normally distributed across ability θ (Muraki, 1993).

The shape of the item information function for Item 5 in Figure 7 resembles that of dichotomous item responses. The item information peaks over a very short range of the lower abilities and the information for students of higher abilities is lost (Muraki, 1993).

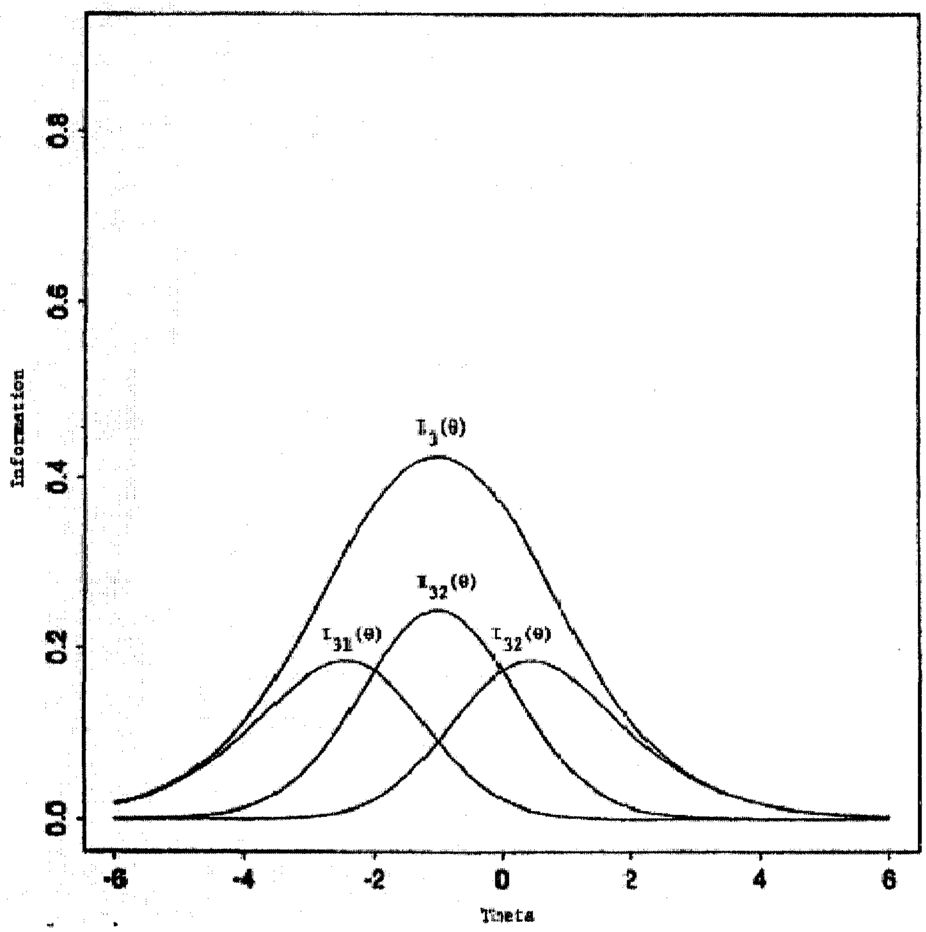

Figure 6. Item Information Function and Item-Category Information Function for Item 3 (Muraki, 1993, p.18).

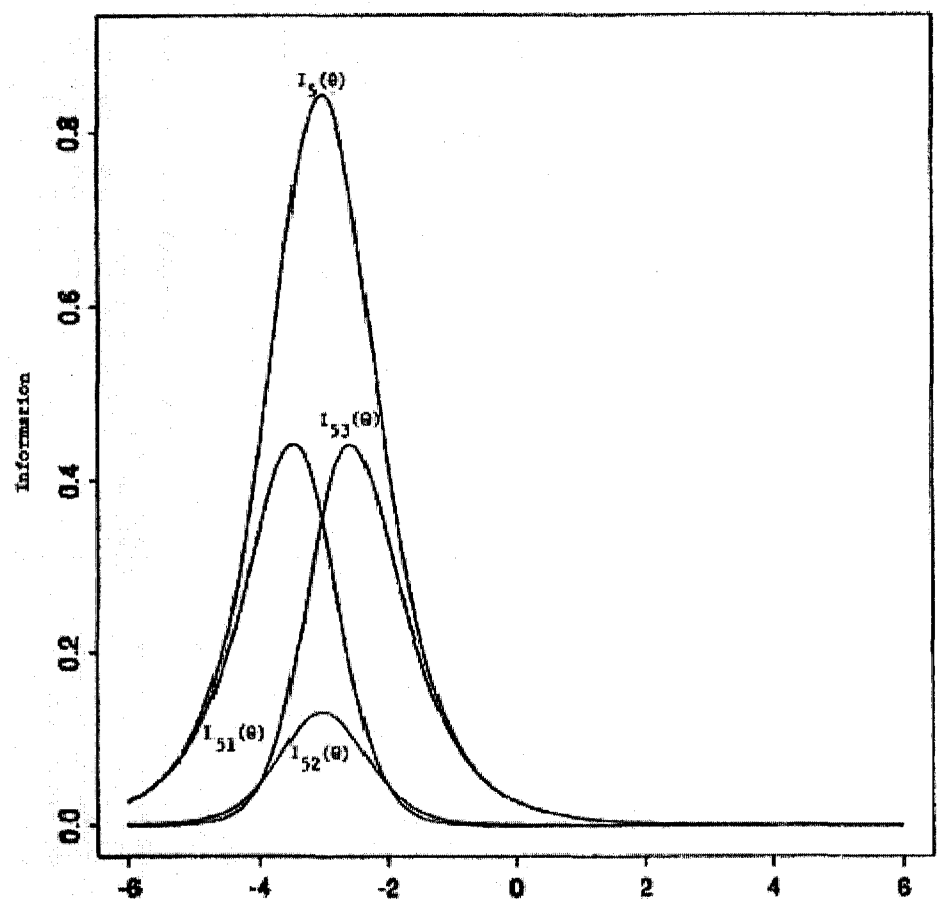

Figure 7. Item Information Function and Item-Category Information Function for Item 5 (Muraki, 1993, p.20).

Test information and standard error of estimate. The test information function for both dichotomous and polytomous items, denoted $I(\theta)$, is the sum of the item information functions at θ (Hambleton et al., 1991):

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) . \qquad (25)$$

Test information is critically important in determining how well a test is performing. This is because $I(\theta)$ has an exact relationship with a student's standard error of measurement (Embretson & Reise, 2000). Specifically, a student's standard error can be written as:

$$SE\left(\hat{\theta}\right)=\frac{1}{\sqrt{I(\theta)}}, \tag{26}$$

where $SE\left(\hat{\theta}\right)$ is the standard error of the ability estimate θ (Hambleton et al., 1991; Embretson & Reise, 2000).

$SE\left(\hat{\theta}\right)$ serves the same role in IRT that standard error of measurement does in CTST. However, the value of $SE\left(\hat{\theta}\right)$ changes with ability level (Hambleton et al., 1991).

The magnitude of the standard error depends, in general, on (a) the number of test items (smaller error with longer tests); (b) the quality of the test items (higher discriminating items with limited guessing result in smaller standard errors); and (c) the match between item difficulty and student ability (tests with items with difficulty parameters close to the ability parameter were associated with smaller standard error) (Hambleton et al., 1991).

Parameter effects on maximum likelihood scoring. There are a number of interesting properties of maximum likelihood scoring. First, when items are equally discriminating all students with the same raw score receive the same θ score and standard error. In Table 1, Test A, where all the items have constant item discriminations of 1.5, the first six students received the same raw score but correctly

Table 1

Maximum Likelihood θ Estimates and Standard Errors When Scored Under Test A and Test B (Embretson & Reise, 2000, p. 167).

Student	*Pattern*	*Test A*		*Test B*	
		θ	*SE*	θ	*SE*
1	1111100000	0.00	0.54	-0.23	0.43
2	0000011111	0.00	0.54	0.23	0.43
3	0000000111	-0.92	0.57	-0.34	0.44
4	1111000000	-0.45	0.55	-0.51	0.46

answered different items. Despite these differences all six students received the same θ and standard error. In this model, maximum likelihood scoring does not take into account the consistency of the student's response pattern as item difficulty is not considered (Embretson & Reise, 2000).

Second, when item discrimination is taken into account, θ estimates are increased according to the discrimination parameters of the item. Therefore, students with the same raw score now may have different θ scores depending on their item response pattern. For example, in Table 1, where the items in Test B have discriminations that go from 1.0 to 1.9 and constant difficulty parameters of 0.0, Student 1 answered the first five items correctly and received a θ of -0.23, whereas Student 2 answered the last five items correctly and received a θ of 0.23. Also, since the items vary in discrimination it is possible for a student to have a high raw score but lower θ than a student with a lower raw score. For example, Student 3 answered three questions correctly and received a θ of -0.34 whereas Student 4 answered four questions correctly and only received a θ of -0.51. This is not to say that the item difficulty does not play a role. Rather, the item difficulty determines the location of the item's log-likelihood function and ultimately, determines where the function is maximized (Embretson & Reise, 2000).

Finally, the θ levels and standard errors are affected as the item discrimination parameters change. In Table 2, we can see the effects of increasing and decreasing item discrimination parameters. The first set of columns shows the θ levels and standard errors where all $\alpha_i = 1.0$. The θ levels in the first column in Table 2 are not equal to those in the first column in Table 2 where the item discrimination is 1.5. This is due to the lower item discrimination parameters which provide less information and therefore the scores are more spread out. Also, the standard errors are much larger. In the second and third columns of Table 2, the item discriminations are increased 0.5 and 1.0 from the original Test A. As the item parameters increase, the θ scores get closer to zero and the standard errors decrease. Figure 8 shows the likelihood function for Student 1 when the discriminations were set to 1.0 and 2.5, respectively. The log-likelihood function with the item discrimination at 2.5 is much steeper than when the discrimination is set to 1.0.

Table 2

Maximum Likelihood θ Estimates and Standard Errors When Scored Under Test A With Item Discriminations Set at 1.0, 2.0 and 2.5 (Embretson & Reise, 2000, p.170).

Student	*Pattern*	$\alpha = 1.0$		$\alpha = 2.0$		$\alpha = 2.5$	
		θ	*SE*	θ	*SE*	θ	*SE*
1	1111100000	0.00	0.73	0.00	0.45	-0.00	0.39
2	0000011111	0.00	0.73	0.00	0.45	-0.00	0.39
3	0000000111	-1.11	0.78	-0.85	0.48	-0.81	0.42
4	1111000000	-0.54	0.74	-0.41	0.45	-0.39	0.40

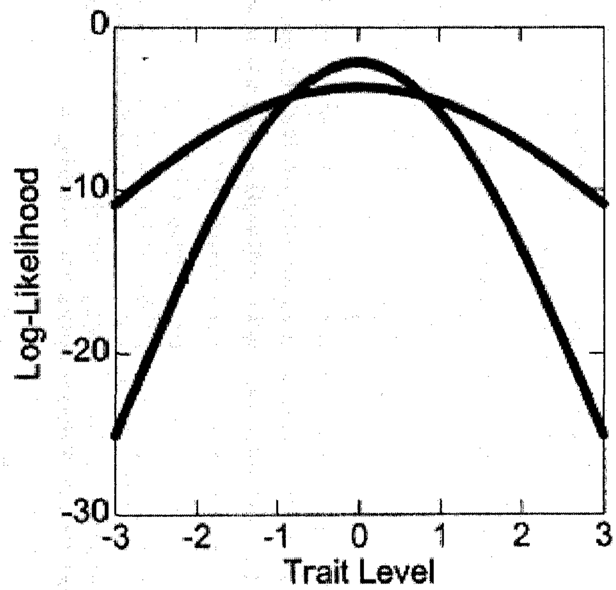

Figure 8. Example of log-likelihood functions when item discriminations vary in size (Embretson & Reise, 2000, p. 171).

The standard error would be smaller with the student being measured with more precision with the item discrimination at 2.5 than at 1.0 (Embretson & Reise, 2000).

Multiple Scoring Procedures for Tests with Combined Response Formats

There have been several studies that have addressed multiple scoring procedures, but with only one item type (Wang, Kolen & Harris, 2000) or different item types using only one (i.e., IRT) scoring procedure (Ercikan et al., 1998; Fitzpatrick, Link, Yen, Burket, Ito, & Sykes, 1996; Sykes & Yen, 2000). Only two studies (Schaeffer et al., 2002; Sykes & Hou, 2003) were found in the literature that addressed multiple scoring procedures with tests that contained both SR items and CR items. These two studies are discussed below.

Schaeffer et al. (2002) compared three different ways of scoring tests that contained both SR items and CR items. Field tests from a low-stakes Grade 9 statewide assessment were used with 1,463 Biology and 1,537 English student results. Both tests contained SR items and CR items. The item parameters for each test were simultaneously calibrated on the same scale. The three-parameter logistic model (3PL: Lord 1980) was used for the SR items. A generalization of Masters' (1982) Partial Credit Model, which is the same as Muraki's (1992) "generalized" partial credit model, was used for the CR items.

The computer program PARDUX (Burket, 1998), which uses marginal maximum likelihood procedures, was used to estimate the parameters. The program WINFLUX (Burket, 1999) was used to place the item parameters onto a single score scale. A multiplier of 50 and an additive constant of 500 were used as scaling parameters.

The resulting student response strings were then scored in each of three ways:

1. *Unweighted raw score procedure*. In the unweighted raw score procedure, a student's observed score was equal to the number of points earned from the SR items plus the number of points earned from the CR items. The focus of this procedure was the total number of points obtained by the student on the test as a whole (Schaeffer et al., 2002).
2. *Weighted raw score procedure*. In the weighted raw score procedure, the CR items were purposely weighted to a predetermined level so that the SR items and CR items contributed the same number of points toward the total score. A student's score was equal to the number of SR

items answered correctly plus *n/m* times the number of points earned from the CR items where n = number of SR items and m = number of CR possible points.

3. *IRT pattern scoring.* With IRT pattern scoring, each student's score was based on a maximum-likelihood estimate derived from the student's item-response vector. This procedure used the optimal item weights determined in terms of item information.

The means and standard deviations of the scaled scores were compared. The scoring procedures were compared for total group and subgroups defined in terms of gender (Female Biology n = 730, English n = 767; and Male Biology n = 720, English n = 751) and ethnicity. For the ethnic subgroups, only African American (Biology n = 367; English n = 533) and White students (Biology n = 881; English n =727) were examined due to insufficient numbers in other ethnic groups. The scaled score distributions, standard errors of measurement, and proficiency-level classifications were compared. Scale scores from the IRT pattern scoring and from raw scoring procedures were found to be tau-equivalent. The tau-equivalence also held for IRT pattern scoring and raw scoring procedures within ethnic subgroups (Schaeffer et al., 2002).

Schaeffer et al. (2002) reported that the three scoring procedures they examined yielded similar results. The lowest correlations were between the SR points and CR points, 0.66 for both tests, suggesting that they might be assessing different constructs. To address these low correlations, disattenuated correlations were computed (Allen & Yen, 1979, p. 98). The resultant disattenuated correlations between SR points and CR points, 0.79 for Biology and 0.76 for English, supported the assumption of unidimensionality. The score distributions and correlational patterns for the total group and the gender and ethnicity subgroups were similar. The scoring procedures were also evaluated by comparing the scores that individual students obtained under each procedure. Differences were computed by subtracting the pattern score from the weighted score. The results revealed that the scale scores resulting from the three scoring procedures were also very similar in value for the students in each subgroup.

Of the three scoring procedures, pattern scoring provided the smallest standard errors, particularly at the lower end of the score scale. This would help ensure that the scores are more precise estimates of student ability, especially for students at the lower end of the scale. However, it has been noted that as more CR items are added, the benefit of lower standard error diminishes (Sykes & Hou, 2003).

Sykes and Hou (2003) also examined the concurrent use of SR and CR items with multiple scoring procedures. Using a procedure similar to the procedure used by Schaeffer et al. (2002), Sykes and Hou (2003) examined the effect of the scoring procedures on a low-stakes Grade 8 writing examination. In addition to the three scoring procedures used by Schaeffer et al. (2002), they added four additional scoring procedures: (i) a weighted score that deliberately increased the weighting of the CR items by a factor of two (CRx2); (ii) a summed score that involved the sum of the two raters scores on the CR items; (iii) a long form in which 18 SR items and eight CR items were added to the examination; and (iv) an all SR long form in which 20 additional SR items were added and the CR items were removed.

Unweighted and Weighted Scoring Procedures.

The unweighted and weighted scores were given as:

$$E = w_m \left\{ \sum_{i=1}^{s.r.} w_i P_i \left(\hat{\theta} \right) + \sum_{j=1}^{c.r.} w_j \sum_{k=1}^{m_j} (k-1) P_{jk} \left(\hat{\theta} \right) \right\}, \tag{27}$$

where the predicted total score has been partitioned into components for the SR items, $s.r.$ and the CR items, $c.r.$. For the unweighted raw score procedure, all weights, w_i and w_j, were equal to one (Sykes & Hou, 2003). The weight w_m, which multiplies each item probability, was used to establish the total number of points in the total score. For the equal weighting scheme, w_j was set to 1 and w_i was set to 1 once the scores were converted so that the SR items and CR items contributed the same number of points toward the total score. For the CRx2 weighting scheme w_j was set to 2 and w_i was set to 1 for each SR item. A conversion table was then produced for each content area that relates the weighted raw score to a non-maximum-likelihood trait estimate using the inverse of the test characteristic function $E\left(X \mid \hat{\theta} \right)$ (Sykes & Hou, 2003).

IRT Pattern Scoring.

For the item scores with the generalized 3PL/2PPC model, the information of the raw score at ability θ is (Sykes & Hou, 2003):

$$I\left(\theta, \sum_l w_l X_l \right) = \frac{\left[w_m \sum_{l=1}^{n} w_l \sum_{k=1}^{m_l} (k-1) P'_{lk}(\theta) \right]^2}{\sum_{l=1}^{n} \sigma^2 \left(w_m w_l X_l \mid \theta \right)}. \tag{28}$$

Pattern scores produced by the 3PL/2PPC model employ implicit item scoring weights (w_i) that are optimal in maximizing reliability and test information (Sykes & Hou, 2003). Employing the optimal weights, test score information is the sum of the test information functions (Sykes & Hou, 2003):

$$I\left(\theta, \sum_{I} w_i X_i\right) = \sum_{i=1}^{n} \sum_{k=1}^{m_i} \frac{\left[P_{ik}^{'}(\theta)\right]^2}{P_{ik}(\theta)} \quad . \qquad (29)$$

Total information for the explicitly weighted items (unweighted, weighted equally, CRx2), and the implicitly weighted items in the IRT pattern scoring were obtained by accumulating the values yielded by Equation 28 and Equation 29, respectfully, over the range of abilities (Sykes & Hou, 2003).

Sykes and Hou (2003) found that pattern scoring provided the smallest standard errors (SEs) across the ability range of all the forms containing CR items. The solely SR long form was found to have the highest test reliability ($\hat{\alpha}$ = 0.90) with the two summed CR form and the CRx2 form having the lowest reliabilities ($\hat{\alpha}$ = 0.84 for both forms). Although increasing the weight of the CR items in the CRx2 form reduced the overall test reliability, the weighting improved the efficiency of the measure in the lower tail of the ability scale. The SEs in the CRx2 form were reduced to less than the SEs obtained from the long form, which represented weighting of CR items through increasing the number of CR items that is possible only when testing time is not constrained.

In summary, of the scoring procedures used in both studies, pattern scoring provided the smallest standard errors, particularly at the lower end of the score scale. This would help ensure that the scores are more precise estimates of student ability, especially for students at the lower end of the scale. The solely SR long form was found to have the highest reliability, which is preferred. However, the CR items were not included in this test, which is a requirement of the current testing protocol. It is also important to note that Schaeffer et al. (2002) and Sykes and Hou (2003) focused only on Grade 8 and Grade 9, low-stakes examinations. It may be possible that high-stakes high-school examinations, in which students are potentially more motivated to perform, may result in significant differences between scoring procedures.

Low-Stakes Tests versus High-Stakes Examinations

Low-stakes tests are tests in which the consequences perceived by the students are low. Student grades are not affected and it is generally perceived as an exercise that must be completed because it is required by government mandate. Students may perceive low-stakes tests as inconsequential to their personal achievement and as a result may not be motivated to work as hard to achieve their best as they would on high-stakes examinations (DeMars, 2000; Kiplinger & Linn, 1992; Paris, Lawton, & Turner, 1992; Wolf, Smith, & Birnbaum, 1995; Wolf & Smith, 1995).

Conversely, high-stakes examinations are examinations in which the consequences of performance directly affect the achievement of the students writing the examinations. It has been demonstrated that average scores on high-stakes examinations are generally higher than average scores on low-stakes examinations (Brown & Walberg, 1993; DeMars, 2000; Kiplinger & Linn, 1993; Wolf & Smith, 1995; Wolf, Smith & Birnbaum, 1995).

For example, Brown and Walberg (1993) examined the effects of motivation on elementary student performance. Two heterogeneously grouped classes at Grades 3, 4, 6, 7 and 8 within each of three schools were sampled for a total sample of 406 students. One class at each grade was assigned to a control condition and the other to an experimental condition. The Mathematics Concepts subtest (Form 7) of the 1978 Iowa Test of Basic Skills (ITBS) was used to measure student achievement. The number of SR items on the test was not noted. Teachers in the experimental condition read an extra set of instructions to the students that called for the students to do their very best "for yourself, your parents, and me [teacher]" and that their scores would be compared to other students in their school and in other schools in Chicago. An analysis of variance showed a significant effect of experimental condition ($F = 10.59$, $p < 0.01$). The mean normal curve equivalent test score of the 214 students in the experimental condition was 41.37 ($SD = 15.41$), and the mean of the control group was 36.25 ($SD = 16.89$). The motivational effect was 0.30 standard deviations, which suggests that the extra set of instructions increased the average students' scores from the 50th to the 62nd percentile.

Wolf and Smith (1995) examined the influence of test consequences on achievement. Two parallel forms of a 40-item SR test were administered to 158 college students in an undergraduate child development class. One form affected the students' grade and was therefore a high-stakes examination; the other form did not and was therefore a low-stakes examination. Form and order of presentation were counterbalanced. Using a repeated measures analysis of variance with test consequence as the within-subjects factor, a significant main effect was found for the condition of consequence versus no consequence ($p < 0.001$).

Kiplinger and Linn (1993) also investigated the effects of test consequence on achievement. Seventeen SR items from the low-stakes 1990 National Assessment of Educational Progress (NAEP) Grade 8 mathematics assessment were embedded in four forms of the high-stakes 1992 Georgia Curriculum-Based Assessment (CBA). The first nine items were included in Test Forms One and Four, while the remaining eight items were included in Test Forms Two and Three. The NAEP items were preceded by different content areas in each test form. A total of 80,836 student records were available for use. The mean for the first nine items was 5.24 (SD = 2.28) in the 1992 high-stakes administration and was higher than the mean of the same items (M = 4.84; SD = 2.16) administered on the low-stakes 1990 NAEP (effect size = 0.18). No significant differences were found between the means for the second set of eight items; the corresponding effect was -0.4. It was suggested that the difference in results from the first nine to the last eight items may be due to (a) the relative difficulty of the items; (b) contextual differences in the administration of the items; or (c) real year-to-year differences in student achievement.

Wolf et al. (1995) also explored the consequence of performance on a low-stakes and a high-stakes math test. The subjects were 168 students in Grade 10 and 133 students in Grade 11. Due to a change in administration in New Jersey, a Grade 9 mathematics test that students were required to pass for high school graduation was moved to Grade 11. During "due-notice" testing in 1992 and 1993, the Grade 11 students wrote the test. However, since they had already written the test in Grade 9, the test held no consequence for them. In some schools, students in Grade 10 were administered the test and the results were used as a major determinant of 11th grade placement into remedial programs. The test consisted of 30 SR items and ten CR items. The data for the CR items were not available for this study and therefore not analyzed.

Wolf et al. reported that the overall performance for the two grade levels was not significantly different. However, the fact that the students in Grade 11 had one more year of math course work and should have performed significantly better than those students in Grade 10 makes the results suspect. The effect of test consequence was noted as a possible variable in this discrepancy. After the test each student was required to answer a question regarding motivation in a list of attitudinal questions. The students were asked to choose, on a four point Likert scale, how hard they worked to answer the questions on the test. The students in Grade 10 showed significantly more motivation on the high-stakes test than the students in Grade 11 on the low-stakes test ($p < 0.001$).

DeMars (2000) examined how scores changed on the science and math sections of Michigan's High School Proficiency Test (HSPT; Michigan Department of Education, 1995) when the stakes of the test were changed. Students participated in the 1994-1995 low-stakes piloting of the test forms or the 1997 high-stakes test for diploma endorsement. The sample included 3,596 students for the low-stakes examination and 8,334 students for the high-stakes examination. There were 34 SR items and eight CR items on the science test and 32 SR items and six CR items on the math test. Two composite scores were estimated for each student, one based on the

SR items and one on the CR items. These estimates were based on the 1-PL and one-parameter partial credit model. A hierarchical linear model HLM4 (Bryk, Raudenbush, & Congdon, 1996) that included both students and schools was used. In both math and science, students scored significantly higher on the high-stakes forms than on the low-stakes forms ($p < 0.001$).

In summary, research demonstrates that student motivation is generally higher in high-stakes testing situations than low-stakes testing situations. It was also demonstrated that this resulting motivation results in higher performance on high-stakes assessments than low-stakes assessments. However, the effect of scoring procedure on student performance has been addressed only on low-stakes examinations (Schaeffer et al., 2002; Sykes & Hou, 2003). It may be possible then, that the scoring procedures used may have differential results when applied to low-stakes and high-stakes examinations with multiple formats. This has not been addressed in the literature and was the purpose of the present study.

High and Low-stakes Testing in Alberta

To address this purpose, two low-stakes tests and two high-stakes examinations were used. The use of two tests at each level allowed an assessment of the stability of the scores yielded by the three procedures to be considered.

The low-stakes tests and high-stakes examinations included the 2003 PATs in Language Arts and Mathematics and the 2003 DIP Examinations in English and Mathematics administered in Alberta. The PATs, which are administered at Grade 3 (Language Arts and Mathematics) and at Grades 6 and 9 (Language Arts, Mathematics, Science, and Social Studies), are low-stakes tests. They are used to provide information to teachers, administrators, school trustees and Alberta Education on how well the students and schools have achieved the learning outcomes set out in the *Programs of Study*, permit comparison of the results of teachers' assessments to the provincial achievement test results, and provide additional feedback to teachers on the effectiveness of their teaching procedures. Another purpose of the PATs is to provide feedback to students and their parents/guardians on how well the students have learned curriculum-based learning outcomes as defined in the *Programs of Study* (Alberta Education, 2004a). The items included in the PATs measure knowledge and skills that are identified in the corresponding provincial curriculum guides. The PATs are administered in May and June in all public and provincially funded independent schools in Alberta.

In contrast, the Alberta DIPs are considered to be high-stakes examinations. The scores from the examinations are combined with a school awarded mark and the blended marks (50% and 50%, respectively) are used to determine whether each student enrolled in a Grade 12 examinable course has passed or not passed the course and to determine scholarship winners. The DIPs are used to certify the level of individual student achievement in the selected Grade 12 courses in which the student is enrolled and in terms of the expected learning outcomes provided in the *Programs of Study*; ensure that province-wide standards of achievement are maintained; and report individual and group results to assist schools, authorities, and the province in monitoring and improving learning. The items included in the high-stakes Alberta Grade 12 diploma examinations are referenced to the learning outcomes that are identified in the corresponding provincial subject area *Programs of Study*. The examinations are scheduled in January, June and August of each year. Each student in an examinable course is required to write the diploma examination for that course (Alberta Education, 2004c).

As mentioned above, the PATs and the DIPs in the areas of language arts and mathematics at the Grade 9 and 12 levels in Alberta were used in the present study (see Tables 3 and 4). By doing so, the comparisons made between the two pairs of tests will not be confounded by different subject matter. However, there were differences in some topics and the level or complexity of the common topics covered between the two grade levels.

Table 3

Low Stakes Tests

Language Arts Grade	Test Name
9	Language Arts PAT
Mathematics **Grade**	**Test Name**
9	Mathematics PAT

Table 4

High Stakes Examinations

Language Arts Grade	Exam Name
12	English Language Arts 30
Mathematics **Grade**	**Exam Name**
12	Pure Math 30

Examination Items and Content

All of the tests contained SR items and CR items. The SR items contained four or five distractors. The CR items were in the form of short answer, computation, extended response, and essays (Alberta Education, 2004a). Selected response items were dichotomously scored while the CR items were polytomously scored by trained raters. The CR items on the Grade 9 mathematics PAT required the students to compute the answers and fill in their responses on a numerical response sheet. However, these items were dichotomously scored. Tables 5 and 6 provide the format, weighting, administration, and writing time information for the low-stakes tests. Tables 7 and 8 provide the format, weighting, administration, and writing time for the high-stakes tests.

The Tables of Specifications for the low-stakes examinations are provided in Appendix A. In Language Arts there were 55 SR items which assess the students' ability to identify and interpret main ideas and make critical analyses by associating meaning and synthesizing ideas. There was also a focus on informational, narrative and poetic texts. The CR items involved personal/narrative and functional writing tasks. The Language Arts PAT was a power test with 120 minutes allotted for Part A and 75 minutes for Part B. An extra 30 minutes was allotted for each component if necessary.

The low-stakes Mathematics PAT items were focused on four main areas: Number, Patterns and Relations, Space and Shape, and Statistics and Probability. Students apply knowledge while interpreting, analyzing, and expressing simple and complex problems. The test was comprised of 44 SR items and 6 numerical response CR items. Students writing the PAT were required to work through each CR problem to a solution and then provide their solution on a machine-scored answer sheet. Marks were not awarded for the process on the PAT, rather only one mark was awarded for each correct solution. The Math PAT was a power test with an allotted 90 minutes and an extra 30 minutes if required.

Table 5

Description of Low-Stakes Grade 9 Language Arts Tests

	Language Arts PAT		
Student Mark	The PATs are not counted toward the students' school marks.		
Format and Weightings	The Grade 9 English PAT referenced to three main topic areas: Narrative/Essay Writing; Functional Writing; and Reading. The examination is made up of two parts:		
	Examination Part	Item Type	Marks
	Part A: Narrative/Essay and Functional Writing (50% of total mark)	1 narrative or essay 1 functional piece	25 20
	Part B: Reading (50% of total mark)	55 multiple-choice	55(one each)
Administration	May and June: The two parts of the exam are written on separate days.		
Writing Time	Part A: 120 minutes; Part B: 75 minutes. An additional 30 minutes is allowed for students to complete each component of the examination		

(Alberta Education, 2004e)

Table 6

Description of Low-Stakes Grade 9 Mathematics Tests

	Mathematics PAT
Student Mark	The PATs are not counted toward the students' school marks.
Format and Weightings	The Mathematics PAT is referenced to four main topic areas: Number, Pattern and Relations; Shape and Space and Statistics; and Probability. The examination consists of: Item Type — Marks 44 multiple-choice — 44 (one each) 6 numerical response — 6 (one each)
Administration	June.
Writing Time	90 minutes: An additional 30 minutes is allowed for students to complete the examination.

(Alberta Education, 2004f)

Tables of Specifications for the high-stakes examinations are provided in Appendix B. Both examinations were in courses provided to students planning to pursue further studies at post-secondary institutions. The English DIP involved a combination of two CR items in Part A and 70 SR items in Part B. The SR and CR items assessed contextual knowledge, comprehension, application and higher level processes. The SR items required the students to respond to a variety of literary texts including poetry and prose. The CR items required the students to respond to personal and critical/analytical queries in paragraph and essay formats. The set of SR and set of CR items were weighted equally (each counts 50% of the total test mark).

Table 7

Description of High-Stakes Grade 12 Language Arts Examinations

	Alberta English Language Arts 30
Standards	Students will develop an understanding and appreciation of the significance and artistry of literature. Students will understand and appreciate language and use it confidently and competently for a variety of purposes, including entry into post-secondary studies or the workplace.
Student Mark	The diploma examination mark and the school-awarded mark each constitute 50% of a student's final mark in English Language Arts 30–1.
Format and Weightings	The English Language Arts 30–1 diploma examination is made up of two parts:

Examination Part	Item Type	% Test (Marks Each)
Part A: Written Response	1 Written Response To Text	10%
	1 Critical/ Analytical Response	20%
Part B: Reading	70 Multiple-Choice	70%

Administration	January, June and August. Part A and Part B are administered on separate days, except for the August administration when they are both written on the same day, at different times.
Writing Time	Part A: 150 minutes; Part B: 150 minutes. An additional 30 minutes is allowed to complete each component.

(Alberta Education, 2004d)

The English DIP was a power test with 150 minutes allotted for both Part A and Part B. An extra 30 minutes was allotted for each component if required.

Table 8

Description of High-Stakes Grade 12 Mathematics Examinations

	Alberta Pure Mathematics 30
Standards	The Pure Mathematics 30 course emphasizes mathematical theory. In pure mathematics, an algebraic and graphical approach is used to solve problems. Deductive and symbolic procedures are used to determine if and under what conditions a concept is true.
Student Mark	The diploma examination mark and the school-awarded mark each constitute 50% of a student's final mark in Pure Mathematics 30.
Format and Weightings	The Pure Math 30 diploma examination is made up of two parts: Part A: Written Response (35%) and Part B: Machine-Scoreable (65%). Examination Part / Item Type / Marks Part A / 3 Written- Response / 15 (five each) Part B / 33 Multiple-Choice / 33 (one each) 6 Numerical Response / 6 (one each)
Administration	January, June and August. Part A and Part B are administered at separate times on the same day.
Writing Time	Part A: 60 minutes, Part B: 90 minutes; An additional 30 minutes is allowed to complete each component.

(Alberta Education, 2004b, 2004i)

The high-stakes Mathematics DIP emphasized mathematical theory and the items assess four main areas: Problem Solving, Patterns and Relations, Shape and Space, and Statistics and Probability. Procedural knowledge, understanding, and application were emphasized. There were 33 SR and nine CR items included on the Mathematics DIP. The CR items included three written response questions and six numerical response items that required the students to work through problems to a solution and record the solution on a dichotomously scored answer sheet. Marks were not awarded for the process, rather only one mark was awarded for each correct solution. The Mathematics DIP was a power test with 60 minutes allotted for both Part A which included the three written response questions and 90 minutes for Part B which included the numerical response and SR items. An extra 30 minutes was allotted for each component if required.

CHAPTER 3

Method

The procedures followed to compare the scores yielded by the unweighted raw score, weighted raw score, and IRT pattern score scoring procedures are described in the present chapter. The low-stakes tests and high-stakes examinations were described at the end of the previous chapter and the Tables of Specifications are provided in Appendix A and Appendix B, respectively. The present chapter begins with a description of the student samples for each test. The second section describes the classical test theory analyses that were conducted. Section three discusses the assumptions of unidimensionality and local independence underlying the use of IRT pattern scoring. The item calibration associated initially with scoring each of the tests is described in the fourth section, followed by presentation of the four scoring procedures. Lastly, the comparative analyses that were conducted are provided.

Student Samples

The numbers of students in language arts and mathematics for the low-stakes tests are indicated in Table 9. Table 10 shows the numbers of students in language arts and mathematics for the high-stakes Grade 12 examinations. Two random samples of 2,000 students were selected without replacement for each test and examination. The scoring and analyses was repeated in each sample to allow estimation of the stability of the results across samples selected from the same population.

Table 9

Low-Stakes Provincial Achievement Test Participation

	Number of Students	
Test	Language Arts	Mathematics
Grade 9 PATs	39,493	39,604

(Alberta Education, 2003d)

Table 10

High-Stakes Grade 12 Diploma Examination Participation

	Number of Students	
Tests	Language Arts	Mathematics
Diploma Examinations	26,566	21,114

(Alberta Education, 2004e)

Classical Test Score Analyses

Descriptive statistics for the two tests and two examinations were computed. This included means, standard deviations of the raw scores, item-test correlations, and reliability for the SR and CR items.

IRT Assumptions

In order to employ IRT models, the assumptions of unidimensionality, local independence, and speededness had to be met. The SR items were evaluated using the three-parameter model which includes a guessing parameter, thus the assumption of lack of guessing was not addressed. The dimensionality of the items was assessed using principal components analysis. The number of eigenvalues greater than one was examined as were the Scree plots. The dimensionality of the SR items was further assessed using NOHARM, which is a non-linear approach (Fraser, 1988). This solution began with one component and then additional components were added to see if a better solution could be attained. Tanaka's (1993) unweighted least squares goodness-of-fit index and the root mean square residual (RMSR) were used to judge the number of components. Tanaka's index has a value of 1.0 if the data fits the model perfectly and 0.0 if the fit is no better than chance. Tanaka's index has no interpretive guidelines, except that a higher value means a better fit. The RMSR has a value of 0.00 if the data fits the model perfectly and has no upper bound. A RMSR equal to or less than four times the reciprocal of the square root of the sample size suggests good model fit (Fraser, 1988).

When the assumption of unidimensionality is met, the assumption of local independence is obtained (Hambleton, Swaminathan, & Rogers, 1991). Speededness was assessed by determining the number of students who did not complete the last three items. Speededness was not to be considered a factor if 95% of the students completed the last three items of the test (Lord, 1980).

Item Calibration

The item parameters for each test were calibrated on the same scale simultaneously (Ercikan et al., 1998; Fitzpatrick et al., 1996; Schaeffer et al., 2002; Sykes & Hou, 2003; Sykes & Yen, 2000). The three-parameter logistic model (3PL: Lord 1980) was used for the SR items. A generalization of Masters' (1982) Partial Credit Model was used for the CR items.

The parameters were estimated using PARDUX (Burket, 1998), a proprietary computer program developed at CTB-McGraw Hill. PARDUX, as described by Schaeffer et al. (2002), uses a marginal maximum likelihood procedure implemented with the EM algorithm (Bock & Aitken, 1981). WINFLUX (Burket, 1999), also developed at CTB-McGraw Hill, was used to place the item parameters onto a common score scale. The scaling parameters, a multiplier of 50 and an additive constant of 500, were used (Schaeffer et al., 2002). The lowest obtainable score (LOSS) and highest obtained scale scores (HOSS) were set at 300 and 700, respectively, to allow for a range of scale scores sufficiently wide to accommodate different weightings of the CR items (Schaeffer et al., 2002; Sykes & Hou, 2003).

Four Scoring Procedures

The SR items and CR items on each examination were scored according to the following four scoring procedures: unweighted raw score (UNW); weighted raw score with CR items worth twice as much as SR items (WCRX2); weighted raw score where the SR items and CR items were weighted equally (WN/M); and IRT pattern score (PTRN).

Unweighted raw score procedure

In the unweighted raw score procedure, a student's score is equal to the number of points earned from the SR items plus the number of points earned from the CR items. The focus of this procedure is the total number of points obtained by the student on the test as a whole (Schaeffer et al., 2002).

Weighted raw score procedure

In the weighted raw score procedure, the CR items are purposefully weighted to a predetermined level so that they count a prescribed amount toward the scale score. Two weighting schemes were examined: one that equally weighted the SR items and CR items (Schaeffer et al., 2002) and the other that doubled the weight (WCRX2) of the CR items (Sykes & Hou, 2003).

For the equal weighting schemes, the students' raw scaled scores were first converted so that the scores were equal to the number of SR items answered correctly plus *n/m* times the number of points earned from the CR items, where n = number of SR items, and m = number of possible CR points (Schaeffer et al., 2002). For Grade 9 Language Arts and Mathematics, the *n/m* weights were, respectively, 55/45 and 88/12. For Alberta DIPs in English and Mathematics, the *n/m* weights were 70/30 and 33/21, respectively. For the equal weighting scheme, Equation (23) with w_j set to 1 and w_i set to 1 were used once the scores are converted so that the SR items and CR items contributed the same number of points toward the total score. The WCRX2 weighting scheme w_j was set to 2 and w_i was set to 1 in Equation (23) for each SR item.

IRT pattern scoring

Pattern scores produced by the 3PL/2PPC model employ implicit item scoring weights (w_i) that are optimal in maximizing reliability and test information (Sykes & Hou, 2003).

Analyses

Group level

Means and standard deviations of scaled scores for the four scoring procedures for each of the low-stakes tests and high-stakes examinations were computed. Differences between the means and variances of the SR items, CR items, and total scores were discussed for both samples. Item-total correlations were computed. Scale scores correlations within language arts and mathematics at low-stakes and high-stakes levels from the each of the four scoring procedures were compared. The variables were as follows:

UNW – The scaled score based on the unweighted raw score.

WCRX2 – The scaled score based on the weighted raw scores, where the CR items contribute twice the number of points possible toward the total score as the SR items.

WN/M – The scaled score based on the weighted raw scores, where the SR items and CR items contribute the same number of points possible toward the total score.

PTRN – The scaled score based on IRT pattern scoring.

SR PTS – The number of SR points earned.

CR PTS – The number of CR points earned.

EXAM – Total number of points earned without weighting (SR PTS + CR PTS)

Given the large sample size, inferential statistical procedures were not employed. As illustrated in the next chapter, the power of these analyses was close to or equal to one. Consequently, small differences were found to be significant. As such an alternative measure of significance was required. The standard error of equating (SEE) first discussed by Lord (1950) is the oldest measure of the statistical accuracy of estimated linking functions. However, in this study the interest was in the accuracy of differences between the four scoring procedures and if there were any important consequences for the reported scores. This issue was addressed by Dorans and Feigenbaum (1994) in their discussion on test equating. Dorans and Feigenbaum

(1994) called a difference in reported score points a "Difference that Matters" (DTM). DTM was defined to evaluate the differences between the scale score means, standard deviations, and pairs of scores (Kolen & Brennan, 2004; Dorans, 2004; Dorans & Feigenbaum, 1994). As a score approaches and crosses a grade threshold, as in the criterion referenced examinations used by Alberta Education, a difference of one scale score point may mean a difference between passing or failing, or deciding on scholarship eligibility. Hence, half a scale score difference defines the DTM (Dorans, 2004). In the case of the correlations, the DTM criterion was set at 0.10, which represents a percentage difference of approximately 10% on the upper half of the correlation scale ($0.00 < r_{XY} \leq 1.000$). Differences equal to or greater than one DTM were not claimed if there was lack of transitivity (e.g., $a < b$, $b < c$, and $a = c$).

The scoring procedures were compared at the group level in terms of the precision of the scores yielded by each procedure. Plots of conditional standard errors of measurement for each were compared at selected points along the ability scale.

Student level

To gain a further understanding of the differences among the scores yielded by the four scoring procedures, scores at the student level were examined. First, differences among the four scores were compared for each student at the 10^{th}, 50^{th}, and 90^{th} percentile score points. Second, the root mean square (RMS) was used to provide a measure of overall fit:

$$RMS = \sqrt{\frac{\sum_{i=1}^{n}(X_{i1} - X_{i2})^2}{n-1}}, \qquad (30)$$

where X_{i1} is the score for one scoring procedure for person i,

X_{i2} is the score for a second scoring procedure for person I, and

n is the number of observations.

Third, all of the tests and examinations were classified into proficiency levels. For the purposes of this study, the proficiency levels were associated with the unweighted raw scores of 50%, 70%, and 85% respectively as in Schaeffer et al. (2002). However, unlike Schaeffer et al., (2002) who examined the proficiency scores at the group level, this study examined how the proficiency levels affected individual students.

Finally, the scoring procedures were evaluated by a comparison of the scores that individual students obtained under the different scoring procedures at the low-stakes and high-stakes levels. Differences in scaled scores between the unweighted scores, two weighted scores and pattern scoring were computed for each student across score points again using a DTM of 0.50.

CHAPTER 4

Analyses and Results of Low Stakes Tests

The analyses and results for the English 9 and Math 9 tests, which, as described in Chapter 2, were considered to be low-stakes tests, are reported in the present chapter. The corresponding results for the two examinations that were considered to be high-stakes, English 30 and Pure Math 30, are presented in Chapter 5. The two chapters are organized in three sections. First, classical test score statistics were examined to determine if the two random samples were randomly equivalent. As will be shown, this was the case for each of the pair of samples for each test and examination. Second, the assumptions of IRT were tested and found to be met for each test and examination. The pattern scores and the unweighted and weighted scores were then computed. Lastly, the degree of fit between each of the pairs of scores yielded by the UNW, WCRX2, WN/M, and PTRN scoring procedures was examined at the group and individual student levels. Both chapters conclude with a summary of results.

English 9

Comparability of Samples.

The summary classical test score statistics for the English 9 samples are reported in Table 11. The variables are as follows:

SR PTS = number of selected response points earned;

CR PTS = number of constructed response points earned;

EXAM = Total number of points earned without weighting (SR PTS + CR PTS).

Table 11

Summary Classical Test Score Statistics: English 9

	Sample 1				Sample 2			
	Mean	SD	Skew	Kurtosis	Mean	SD	Skew	Kurtosis
SR	36.63	8.23	-0.51	-0.24	36.56	8.32	-0.58	-0.26
CR	23.75	4.76	0.14	-0.10	23.76	4.79	0.07	-0.02
Exam	60.38	11.67	-0.28	-0.31	60.32	11.73	-0.39	-0.29

Note: SR PTS = selected response points; CR PTS = constructed response points; EXAM = Total number of points (SR PTS + CR PTS).

The means and standard deviations between the two samples were similar. The differences between the means and standard deviations (sd) were less than one DTM (0.5) for the SR and CR items and exam total for the two samples (0.07 and 0.09 for the means and sd of the SR items, 0.01 and 0.03 for the means and sd of the CR items, and 0.06 for the both the means and sd of the total exam). On average, the students earned slightly higher scores on the SR items about 67% of the maximum SR points possible (36.6 out of 55) than on the CR items, 53% of the maximum CR points (23.8/45). The negative skewness and kurtosis for both samples indicate that English 9 was a relatively easy exam. This finding suggests that there may be problems in obtaining an ability distribution using IRT that is centered on zero with a standard deviation of one.

Given total scores and not item scores were available for the selection and constructed items, it was not possible to compute the reliabilities (internal consistencies) for these scores. However, the internal consistency of the selection items for the total population from which the samples were drawn was 0.86 (Ping Yang, Personal Communication, October 18, 2007). The inter-rater reliability for the constructed response items was not available for the two PATs and two DIPs used in this study. Lastly, the correlations between the selection and constructed responses scores, 0.59 for sample 1 and 0.57 for sample 2, were moderate, suggesting that each item type was measuring, in part, something different (see Table 12). Thus, taken together, these results indicated that the two samples were randomly equivalent and that non-overlapping information was yielded by the selection and constructed response items.

Table 12

Correlations Classical Test Score Statistics: English 9

	Sample 1			Sample 2		
	SR Pts	CR Pts	Exam	SR Pts	CR Pts	Exam
SR Pts	1.00	0.59	0.94	1.00	0.57	0.92
CR Pts	0.59	1.00	0.82	0.57	1.00	0.81
Exam	0.94	0.82	1.00	0.92	0.81	1.00

Note: SR PTS = selected response points; CR PTS = constructed response points; EXAM = Total number of points (SR PTS + CR PTS).

Assumptions of IRT

Unidimensionality. The assumption of unidimensionality was assessed separately for the subset of selection items and subset of constructed response items given each was analyzed separately using item response theory. Principal component analysis yielded 18 components with eigenvalues greater than 1.0 for the SR items in the English 9 test for Sample 1. The eigenvalue for the first component, 6.74, was 5.11 times greater than the eigenvalue of the second component 1.45. Further, the successive differences between remaining components were small (0.18, 0.01, 0.06, 0.01, 0.03, 0.02, 0.02, 0.01, 0.01, 0.0, 0.03, 0.01, 0.02, 0.02, 0.0, and 0.02).

Principal component analysis for the CR items included in the English 9 test yielded one component with an eigenvalue greater than 1.0 for Sample 1. The eigenvalue for the first component was 4.58, which was 5.52 times greater than the eigenvalue of the second component 0.83. Further, the successive differences between remaining components were small (0.37, 0.13, 0.04, 0.02, and 0.02).

In Sample 2, the principal component analysis yielded 14 components with eigenvalues greater than 1.0 for the SR items in the English 9 test. The eigenvalue for the first component, 6.84, was 4.75 times greater than the eigenvalue of the second component 1.44. Further, the successive differences between remaining components were small (0.08, 0.14, 0.04, 0.01, 0.02, 0.02, 0.01, 0.05, 0.01, 0.01, 0.02, 0.01, 0.01, and 0.02).

Principal component analysis for the CR items in Sample 2 yielded one component with an eigenvalue greater than 1.0. The eigenvalue for the first component, 4.65, was 5.71 times greater than the eigenvalue of the second component 0.81. Further, the successive differences between remaining components were small (0.38, 0.10, 0.05, 0.02, and 0.03).

The scree plots (see Appendix C1 to Appendix C4), confirmed the dominance of the first principal component for the SR items and CR items in both Sample1 and Sample 2.

Non-linear factor analysis. To further examine the factor structure of the SR items, non-linear factor analysis (McDonald, 1967) was conducted using NOHARM (Fraser, 1988). The fit indices for the two English 9 samples are presented in Table 13. For both samples, the unidimensional model fit the data well: the changes in the

Table 13

NOHARM Fit Indices for English 9

	Sample 1		Sample 2	
No. of Factors	Tanaka	RMSR	Tanaka	RMSR
1	0.980	0.005	0.982	0.005
2	0.983	0.005	0.985	0.004

fit statistics were marginal when the number of factors was increased from 1 to 2. For example, Tanaka values went up by 0.003 in Sample 1 and 0.002 in Sample 2, and the RMSR remained the same in Sample 1 and decreased by 0.001 in Sample 2.

The results of the principal component analysis, the scree plots, and NOHARM suggested that there was a dominant component underlying the student responses to the CR and SR items on the English 9 examination. Consequently, the assumption of essential dimensionality was met for both sets of items.

Local independence. Given that the assumption of essential unidimensionality was met for both the SR and CR items, the assumption of essential item independence was obtained (Hambleton, et al., 1991) for both the selected response and constructed items in both samples.

Speededness. The percentage of students who did not complete the last three items was calculated. Less than 1% of the students did not complete the last three questions. Thus, it was concluded that speededness was not a factor (Hambleton et al., 1991).

Taken together, the three sets of results presented above indicated that the assumptions for the use of IRT were met for the English 9 test.

Fit among Weighted, Unweighted, and Pattern Scores at the Group Level

The scores yielded by the unweighted scoring procedure (UNW), the weighted scoring procedure in which the constructed response scores were double weighted (WCRX2), the weighted scoring procedure in which the constructed response scores were adjusted so that the constructed response scores counted the same as the selected responses scores (WN/M), and the pattern scoring procedure (PTRN) were compared to each other. The means and standard deviations of the four score distributions are provided in Table 14 and the correlations are provided in Table 15. Given that the scores are correlated and the large sample size, the power of the statistical tests for testing differences among the means, among the variances, and among the correlations is close to one (see Appendix D1).

Inspection of the means in Table 15 reveals that the four scale means for both samples were all less than 500. This occurred because the mean ability estimate on the IRT theta scale was less than zero (-0.04). As suggested earlier, this finding is attributable to the large number of easy items. Consequently, when these scores were transformed, the means were less than 500. Although the test was relatively easy, this transformation yields scores that suggest that the test was not easy, but somewhat difficult. This is an undesirable artifact of the transformation process. However, the transformation was retained given the two previous studies in which the scoring procedures were compared used this transformation (Schaeffer et al. 2002; Sykes & Hou, 2003). Further, for the purposes of this study, which was to determine if the scores yielded by the same procedures were interchangeable, this behavior did not adversely influence the comparisons made.

Table 14

Measures of Central Tendency for the Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 9

	Sample 1				Sample 2			
	Mean	SD	Skew	Kurtosis	Mean	SD	Skew	Kurtosis
UNW	494.31	57.00	-0.07	0.10	497.84	56.36	-0.20	-0.01
WCRX2	493.10	56.29	-0.17	-0.05	496.65	55.48	-0.29	-0.15
WN/M	493.97	56.89	-0.10	0.06	497.48	56.08	-0.22	-0.05
PTRN	494.71	54.68	0.11	0.06	498.29	53.98	0.00	0.05

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two; WN/M – Weighted SR/CR.

While the mean for WCRX2 is less than each of the other means, no significant differences are claimed among the other three scoring procedures due to the lack of transitivity (see Table 14). The same held true for Sample 2.

The standard deviations of the four distributions all exceeded 50 for both samples. Further, for both samples the standard deviation of the distribution of UNW scale scores exceeded the standard deviations of the distributions of the remaining three scale scores by more than one DTM; the differences among the standard deviations of the remaining scale scores were within one DTM. The four score distributions were negatively skewed and leptokurtic. The negative skewness reflects the easiness of the test, and explains why the means of the transformed scores were less than 500. Lastly, the correlations (see Table 15) among the four sets of scores were all above 0.96 for both samples; the differences among the six pairs of correlations were all less than one DTM.

Table 15

Correlations of the Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 9

	Sample 1				Sample 2			
	UNW	WCRX2	WN/M	PTRN	UNW	WCRX2	WN/M	PTRN
UNW	1.00	1.00	1.00	0.97	1.00	1.00	1.00	0.96
WCRX2	1.00	1.00	1.00	0.96	1.00	1.00	1.00	0.96
WN/M	1.00	1.00	1.00	0.97	1.00	1.00	1.00	0.96
PTRN	0.97	0.96	0.96	1.00	0.96	0.96	0.96	1.00

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two; WN/M – Weighted SR/CR.

Standard error. The four scoring procedures were also compared in terms of the precision of the scores yielded by each procedure. Plots of conditional standard errors of measurement for each scoring procedure are shown in Figure 9 and Figure 10 for Samples 1 and 2, respectively.

As shown in both figures, the overall magnitudes of the standard errors of measurement varied from five to 10 transformed score points for the majority of score points. The minimum standard error for all the procedures occurred around the means (between 475 and 575). The standard errors then increased, but much more rapidly for scores above 575 than for scores below 475; at a scale score of 700, the SE grouped around 50 scale points while at a scale score of 300, the SE grouped around 20 score points. Across the scale scores, the four scoring procedures were similar with UNW scoring resulting in marginally higher amounts of error and PTRN scoring resulting in marginally lower amounts of error than the remaining score procedures, particularly for scores at the lower end of the scale score distribution.

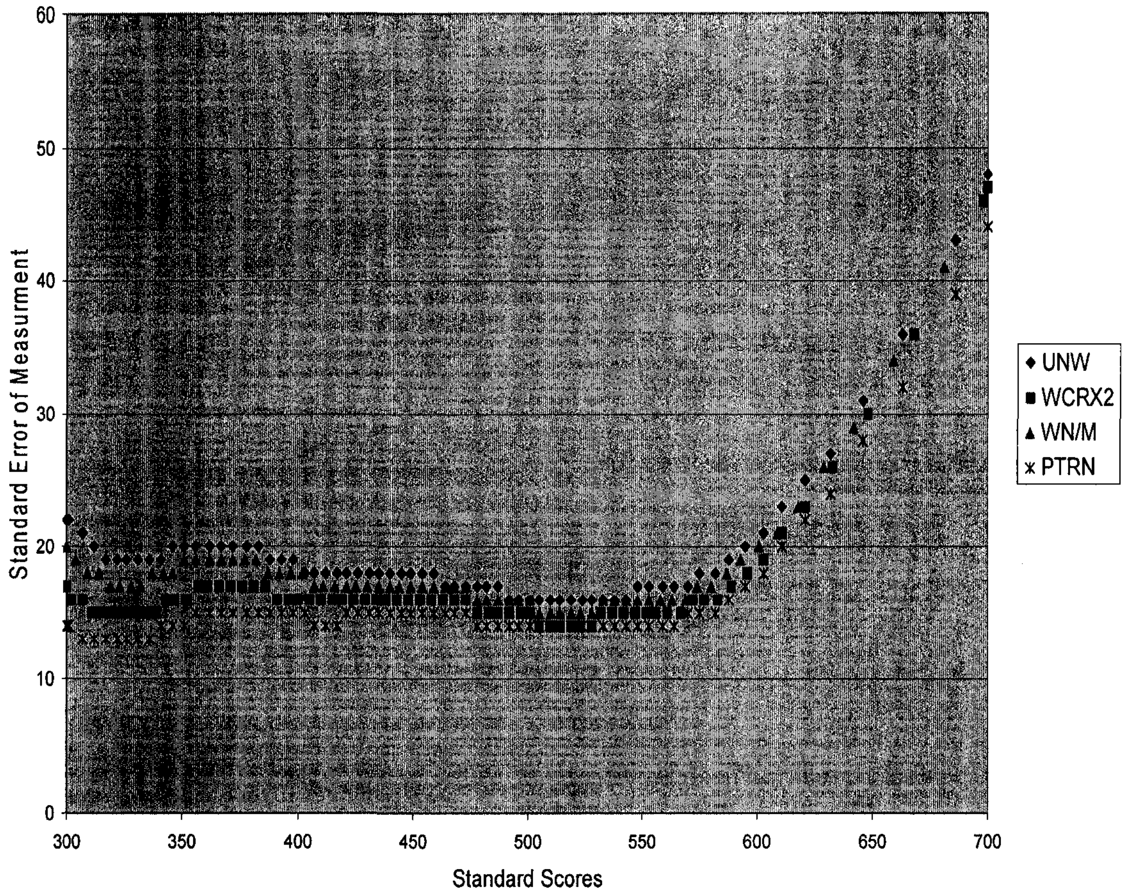

Figure 9. Standard Error of Measurement for English 9 Sample 1

Fit among Weighted, Unweighted, and Pattern Scores at the Student Level

The four scaled scores for each student were compared in four ways. First, the differences among the four scores were compared at the 10th, 50th, and 90th percentile points. Second, the differences between pairs of scores were summarized using the Root Mean Square. Third, the comparability of criterion-referenced decisions was assessed with respect to three cut-score points in the distributions of scores. Finally, individual student score differences were examined.

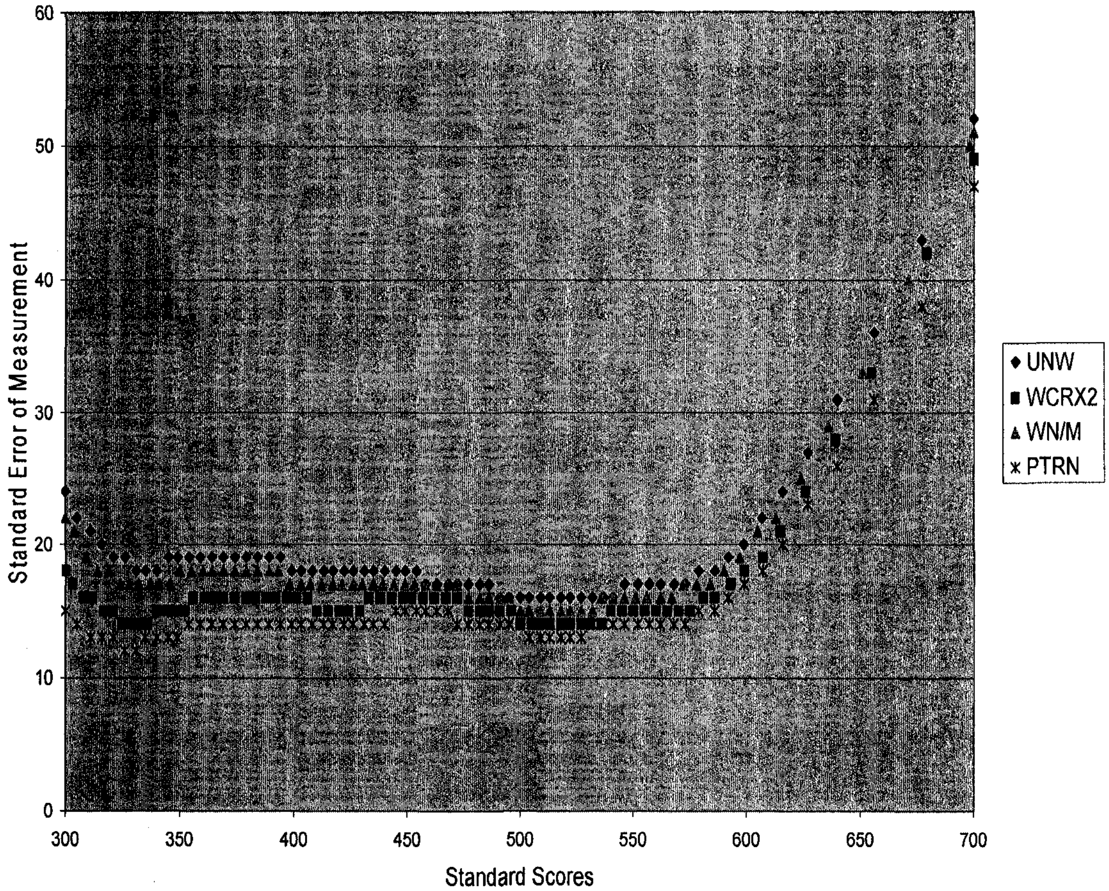

Figure 10. Standard Error of Measurement for English 9 Sample 2

Scale score differences at the 10th, 50th and 90th percentiles. To gain a further understanding of the differences among the scores yielded by the four scoring procedures at the student level, the 10th, 50th, and 90th percentile score points were compared. The results are reported in Table 16 for Sample 1 and Sample 2.

The pattern of percentile scale score differences was similar in both samples. Further, the magnitudes of the differences between pairs of scoring procedures were comparable between the two samples at the three percentile points. Using a DTM of 0.50, the UNW and WN/M scoring procedures were similar at the 10th and 50th

Table 16

Scale Score Differences at the 10th, 50th, and 90th Percentiles: English 9

	Sample 1				Sample 2		
	10th Percentile	50th Percentile	90th Percentile		10th Percentile	50th Percentile	90th Percentile
UNW vs. WCRX2	0.00	1.00	3.00	UNW vs. WCRX2	0.00	1.00	3.00
UNW vs. WN/M	0.00	0.00	1.00	UNW vs. WN/M	0.00	0.00	1.00
WCRX2 vs. WN/M	-2.00	-1.00	0.00	WCRX2 vs. WN/M	-2.00	-1.00	0.00
UNW vs. PTRN	-19.00	0.00	18.00	UNW vs. PTRN	-20.00	0.00	19.00
WCRX2 vs. PTRN	-21.00	-1.00	17.00	WCRX2 vs. PTRN	-22.00	-1.00	18.00
WN/M vs. PTRN	-20.00	0.00	18.00	WN/M vs. PTRN	-20.00	-1.00	19.00

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two; WN/M – Weighted SR/CR.

percentile points but not at the 90th percentile point in both samples, while the UNW and PTRN scoring procedures were similar at the 50th percentile point in both samples and the UNW and WCRX2 were similar at the 10th percentile point in both samples. However, there was lack of transitivity at the 10th and 50th percentiles. Hence, no significant differences among the scores yielded at the 10th and 50th percentiles are claimed. Lastly, the two weighted procedures yielded scores greater than the scores yielded by the UNW procedure at the 90th percentile point, but not to as great an extent as observed when pattern scoring was considered (e.g., 4.00 vs. 18.00, 17.00, and 18.00 in Sample 1 for WCRX2, WN/M and PTRN, respectively).

Root mean square. Table 17 shows that the UNW and WN/M scoring procedures produced very low RMS values, 0.74 for Sample 1 and 0.77 for Sample 2, indicating close agreement between the two sets of scores. This is consistent with the percentile point findings presented above, and again is attributable to the small difference in weights (1.00 vs. 1.22). The agreement between the UNW and WCRX2 scoring procedures and the WCRX2 and WN/M scoring procedures is less: the RMS values were, respectively, 2.37 and 2.40 for Sample 1 and 1.72 and 2.40 for Sample 2. Lastly, the RMS values for the PTRN scoring procedure versus the UNW, WCRX2, and WN/M scoring procedures were much larger, ranging from 14.95 to 15.14 for Sample 1 and 15.61 to 15.86 for Sample 2. The lack of agreement between the PTRN scoring procedure and the other scoring procedures corresponds with the differences reported at the 10th and 90th percentiles reported above.

Table 17

Root Mean Squares of the Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 9

	Sample 1	Sample 2
UNW vs. WCRX2	2.37	2.40
UNW vs. WN/M	0.74	0.77
WCRX2 vs. WN/M	1.72	2.40
UNW vs. PTRN	14.95	15.61
WCRX2 vs. PTRN	15.14	15.86
WN/M vs. PTRN	14.97	15.65

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two; WN/M – Weighted SR/CR.

Proficiency levels. For the purposes of this study, the cut points for the proficiency levels were the scores in the UNW score distribution corresponding, respectively, to percentage scores of 50%, 70%, and 85 % in the observed score distribution. The observed score to UNW scale score conversion table indicated that the corresponding scale scores would be 421, 505, and 575 for Sample 1 and 426, 509, and 579 for Sample 2. These same cut points were used in each of the other three score distributions. The number and percentage of students in each of the four performance levels for each of the four scoring procedures are shown in Table 18 for Sample 1 and Sample 2.

The UNW and WN/M scoring procedures classified the same number of students at Level 1 and Level 2. This was expected as the previous results have all suggested that the UNW and WN/M scoring procedures yielded scores that were similarly distributed. The WCRX2 scoring procedure resulted in 26 more students placed in the first level and 26 fewer in the second level than the UNW and WN/M

Table 18

Proficiency Levels of the Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 9

	Sample 1									Sample 2							
	UNW		WCRX2		WN/M		PTRN			UNW		WCRX2		WN/M		PTRN	
	#	%	#	%	#	%	#	%		#	%	#	%	#	%	#	%
Level 1									Level 1								
(300 – 420)	193	9.7	219	11.0	193	9.7	169	8.5	(300 – 425)	220	11.0	240	12.0	220	11.0	188	9.4
Level 2									Level 2								
(421 – 504)	889	44.5	863	43.2	889	44.5	991	49.6	(426 – 508)	873	43.7	853	42.7	873	43.7	973	48.7
Level 3									Level 3								
(505 – 574)	766	38.3	805	40.3	805	40.3	687	34.4	(509 – 578)	766	38.3	792	39.6	792	39.6	707	35.4
Level 4									Level 4								
(575 – 700)	152	7.6	113	5.7	113	5.7	153	7.7	(579 – 700)	141	7.1	115	5.8	115	5.8	132	6.6

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two; WN/M = Weighted SR/CR; # = number of students; % = percentage of Students

procedures in Sample 1; the comparable numbers in Sample 2 were slightly different: 20 greater and 20 fewer. The PTRN scoring procedure classified 24 fewer students at level 1 and 102 more students at level 2 than the UNW and WN/M scoring procedures in Sample 1; the comparable numbers in Sample 2 were 32 fewer at level 1 and 100 more. The WCRX2 and WN/M scoring procedures classified the same number of students at Level 3 and Level 4. The UNW scoring procedure resulted in 39 fewer students placed in the third level and 39 more in the fourth level than the WCRX2 and WN/M scoring procedures in Sample 1; the comparable numbers in Sample 2 were slightly different: 26 fewer and 26 greater. For the PTRN scoring procedure there were: 118 fewer at third level and 40 greater at Level 4 the WCRX2 and WN/M scoring procedures in Sample 1; the comparable numbers in Sample 2 were 85 fewer at Level 3 and 17 greater at Level 4. While the differences in the corresponding percentages are relatively small (all less than seven percent), the number of students placed in the levels did vary, with up to 100 students being placed at a different level if the UNW or WN/M procedures were used in place of PTRN procedure at the first and second levels, and up to 118 students if the PTRN procedure was used in place of the WCRX2 and WN/M procedures at the third and fourth levels.

Difference in individual student scaled scores. To gain a further understanding of the differences among the four scoring procedures, the distributions were compared across score points. A graphical representation of the distribution of differences between scaled scores yielded by the UNW, WCRX2, WN/M, and PTRN scoring procedures is provided in Figure 11 for Sample 1 and Figure 12 for Sample 2.

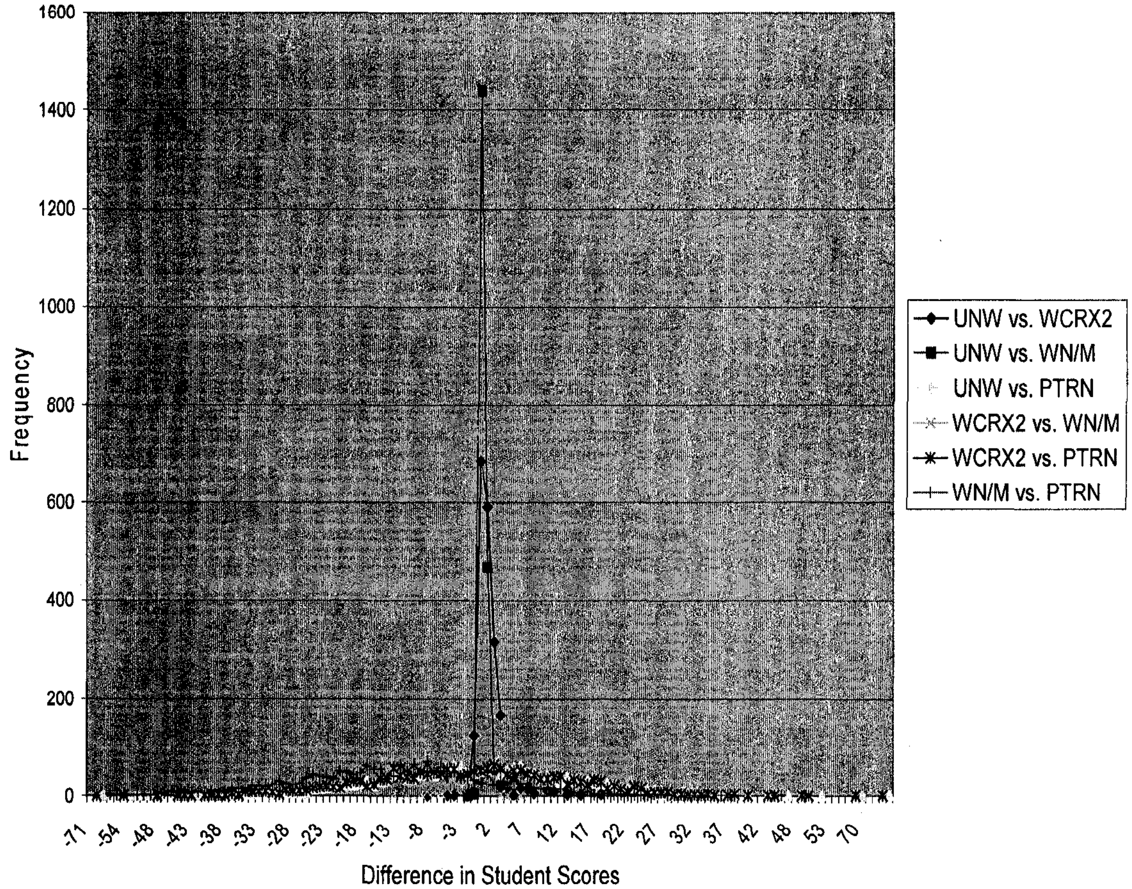

Figure 11. Differences Between Unweighted, Weighted CRx2, Weighted N/M and Pattern Scores: Sample 1

The corresponding tables are provided in Appendix E1 and E2. As shown, the greatest differences occurred when the pattern scores were involved. In these cases, the differences ranged from -59 to 85 in Sample 1 and -72 to 96 in Sample 2 for the UNW procedure, -60 to 84 in Sample 1 and -70 to 94 in Sample 2 for the WCRX2 procedure, and -59 to 85 in Sample 1 and -71 to 95 in Sample 2 for the WN/M procedure. Individual student scores did vary somewhat between the UNW and WCRX2 procedures, which ranged from -8 to 18 in Sample 1 and -10 to 22 in Sample 2; and the WCRX2 and WN/M procedures, which ranged from -12 to 6 in

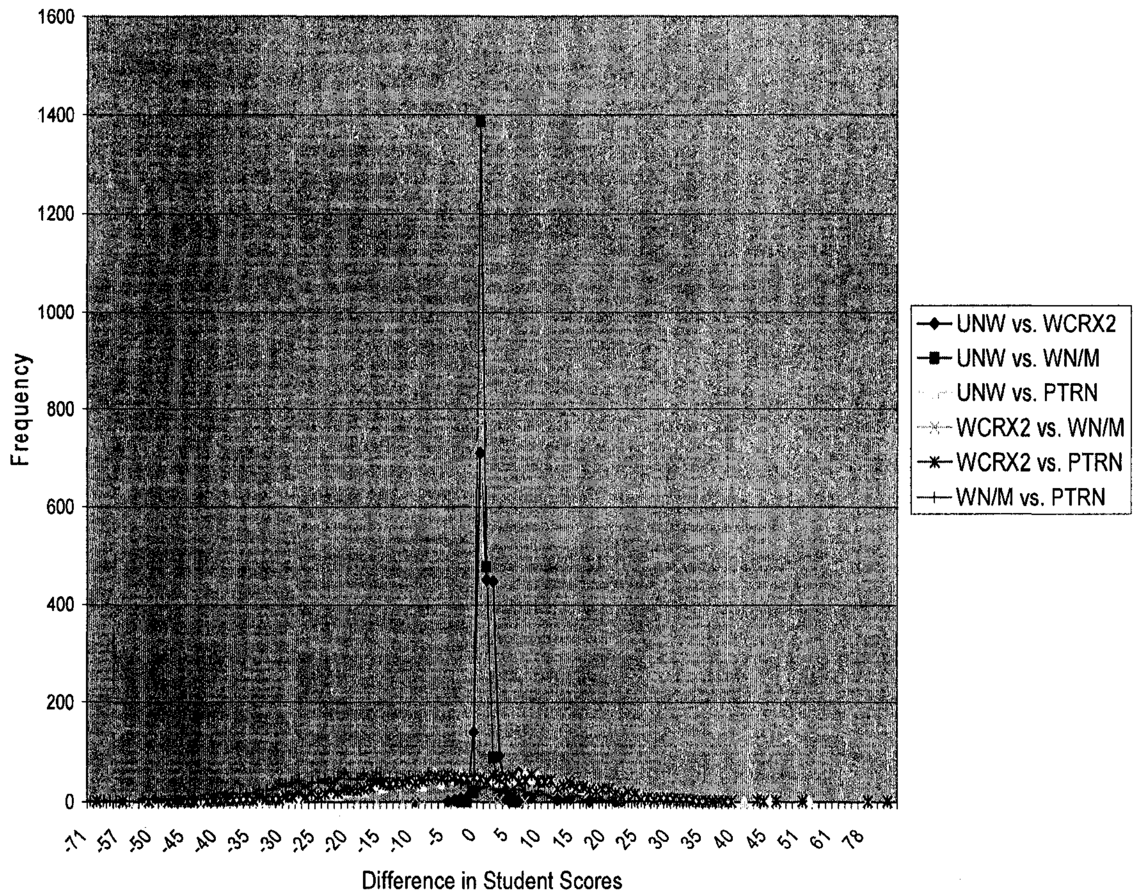

Figure 12. Differences Between Unweighted, Weighted CRx2, Weighted N/M and Pattern Scores: Sample 2

Sample 1 and -19 to 7 in Sample 2. The differences between the UNW and WN/M scoring procedures were smaller, varying from -2 to 5 in Sample and -3 to 6 in the second sample. This latter finding is again attributable to the small difference in weights (1.0 vs. 1.2).

When using a DTM of 0.50, the differences yielded by the four scoring procedures were significantly different for the vast majority of students for all pairs of scoring procedures except the UNW and WN/M pair. For example, in Sample 1 the scores yielded by the UNW procedure and the PTRN scoring procedure were within one DTM for 52 (2.6%) students. The corresponding numbers for each of the weighted procedures and the PTRN procedure were 53 (2.7%) and 64 (3.2%) for the WCRX2 and WN/M respectively. In contrast, the UNW and WCRX2 scores for 684 (34.2%) students, the WCRX2 and WN/M scores for 734 (36.7%), and the UNW and WN/M scores for 1439 (72.0%) students were within one DTM. If the DTM is relaxed to 1.00, the corresponding numbers, taken in the same order, are 153 (7.7%), 163 (8.2%), 156 (7.8%), 1399 (70.0%) 1626 (81.3%), and 1911 (95.6%). The results are similar in Sample 2.

Mathematics 9

Comparability of Samples

The classical test score statistics for the Math 9 samples are reported in Table 19. The means and standard deviations between the two samples were similar. The differences between the means and standard deviations (sd) were less than one DTM for the SR and CR items and exam total for the two samples (0.19 and 0.15 for the means and sd of the SR items, 0.02 and 0.04 for the means and sd of the CR items, and 0.19 and 0.22 for the means and sd of the total exam). On average, the students earned about 68% of the maximum SR points possible (29.7 out of 44) and earned about 65% of the maximum CR points (3.9/6). These results combined with the negative skewness and kurtosis for both samples indicate that Math 9 was a relatively easy exam. As with English 9, this finding suggests that there may be problems in

Table 19

Summary Classical Test Score Statistics: Math 9

	Sample 1				Sample 2			
	Mean	SD	Skew	Kurtosis	Mean	SD	Skew	Kurtosis
SR	29.79	8.45	-0.39	-0.70	29.62	8.60	-0.38	-0.72
CR	3.88	1.75	-0.58	-0.63	3.86	1.79	-0.56	-0.67
Exam	33.66	9.84	-0.43	-0.68	33.47	10.06	-0.43	-0.65

Note: SR = selected response; CR = constructed response; EXAM = Total number of points (SR + CR).

obtaining an ability distribution using IRT that is centered on zero with a standard deviation of one. Like English 9, total scores and not item scores were available for the selection and constructed items, it was not possible to compute the reliabilities (internal consistencies) for these scores. The internal consistency of the selection items for the total population from which the samples were drawn was 0.92 (Ping Yang, Personal Communication, October 18, 2007). Lastly, the correlations between the selection and constructed responses scores, 0.77 for Sample 1 and 0.76 for Sample 2, were moderately large, suggesting that each item type was measuring, in part, something different (see Table 20).

Table 20

Correlations Classical Test Score Statistics: Math 9

	Sample 1			Sample 2		
	SR	CR	Exam	SR	CR	Exam
SR	1.00	0.77	0.99	1.00	0.76	0.99
CR	0.77	1.00	0.84	0.76	1.00	0.83
Exam	0.99	0.84	1.00	0.99	0.83	1.00

Note: SR = selected response; CR = constructed response; EXAM = Total number of points (SR + CR).

Taken together, these results indicate that the two samples were randomly equivalent and that non-overlapping information was yielded by the selection and constructed response items.

Assumptions of IRT

Unidimensionality. Principal component analysis yielded 9 components with eigenvalues greater than 1.0 for the SR items in the Math 9 test for Sample 1. The eigenvalue for the first component, 8.41, was 5.63 times greater than the eigenvalue of the second component 1.49. Further, the successive differences between remaining components were small (0.35, 0.03, 0.01, 0.04, 0.02, 0.02, and 0.01).

Principal component analysis for the CR items include in the Math 9 test yielded one component with an eigenvalue greater than 1.0 in Sample 1. The eigenvalue for the first component was 2.39, which was 2.84 times greater than the eigenvalue of the second component 0.84, and the successive differences between remaining components were small (0.04, 0.10, 0.04, and 0.05).

In Sample 2, the principal component analysis yielded seven components with eigenvalues greater than 1.0 for the SR items in Math 9 test. The eigenvalue for the first component, 8.17, was 5.42 times greater than the eigenvalue of the second component 1.51; the successive differences between remaining components were small (0.35, 0.02, 0.07, 0.01, 0.01, and 0.03). Lastly principal component analysis for the CR items in Sample 2 yielded one component with an eigenvalue greater than 1.0. The eigenvalue for the first component, 2.30, was 2.74 times greater than the eigenvalue of the second component 0.84. Again, the successive differences between remaining components were small (0.03, 0.03, 0.13, and 0.04).

The scree plots (see Appendix C5 to Appendix C8), confirm the dominance of the first principal component for the SR items and CR items in both Sample 1 and Sample 2.

Non-linear factor analysis. As with English 9, the non-linear factor analyses results revealed that the fit statistics did not change when moving from one to two factors (see Table 21). For both samples, the unidimensional model fit the data well.

The results of the principal component analysis, the scree plots, and NOHARM suggested that a dominant component underlay the student responses to the SR and CR items on the Math 9 test. Consequently, the assumption of essential dimensionality was met for both sets of items.

Local independence. Given that the assumption of essential unidimensionality was met for both the SR and CR items, the assumption of essential item independence was obtained for both the selected response and constructed items in both samples.

Speededness. Like English 9, less than 1% of the students did not complete the last three questions. Thus, it was concluded that speededness was not a factor.

Table 21

NOHARM Fit Indices for Math 9

	Sample 1		Sample 2	
No. of Factors	Tanaka	RMSR	Tanaka	RMSR
1	0.988	0.005	0.992	0.004
2	0.988	0.005	0.992	0.004

Taken together, the three sets of results presented above indicate that the assumptions for the use of IRT were met for Math 9.

Fit among Weighted, Unweighted, and Pattern Scores at the Group Level

The means and standard deviations for the UNW, WCRX2, WN/M, and PTRN scoring procedures are provided in Table 22 and the correlations are provided in Table 23.

As shown in Table 22, the four scale means were again less than 500 for both samples. Again, this occurred because the mean ability estimate on the IRT theta scale was less than zero (-0.20 in Sample 1). The means for the UNW vs. PTRN differed by less than one DTM (0.06) in Sample 1; the difference between the other members in each pair was greater than one DTM (e.g., 1.19 for WCRX2 vs. PTRN in Sample 1). The same held true for Sample 2 with a DTM less than one for UNW vs. PTRN (0.41) and the difference between the other members in each pair greater than one DTM (e.g., 0.92 for WCRX2 vs. PTRN). The standard deviations of the four distributions all exceeded 50 for both samples. Further, for both samples the standard

Table 22

Measures of Central Tendency for the Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Math 9

	Sample 1				Sample 2			
	Mean	SD	Skew	Kurtosis	Mean	SD	Skew	Kurtosis
UNW	492.23	63.87	-0.18	1.43	489.98	63.74	-0.30	1.22
WCRX2	493.48	60.85	-0.03	1.47	491.31	61.13	-0.16	1.19
WN/M	496.33	54.25	0.48	1.57	494.08	54.89	0.26	1.09
PTRN	492.29	64.21	-0.20	1.40	490.39	63.45	-0.25	1.09

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two; WN/M – Weighted SR/CR.

deviation of the distributions of UNW and PTRN scale scores were within one DTM, the differences among the standard deviations of the remaining scale scores exceeded one DTM. With the exception of WN/M the scoring distributions were negatively skewed and all four scoring distributions were leptokurtic. The correlations (see Table 23) among the four sets of scores were all above 0.98 for both samples; the differences among the six pairs of correlations were all less than one DTM. Taken together, the results reveal that the scoring procedures at the group level tended to rank the students the same but differed in their central tendency and variability with the exception of the UNW and PTRN scale scores which had similar means and standard deviations.

Table 23

Correlations of the Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Math 9

	Sample 1				Sample 2			
	UNW	WCRX2	WN/M	PTRN	UNW	WCRX2	WN/M	PTRN
UNW	1.00	1.00	0.99	0.99	1.00	1.00	0.99	0.99
WCRX2	1.00	1.00	0.99	0.99	1.00	1.00	1.00	0.99
WN/M	0.99	0.99	1.00	0.98	0.99	1.00	1.00	0.99
PTRN	0.99	0.99	0.98	1.00	0.99	0.99	0.99	1.00

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two; WN/M – Weighted SR/CR.

Standard error. As shown in Figures 13 and 14, the standard errors of measurement for each scoring procedure tended to follow a parabolic distribution. The magnitudes of the standard errors of measurement varied from 10 (around the mean) transformed score points to 40 (around the upper and lower ends) transformed

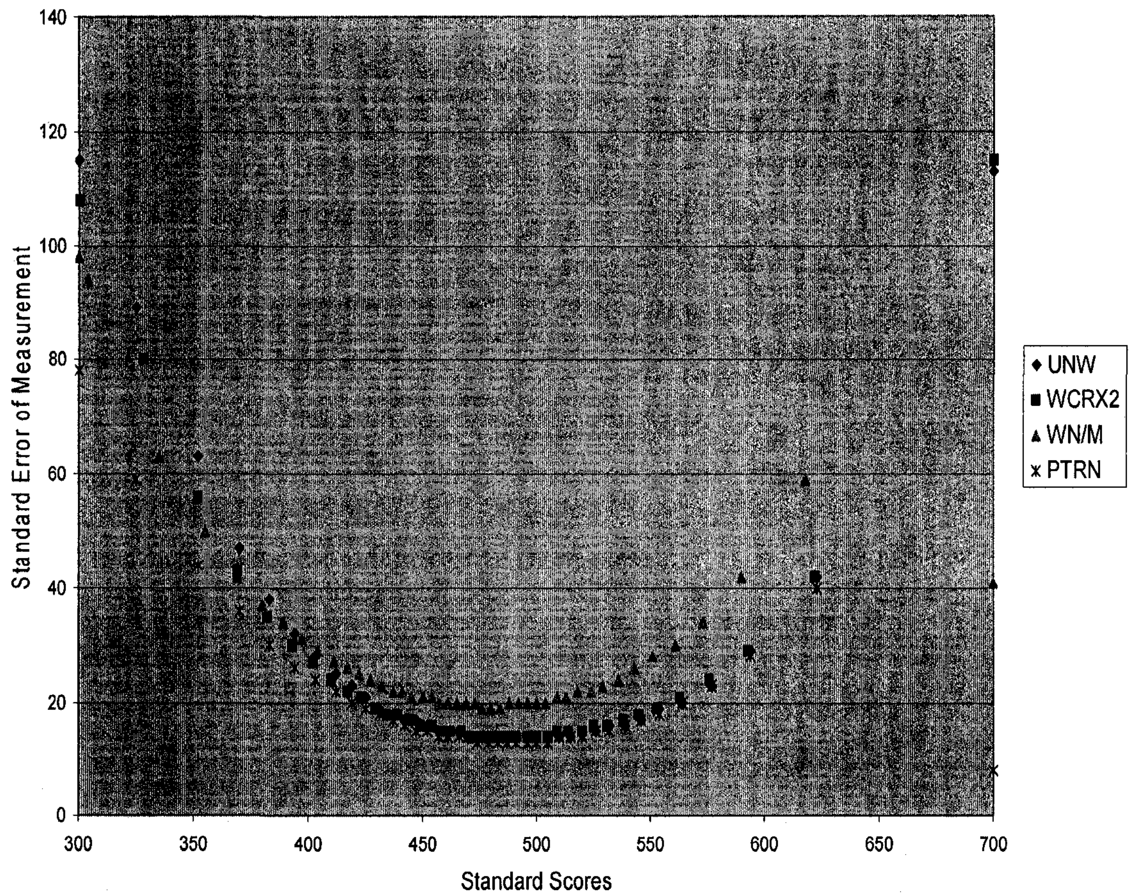

Figure 13. Standard Error of Measurement for Math 9 Sample 1

score points for the majority of score points. The lowest standard errors occurred between 400 and 575, with a sharp increase for scores below 400 and above 575. For scale scores about 400, the UNW and WN/M standard errors crossed with the resulting UNW standard errors comparable with the WCRX2 and PTRN distributions and the WN/M standard errors markedly increasing. Across the scale scores, PTRN scoring resulted in marginally lower amounts of error than the remaining score procedures, particularly at the low end of the scale score distribution. Taken together, these findings suggest that there is a less precise measurement of student ability at the higher and the lower ends of the distribution than at the center of the distribution

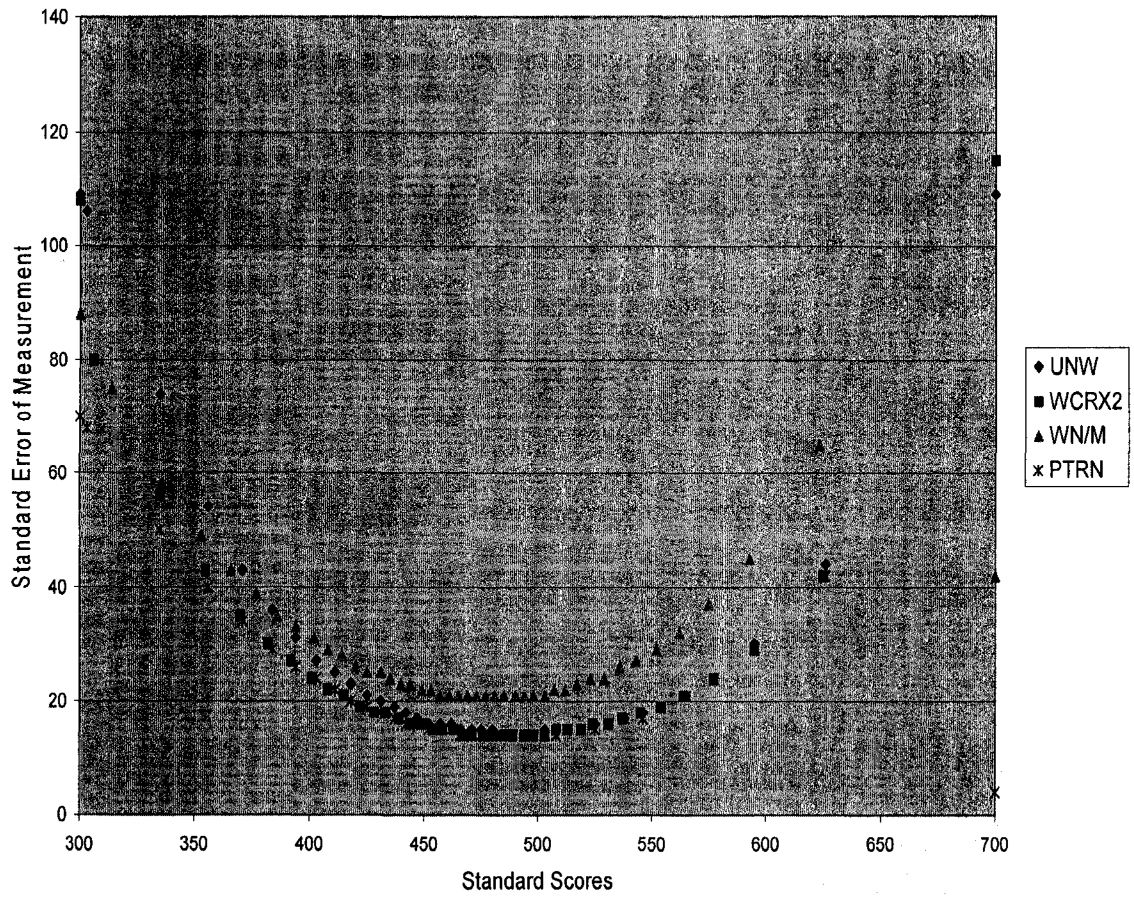

Figure 14. Standard Error of Measurement for Math 9 Sample 2

for all four scoring procedures.

Scale score differences at the 10^th^, 50^th^ and 90^th^ percentiles. The pattern and magnitudes of the scale score differences at the 10^{th}, 50^{th}, and 90^{th} percentile points were similar in both samples (see Table 24). Using a DTM of 0.50, all of the scoring procedures were similar at the 50^{th} percentile with the exception of WN/M vs. PTRN in Sample 1. The differences exceeded 0.50 in absolute value at the 10^{th} and 90^{th} percentile points. In the case of the 10^{th} percentile, the differences were negative, ranging from -4.00 to -15.00; in contrast the differences were positive at the 90^{th} percentile point, ranging from 1.00 to 16.00. That is, the pattern scores at the 10^{th}

Table 24

Scale Score Differences at the 10^{th}, 50^{th}, and 90^{th} Percentiles: Math 9

	Sample 1				Sample 2		
	10^{th} Percentile	50^{th} Percentile	90^{th} Percentile		10^{th} Percentile	50^{th} Percentile	90^{th} Percentile
UNW vs. WCRX2	-4.00	0.00	1.00	UNW vs. WCRX2	-4.00	0.00	1.00
UNW vs. WN/M	-15.00	0.00	3.00	UNW vs. WN/M	-14.00	0.00	2.00
WCRX2 vs. WN/M	-4.00	0.00	1.00	WCRX2 vs. WN/M	-10.00	0.00	2.00
UNW vs. PTRN	-7.00	0.00	7.00	UNW vs. PTRN	-8.00	0.00	7.00
WCRX2 vs. PTRN	-7.00	0.00	9.00	WCRX2 vs. PTRN	-7.00	0.00	8.00
WN/M vs. PTRN	-6.00	0.00	16.00	WN/M vs. PTRN	-6.00	1.00	15.00

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two; WN/M – Weighted SR/CR.

percentile point were less than the scores yielded by the other three scoring procedures, and the WRCX2 and WN/M scores were less than the UNW scores at the 10th percentile; but PTRN and UNW were greater at the 90th percentile point.

Root mean square. Table 25 shows that the UNW and WCRX2 scoring procedures produced relatively lower RMS values, 4.59 for Sample 1 and 4.20 for Sample 2, than the other scoring procedures. However, these values do not indicate much agreement between the two sets of scores. The RMS values between UNW and PTRN, and WCRX2 and WN/M scoring procedures were, respectively, 9.78 and 9.59 for Sample 1 and 9.36 and 8.89 for Sample 2. Lastly, the RMS values for the PTRN scoring procedure versus the WCRX2, and WN/M scoring procedures and those between UNE and WN/M were much larger, ranging from 10.58 to 15.83 for Sample 1 and 9.29 to 13.73 for Sample 2.

Table 25

Root Mean Squares of the Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Math 9

	Sample 1	Sample 2
UNW vs. WCRX2	4.59	4.20
UNW vs. WN/M	13.56	12.49
WCRX2 vs. WN/M	9.59	8.89
UNW vs. PTRN	9.78	9.36
WCRX2 vs. PTRN	10.58	9.29
WN/M vs. PTRN	15.83	13.73

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two;
WN/M – Weighted SR/CR.

Proficiency levels. The observed score to UNW scale score conversion table for Math 9 indicated corresponding cut scores would be, respectively, 452, 496, and 539 for Sample 1 and 447, 494, and 538 for Sample 2. These same cut points were used in each of the other three score distributions. The number and percentage of students in each of the four performance levels for each of the four scoring procedures are shown in Table 26 for Sample 1 and Sample 2.

The UNW and WCRX2 scoring procedures classified the same number of students at Level 2 in Sample 1 and Levels 1 and 2 in Sample 2. The WN/M scoring procedure classified 35 more students in the second level in Sample 1; the comparable numbers in Sample 2 were slightly different: 40 fewer for Level 1 and 40 more students for Level 2 than the UNW and WCRX2 scoring procedures. With the PTRN scoring procedure there were 37 more students at Level 2 in Sample 1; the comparable numbers in Sample 2 were 23 more students at Level 1 and 30 fewer students at Level 2 than the UNW and WCRX2 scoring procedures. The WCRX2 and WN/M scoring procedures classified the same number of students at Level 3 and Level 4. The UNW, WCRX2, and WN/M procedures classified the same number of students at Levels 3 and 4 in Sample 2. The UNW scoring procedure classified 53 fewer students in the third level and 53 more in the fourth level in Sample 1 than the WCRX2 and WN/M scoring procedures. For the PTRN scoring procedure there were 66 fewer students at the third level and 28 more students at Level 4 in Sample 1; the comparable numbers in Sample 2 were 37 fewer students at Level 3 and 37 more student at Level 4 than the UNW, WCRX2, and WN/M scoring procedures in Sample 2. While, the differences in the corresponding percentages are

Table 26

Proficiency Levels of the Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Math 9

	Sample 1									Sample 2							
	UNW		WCRX2		WN/M		PTRN			UNW		WCRX2		WN/M		PTRN	
	#	%	#	%	#	%	#	%		#	%	#	%	#	%	#	%
Level 1 (300 – 407)	436	21.8	409	20. 5	374	18.7	410	20.5	Level 1 (300 – 415)	440	22.0	440	22.0	400	20.0	463	23.2
Level 2 (408 – 485)	574	28.7	574	28.7	609	30.5	611	30.6	Level 2 (416 – 490)	600	30.0	600	30.0	640	32.0	570	28.5
Level 3 (486 – 549)	589	29.5	642	32.1	642	32.1	576	28.8	Level 3 (491 – 553)	587	29.4	587	29.4	587	29.4	557	27.9
Level 4 (550 – 700)	428	21.4	375	18.8	375	18.8	403	20.2	Level 4 (554 – 700)	373	18.7	373	18.7	373	18.7	410	20.5

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two; WN/M = Weighted SR/CR; # = number of students; % = percentage of Students

relatively small (all less than two percent), the number of students classified in the levels did vary, with up to 70 students at a different level if the UNW scoring procedure was used in place of the PTRN scoring at the first and second levels in Sample 2 and up to 66 students if PTRN was used in place of WCRX2 or WN/M at Level 3 in Sample 1.

Difference in individual student scaled scores. The distributions of differences between scaled scores yielded by the four scoring procedures is provided in Figure 15 for Sample 1 and Figure 16 for Sample 2 (see Appendix E3 and E4 for the corresponding tables). As shown, the greatest differences at both extremes occurred between the UNW and PTRN and the WCRX2 and PTRN scoring procedures. In these cases, the differences ranged from -58 to 112 in Sample 1 and -73 to 94 in Sample 2 for the UNW and PTRN procedures, and from -52 to 117 in Sample 1 and -50 to 101 in Sample 2 for the WCRX2 and PTRN procedures. Large negative differences were also found between the scores yielded by the UNW and WN/M procedures and large positive differences between the WN/M and PTRN scores. In these cases, a difference -80 in Sample 1 and -74 in Sample 2 was found for the UNW and WN/M procedures; 127 in Sample 1 and 114 in Sample 2 for the WN/M and PTRN procedures. Individual student scores did vary somewhat between the UNW and WCRX2 scores, which ranged from -28 to 1.

When using a DTM of 0.50, the differences yielded by the four scoring procedures were significantly different for the vast majority of students for all pairs

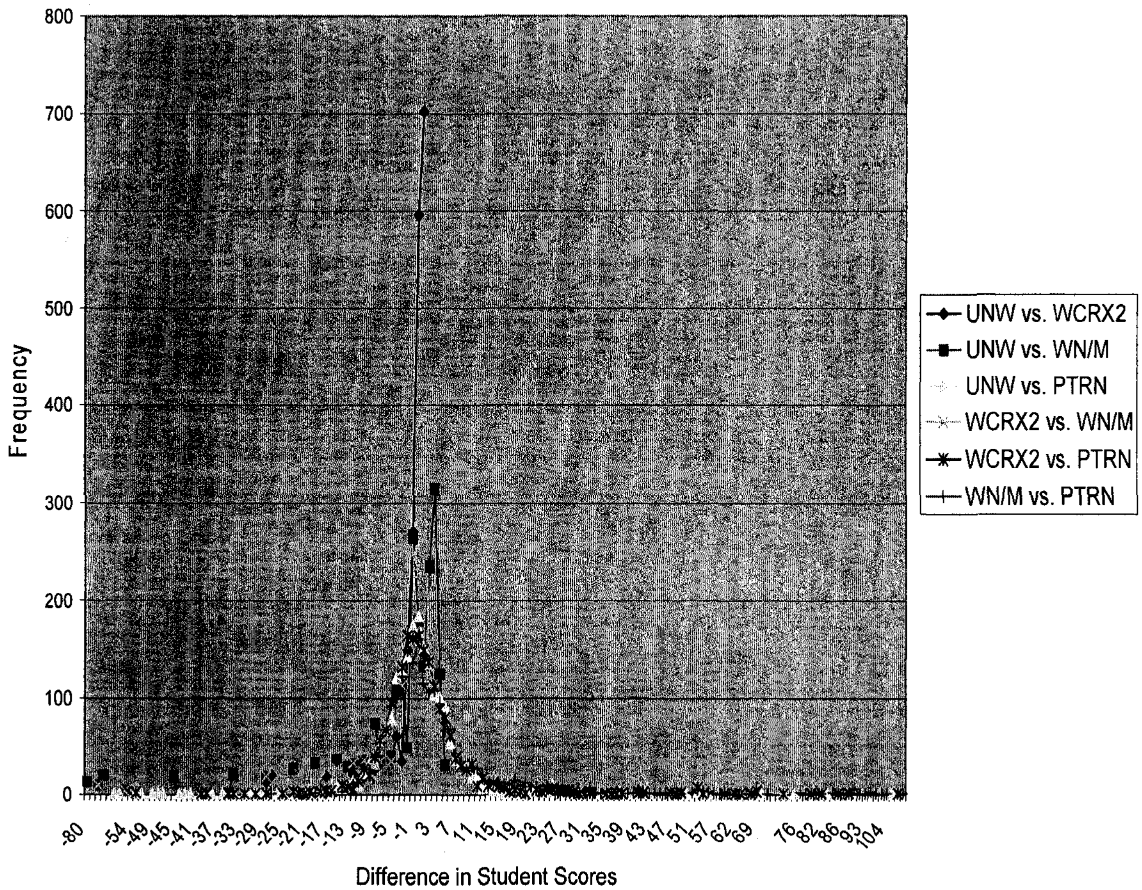

Figure 15. Differences Between Unweighted, Weighted CRx2, Weighted N/M and Pattern Scores: Sample 1

of scoring procedures perhaps with the exception of the UNW and WCRX2 procedures. For example, in Sample 1 the scores yielded by the UNW procedure and the PTRN procedure were within one DTM for 184 (9.2%) students. The corresponding numbers for each of the weighted procedures and the PTRN procedure were 164 (8.2%) for WCRX2 and 159(8.0%) for WN/M. The number of students within one DTM for the UNW and WCRX2 is slightly greater at 596 (29.8%). If the DTM is relaxed to 1.00, the corresponding numbers, taken in the same order, are 510

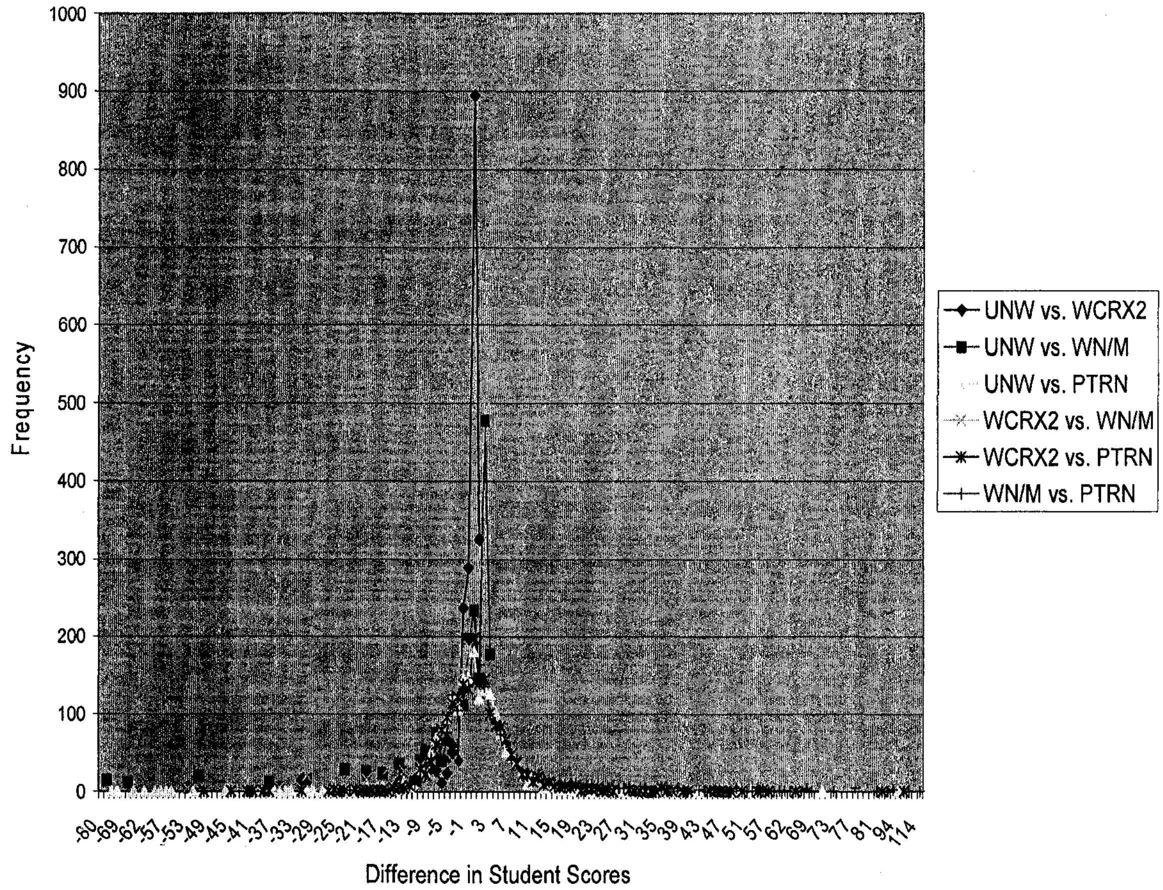

Figure 16. Differences Between Unweighted, Weighted CRx2, Weighted N/M and Pattern Scores: Sample 2

(25.5%), 474 (23.7%), 410(20.5%), and 1568 (78.4%). The results are similar in Sample 2.

Summary

The classical test score statistics for English 9 and Math 9 indicated that the two samples were randomly equivalent and that non-overlapping information was yielded by the SR and constructed CR items, but not to the same extent for both subjects. Through analysis with principal component analysis, scree plots and, in the case of the selection items, NOHARM and the accompanying fit statistics, the assumptions of unidimensionality and item independence were met. The assumption of speededness was also met. Thus the assumptions of IRT were met.

At the group level, the means of the score distributions were all less than 500 varying from 493.10 (WCRX2) and 494.71 (PTRN) for English 9 and from 492.23 (UNW) to 496.33 (WN/M) for Math 9. While the means of the UNW and WN/M and the UNW and PTRN scoring procedures tended to yield more comparable scores, there was a lack of transitivity for English 9. In Math 9, the means for the UNW vs. PTRN differed by less than one DTM and the difference between the other members in each pair was greater than one DTM. The standard deviations of the four distributions all exceeded 50 for both samples on both tests. In English 9, the standard deviation of the distribution of UNW scale scores exceeded the standard deviations of the distributions of the remaining three scale scores by more than one DTM; the differences among the standard deviations of the remaining scale scores were within one DTM. For Math 9, the standard deviations of the distributions of UNW and PTRN scale scores were within one DTM, the differences among the standard deviations of the remaining scale scores exceeded one DTM. With the exception of WN/M in Math 9, the score distributions were negatively skewed and leptokurtic for both samples on both tests. The negative skewness reflects the easiness of the test, and explains why the means of the transformed scores were less than 500. Lastly, the correlations among the four sets of scores were all above 0.96 for English 9 and 0.98 for Math 9; the differences among the six pairs of correlations were all less than one DTM for both tests. Taken together, the results first reveal that the scoring procedures at the group level tended to rank the students the same but differed in their central tendency and variability with the exception of the UNW and PTRN scale scores in Math 9.

The magnitudes of the standard errors of measurement varied from 5 to 10 transformed score points (English 9) and 10 points for Math 9 for the majority of score points. The minimum SE for all procedures occurred around the means with a sharp increase for scores above 575 and a more moderate increase for scores below 475 (English 9) and 400 (Math 9). Across the scale scores, the four scoring procedures were similar with UNW scoring resulting in marginally higher amounts of error. The PTRN scoring resulted in marginally lower amounts of error than the remaining score procedures. However, in Math 9, at a scale score of about 400, the UNW and WN/M standard errors crossed with the resulting UNW standard errors following closely with the WCRX2 and PTRN distributions and the WN/M standard errors markedly increasing. Taken together, these findings suggest that there is a less precise measurement of student ability at the higher and lower ends of the scale score distribution than at the center of the distributions for all four scoring procedures.

The pattern of percentile scale score differences was similar in both samples for both tests. Further, the magnitudes of the differences between pairs of scoring procedures were comparable between the two samples for the two tests. However, there was a lack of transitivity at the 10^{th} and 50^{th} percentiles for English 9. On Math 9, using a DTM of 0.50, all of the scoring procedures were similar at the 50^{th} percentile in Sample 1 and with the exception of WN/M vs. PTRN, the same results were found in Sample 2. The remaining differences exceeded 0.50 in absolute value.

The pattern scores at the 10th percentile point were less than the scores yielded by the other three scoring procedures but greater at the 90th percentile point on both tests. The RMS are consistent with the percentile results with the RMS involving PTRN scores markedly higher than the RMS values for the other three scoring procedures while the RMS values for UNW and WN/M were more comparable.

Proficiency levels results suggest that student scores did not fluctuate between the UNW and WN/M scoring procedures at the first two levels in both samples in English 9 and the UNW and WCRX2 scoring procedures for Math 9 in Sample 2. The WCRX2 and WN/M scoring procedures classified the same number of students at Level 3 and Level 4 for both English 9 and Math 9. The UNW, WCRX2, and WN/M procedures classified the same number of students at Level 3 and Level 4 in Sample 2 for Math 9. However, the number of students placed in the levels did vary, with up to 118 students being placed at a different level if the PTRN procedure was used in place of the other three scoring procedures in English 9 and up to 70 students being placed at a different level if the UNW procedure was used in place of PTRN at the first and second levels in Sample 2 of Math 9.

Lastly, when using a DTM of 0.50, the differences yielded by the four scoring procedures were significantly different for the vast majority of students for all pairs except the UNW and WN/M pair in English 9. For the latter pair, 72% were within one DTM, while for the other pairs of scores, the percentages varied between approximately 3 and 37%. On Math 9, the differences between yielded by the four scoring procedures were significantly different for the vast majority of students for all pairs. The percentage of students within one DTM ranged from 8% (WN/M and PTRN) to 30% (UNW and WCRX2).

Thus, perhaps with the exception of the UNW and WN/M scoring procedures in English 9, the scores yielded by the UWN, WCRX2, WN/M. and PTRN scoring procedures cannot be used interchangeably for English 9 or Math 9.

CHAPTER 5

Analyses and Results of High Stakes Examinations

The results and analyses conducted for the English 30 and Pure Mathematics 30 are reported in the present chapter. Like the previous chapter, this chapter is organized in the three sections for each examination.

English 30

Comparability of Samples

The classical test score statistics for the English 30 samples are reported in Table 27. The means and standard deviations between the two samples were similar. The differences between the means and standard deviations (sd) were less than one DTM for the SR and CR items and exam total for the two samples (0.49 and 0.28 for the means and sd of the SR items, 0.12 and 0.00 for the means and sd of the CR items, and 0.43 and 0.24 for the means and sd of the total exam). On average, the students earned about 68% of the maximum SR points possible (47.4 out of 70) and about 70% of the maximum CR points (21.1/30). These results combined with the negative skewness and kurtosis for both samples indicate that English 30 was a relatively easy exam. As for the low-stakes test, this finding suggests that there may be problems in obtaining an ability distribution for English 30 using IRT that is centered on zero with a standard deviation of one. Total scores and item scores were not available for the selection and constructed items. However, the internal consistency of the selection items for the total population from which the samples were drawn was 0.89 (Ping Yang, Personal Communication, October 18, 2007).

Table 27

Summary Classical Test Score Statistics: English 30

	Sample 1				Sample 2			
	Mean	SD	Skew	Kurtosis	Mean	SD	Skew	Kurtosis
SR	47.36	10.16	-0.37	-0.46	47.85	10.44	-0.42	-0.42
CR	21.14	4.53	-0.01	-0.35	21.26	4.53	-0.06	-0.31
Exam	68.68	13.50	-0.25	-0.48	69.11	13.74	-0.32	-0.37

Note: SR PTS = selected response points; CR PTS = constructed response points; EXAM = Total number of points (SR PTS + CR PTS).

Lastly, the correlations between the selection and constructed responses scores, 0.64 for Sample 1 and 0.63 for Sample 2, were moderate, suggesting that each item type was measuring, in part, something different (see Table 28). Taken together, these results indicate that the two samples were randomly equivalent and that non-overlapping information was yielded by the selection and constructed response items.

Assumptions of IRT

Unidimensionality. Principal component analysis yielded 22 components with eigenvalues greater than 1.0 for the SR items in English 30 for Sample 1. The eigenvalue for the first component, 8.22, was 5.67 times greater than the eigenvalue

Table 28

Correlations Classical Test Score Statistics: English 30

	Sample 1			Sample 2		
	SR Pts	CR Pts	Exam	SR Pts	CR Pts	Exam
SR Pts	1.00	0.64	0.97	1.00	0.63	0.97
CR Pts	0.64	1.00	0.81	0.63	1.00	0.81
Exam	0.97	0.81	1.00	0.97	0.81	1.00

Note: SR PTS = selected response points; CR PTS = constructed response points; EXAM = Total number of points (SR PTS + CR PTS).

of the second component 1.45. Further, the successive differences between remaining components were small (0.10, 0.06, 0.02, 0.04, 0.00, 0.02, 0.02, 0.02, 0.02, 0.01, 0.03, 0.00, 0.01, 0.00, 0.00, 0.00, 0.06, 0.02, 0.01, and 0.01).

Principal component analysis for the CR items on English 30 yielded one component with an eigenvalue greater than 1.0 for Sample 1. The eigenvalue for the first component 4.24, was 7.07 times greater than the eigenvalue of the second component 0.60. Further, the successive differences between remaining components were small (0.15, 0.20, 0.02, and 0.00). In Sample 2, the principal component analysis yielded 21 components with eigenvalues greater than 1.0 for the SR items in English 30. The eigenvalue for the first component, 8.79, was 6.23 times greater than the eigenvalue of the second component 1.41; the successive differences between remaining components were small (0.05, 0.04, 0.04, 0.05, 0.04, 0.02, 0.02, 0.02, 0.08, 0.10, 0.00, 0.01, 0.02, 0.00, 0.01, 0.02, 0.01, 0.02, 0.01, and 0.00).

Principal component analysis for the CR items in Sample 2 yielded one component with an eigenvalue greater than 1.0. The eigenvalue for the first component 4.23, was 6.82 times greater than the eigenvalue of the second component 0.62; the successive differences between remaining components were small (0.17, 0.19, 0.03, 0.02).

The scree plots confirm the dominance of the first principal component for the SR items and CR items in both Sample 1 and Sample 2 (see Appendix C9 to Appendix C12).

Non-linear factor analysis. Non-linear factor analysis (NOHARM, Fraser, 1988) was also used to determine the factor structure of the selection items. The fit indices for the English 30 Sample 1 and Sample 2 are presented in Table 29. For both samples, the unidimensional model fit the data well: the changes in the fit statistics were marginal when the number of factors was increased from 1 to 2. For example, Tanaka values went up by 0.001 in Sample 1 and 0.002 in Sample 2, and RMSR values went down 0.001 in Sample 1 and remained the same in Sample 2.

The results of the principal component analysis, the scree plots, and NOHARM suggested that there was a dominant component underlying the student responses to the CR and SR items on the English 30 examination. Consequently, the assumption of essential dimensionality was met for both sets of items.

Local independence. Given that the assumption of essential unidimensionality was met for both the SR and CR items, the assumption of essential item independence was obtained for both the selected response and constructed items in both samples.

Speededness. The percentage of students who did not complete the last three items was calculated. Less than 1% of the students did not complete the last three questions. Thus, it was concluded that speededness was not a factor.

Taken together, the results of testing the assumptions revealed the assumptions were met for the use of IRT were met for English 30 examination.

Table 29

NOHARM Fit Indices for English 30

	Sample 1		Sample 2	
No. of Factors	Tanaka	RMSR	Tanaka	RMSR
1	0.981	0.005	0.982	0.004
2	0.982	0.004	0.984	0.004

Fit among Weighted, Unweighted, and Pattern Scores at the Group Level

The means and standard deviations for UNW, WCRX2, WN/M, and PTRN scoring procedures are provided in Table 30 and the correlations are provided in Table 31.

Inspection of the means in Table 30 reveals that the four scale means for both samples were all less than 500. As foreshadowed above, the mean ability estimates on the IRT theta scale were again all less than zero because of the easiness of the test. Consequently, when these scores were transformed, the means were less than 500.

Further inspection of the four means in Table 30 reveals that the scoring procedures can be placed in two sets for both samples: UNW with PTRN and WCRX2 with WN/M. The members in each pair differed by less than one DTM and the difference between the two members in one pair and the two members in the second pair differed by more than one DTM (e.g., 2.62 for UNW vs. WN/M in scale Sample 1). The standard deviations of the four distributions all exceeded 50 for both

Table 30

Measures of Central Tendency for the Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English30

	Sample 1				Sample 2			
	Mean	SD	Skew	Kurtosis	Mean	SD	Skew	Kurtosis
UNW	483.02	56.96	0.20	0.17	491.11	57.68	0.15	0.29
WCRX2	480.82	55.56	0.08	0.51	489.08	56.75	0.06	0.31
WN/M	480.40	55.21	0.02	0.01	488.75	56.69	-0.02	0.13
PTRN	483.40	55.59	0.00	-0.02	491.56	56.45	-0.04	0.08

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two; WN/M – Weighted SR/CR.

samples. Further, for both samples the standard deviation of the distribution of UNW scores exceeded the standard deviations of the distributions of the remaining three scale scores by more than one DTM; the differences among the standard deviations of the remaining scale scores were within one DTM. With the exception of WN/M and PTRN in Sample 2, the scoring distributions were slightly positively skewed. The kurtosis suggests a somewhat leptokurtic distribution with the exception of PTRN in Sample 1 which is slightly platykurtic.

Lastly, the correlations (see Table 31) among the four sets of scores were all above 0.98 for both samples; the differences among the six pairs of correlations were all less than one DTM. Taken together, the results first reveal that the scoring procedures at the group level tended to rank the students the same, but differed in their central tendency and variability with the UNW and PTN scoring procedures and the WCRX2 and WN/M scoring procedures yielding comparable means.

Standard error. As shown in Figure 17 (Sample 1) and Figure 18 (Sample 2), the overall magnitudes of the standard errors of measurement were low to moderate

Table 31

Correlations of the Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English30

	Sample 1				Sample 2			
	UNW	WCRX2	WN/M	PTRN	UNW	WCRX2	WN/M	PTRN
UNW	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.98
WCRX2	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.98
WN/M	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.98
PTRN	0.98	0.98	0.98	1.00	0.98	0.98	0.98	1.00

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two; WN/M – Weighted SR/CR.

for the majority of scores (ranging from five to 10 transformed score points). The minimum standard error for the four scoring procedures occurred around the means (between 450 and 550) and, unexpectedly between 320 and 350. Taking into account this latter exception, the standard errors then increased, but much more rapidly for scores above 550 than for scores below 320. For scale scores less than about 430, the unweighted scoring procedure resulted in the highest amount of error; the differences among the three remaining score procedures were more similar and smaller.

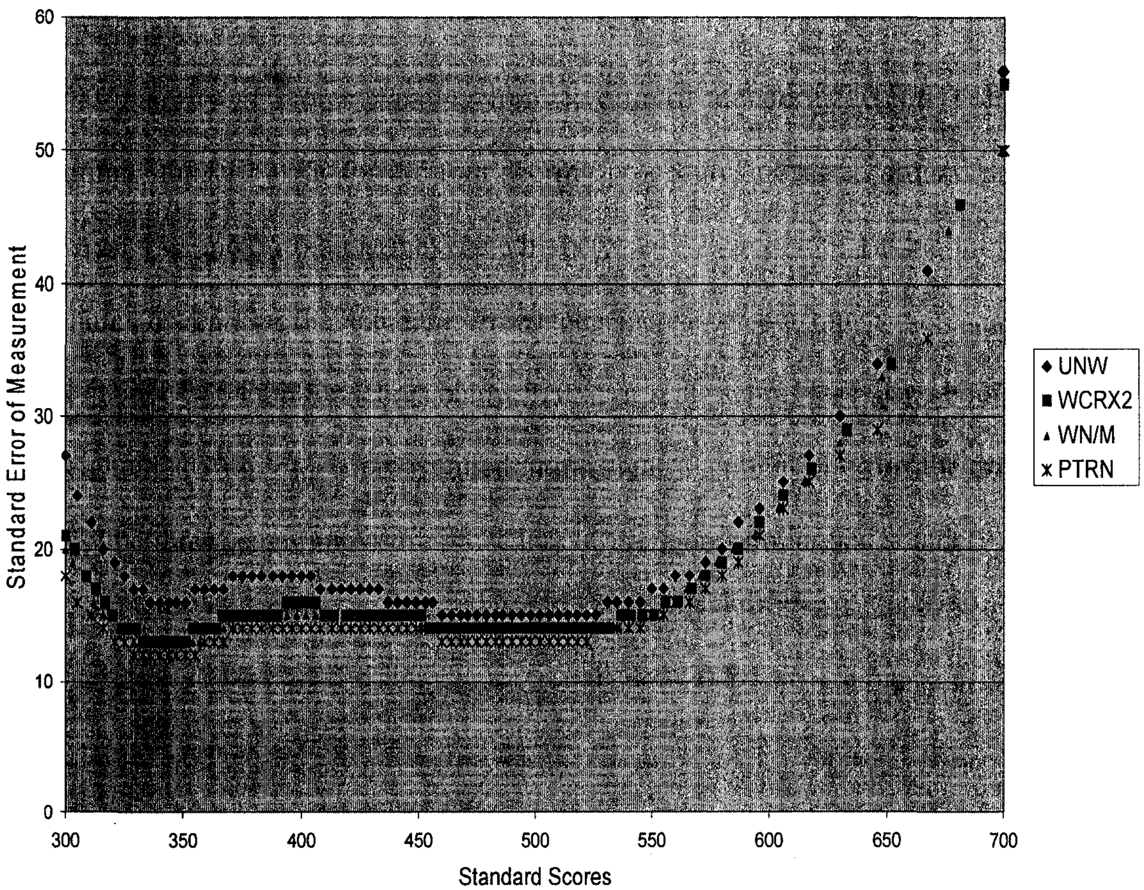

Figure 17. Standard Error of Measurement for English 30 Sample 1

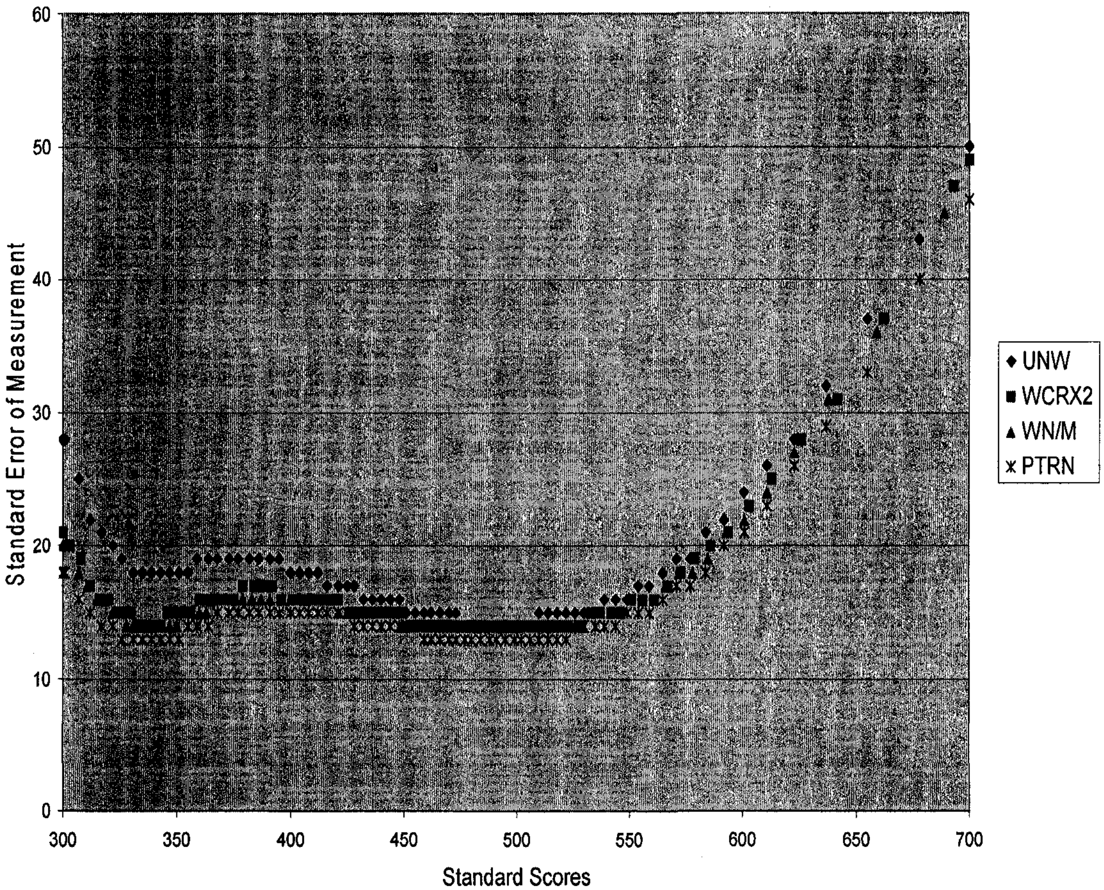

Figure 18. Standard Error of Measurement for English 30 Sample 2

Across the scale scores, the four scoring procedures were similar with UNW scoring resulting in marginally higher amounts of error and PTRN scoring resulting in marginally lower amounts of error than the remaining score procedures, particularly at the low and high end of the scale score distribution. Taken together, these findings suggest that there is a less precise measurement of student ability at the higher end of the scale score than at the lower end for all four scoring procedures.

Scale score differences at the 10^{th}, 50^{th} and 90^{th} percentiles. The results for the differences among the scores yielded by the four scoring procedures at the student level, the 10th, 50th, and 90th percentile score points are reported in Table 32 for both Sample 1and Sample 2.

The magnitude and pattern of percentile scale score differences was similar in both samples. Using a DTM of 0.50, the WCRX2 and WN/M scoring procedures were similar at the 10th and 50th percentile points but not at the 90th percentile point in both samples, while the UNW and PTRN scoring procedures were similar at the 50th percentile point in both samples. The remaining differences exceeded 0.50 in absolute value. In the case of the 10th percentile, the remaining three differences – UNW vs. PTRN, WCRX2 vs. PTRN, and WN.M vs. PTRN – were negative, ranging from -2.00 to -18.00; in contrast these same three differences were positive at the 90th percentile point, ranging from 9.00 to 14.00. That is, the pattern scores at the 10th percentile point were less than the scores yielded by the other three scoring procedures but greater at the 90th percentile point. Lastly, the two weighted procedures yielded scores greater than the scores yielded by the UNW procedure at the 90th percentile point, but not to as great an extent as observed when pattern scoring was considered (e.g., 4.00 vs. 14.00, 12.00, and 12.00 in Sample 1).

Root mean square. Table 33 shows that the WCRX2 and WN/M scoring procedures produced very low RMS values, 0.78 for Sample 1 and 0.67 for Sample 2, indicating close agreement between the two sets of scores. This is consistent with the percentile findings presented above, and may be attributable to the small difference in weights (2.0 vs. 2.3). The agreement between UNW and WCRX2 and WN/M scoring procedures was less: the RMS values were, respectively, 3.03 and 3.69 for

Table 32

Scale Score Differences at the 10^{th}, 50^{th}, and 90^{th} Percentiles: English 30

	Sample 1				Sample 2		
	10^{th} Percentile	50^{th} Percentile	90^{th} Percentile		10^{th} Percentile	50^{th} Percentile	90^{th} Percentile
UNW vs. WCRX2	1.00	2.00	4.00	UNW vs. WCRX2	1.00	2.00	4.00
UNW vs. WN/M	1.00	2.00	5.00	UNW vs. WN/M	1.00	2.00	4.00
WCRX2 vs. WN/M	0.00	0.00	1.00	WCRX2 vs. WN/M	0.00	0.00	1.00
UNW vs. PTRN	-15.00	0.00	14.00	UNW vs. PTRN	-13.00	0.00	12.00
WCRX2 vs. PTRN	-18.00	-2.00	12.00	WCRX2 vs. PTRN	-16.00	-2.00	10.00
WN/M vs. PTRN	-18.00	-3.00	12.00	WN/M vs. PTRN	-16.00	-2.00	9.00

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two; WN/M – Weighted SR/CR.

Table 33

Root Mean Squares of the Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English30

	Sample 1	Sample 2
UNW vs. WCRX2	3.03	2.85
UNW vs. WN/M	3.69	3.38
WCRX2 vs. WN/M	0.78	0.67
UNW vs. PTRN	11.89	10.59
WCRX2 vs. PTRN	12.09	10.92
WN/M vs. PTRN	12.25	11.10

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two; WN/M – Weighted SR/CR.

Sample 1 and 2.85 and 3.38 for Sample 2. Lastly, the RMS values for the PTRN scoring procedure versus the UNW, WCRX2, and WN/M scoring procedures were much larger, ranging from 11.89 to 12.25 for Sample 1 and 10.59 to 12.25 for Sample 2. The lack of agreement between the PTRN scoring procedure and the other scoring procedures corresponds with the findings noted above at the 10^{th} and 90^{th} percentiles.

Proficiency levels. The observed score to UNW scale score conversion table indicated that the corresponding cut scores would be 408, 486, and 550 for Sample 1 and 416, 491, and 554 for Sample 2 in the UNW scale score distribution. These same cut points were used in each of the other three score distributions. The number and percentage of students in each of the four performance levels for each of the four scoring procedures are shown in Table 34 for both Sample 1 and Sample 2.

As expected, the WCRX2 and WN/M scoring procedures classified the same number of students at each level. The UNW scoring procedure resulted in 29 fewer students placed in the first level, 22 fewer in the second level, 14 more at the third level, and 37 more in the fourth level than the two weighted scoring procedures in Sample 1; the comparable numbers in Sample 2 are slightly different: 26, 37, 33, and 30. A similar pattern was found for the PTRN scoring procedure: 44 fewer at Level 1, six more at Level 2, 21 more at Level 3, and 17 more at Level 4 in Sample 1; and 36 fewer at Level 1, 10 fewer at Level 2, 37 fewer at Level 3 and 9 fewer at Level 4. While the differences in the corresponding percentages are small (all less than two percent), the number of students placed in the levels did vary, with up to 100 students being placed at a different level if the WCRX2 or WN/M procedures were used in place of the pattern, and up to 66 students if the UNW procedure was used in place of the PTRN procedure.

Difference in individual student scaled scores. A graphical representation of the distribution of differences between scaled scores yielded by the UNW, WCRX2, WN/M, and PTRN scoring procedures is provided in Figure 19 for Sample 1 and Figure 20 for Sample 2 (see Appendices E5 and E6 for the corresponding tables). As shown, the greatest differences occurred when the pattern scores were involved. In these cases, the differences ranged from -46 to 74 in Sample 1 and -49 to 78 in Sample 2 for the UNW procedure, -44 to 72 in Sample 1 and -48 to 76 in Sample 2 for the WCRX2 procedure, and -46 to 72 in Sample 1 and -49 to 76 in Sample 2 for the WN/M procedure. Individual student scores did vary somewhat between the UNW and WN/M and the WCRX2 and WN/M procedures, which ranged from -13 to 23. The differences between the WCRX2 and WN/M scoring procedures were

Table 34

Proficiency Levels of the Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English30

	Sample 1									Sample 2							
	UNW		WCRX2		WN/M		PTRN			UNW		WCRX2		WN/M		PTRN	
	#	%	#	%	#	%	#	%		#	%	#	%	#	%	#	%
Level 1									Level 1								
(300 – 407)	189	9.45	218	10.9	218	10.9	174	8.7	(300 – 415)	181	9.1	207	10.4	207	10.4	171	8.6
Level 2									Level 2								
(408 – 485)	828	41.4	850	42.5	850	42.5	856	42.8	(416 – 490)	798	39.9	835	41.8	835	41.8	825	41.3
Level 3									Level 3								
(486 – 549)	733	36.7	719	36.0	719	36.0	740	37.0	(491 – 553)	746	37.3	713	35.7	713	35.7	750	37.5
Level 4									Level 4								
(550 – 700)	250	12.5	213	10.7	213	10.7	230	11.5	(554 – 700)	275	13.8	245	12.3	245	12.3	254	12.7

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two; WN/M = Weighted SR/CR; # = number of students; % = percentage of Students

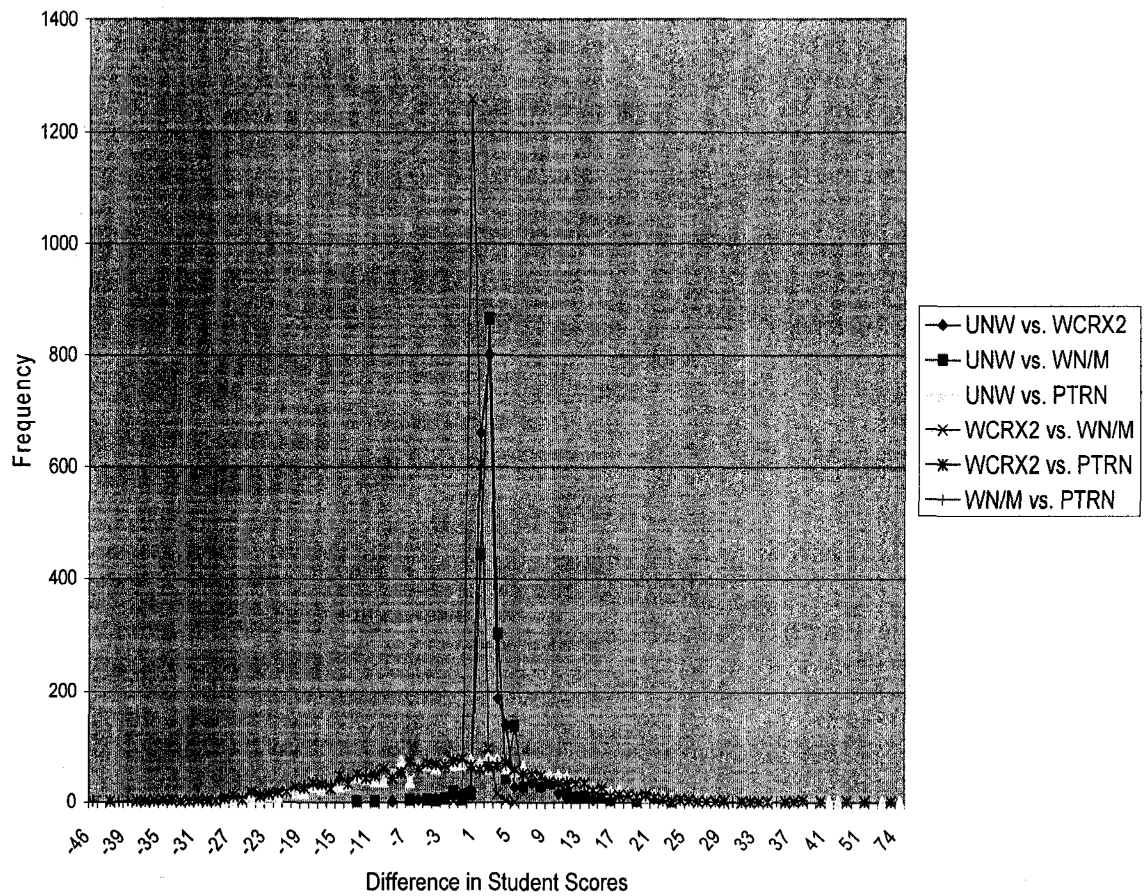

Figure 19. Differences Between Unweighted, Weighted CRx2, Weighted N/M and Pattern Scores: Sample 1

smaller, varying from -2.00 to 5.00. This latter finding again may be attributable to the small difference in weights (2.0 vs. 2.3). When using a DTM of 0.50, the differences between the two pairs of scores yielded by the four scoring procedures were significantly different for the vast majority of students for all pairs of scoring procedures except the WCRX2 and WN/M pair. For example, in Sample 1 the scores yielded by the UNW and WN/M scoring procedures and the UNW and PTRN

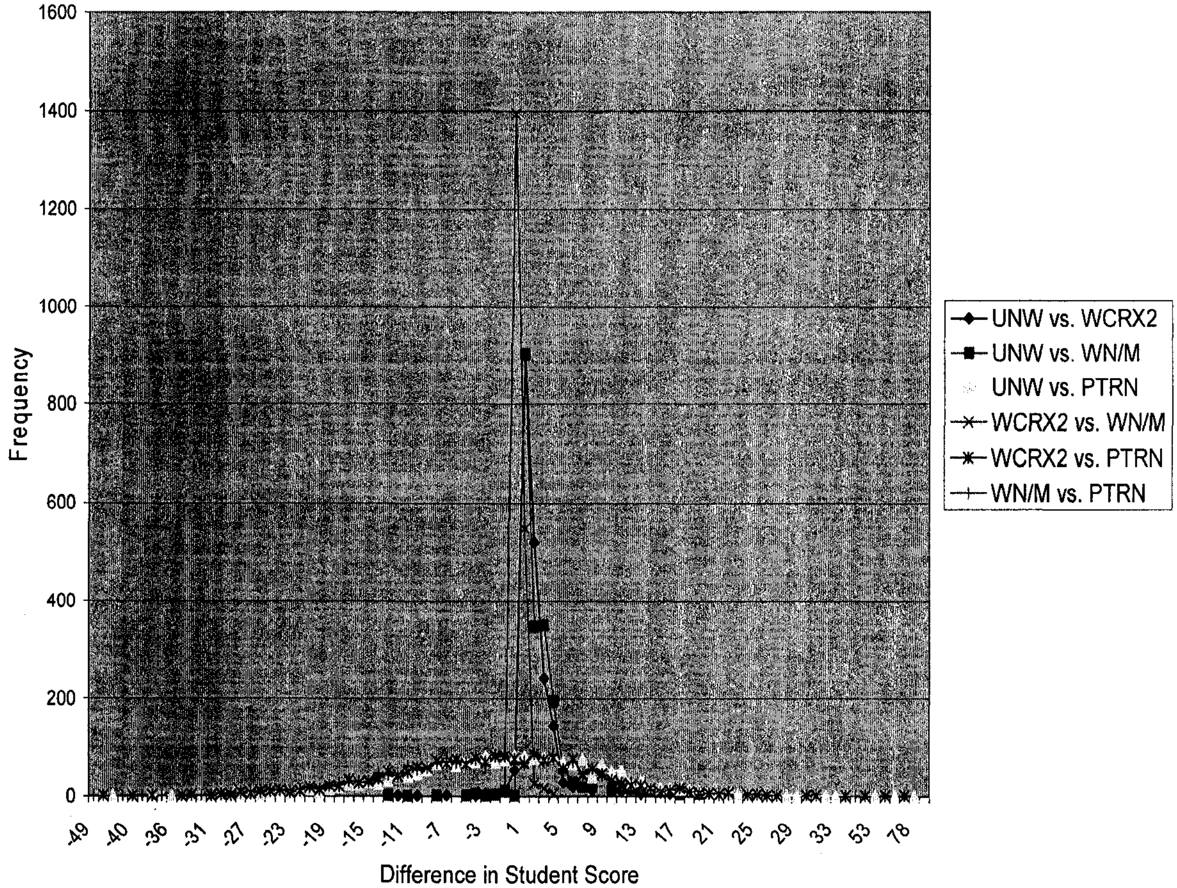

Figure 20. Differences Between Unweighted, Weighted CRx2, Weighted N/M and Pattern Scores: Sample 2

scoring procedure were within one DTM for 15 (1%) and 77 (3.9%) of students respectively. The corresponding numbers for each of the weighted procedures and the PTRN procedure were 66 (3.3%) and 58 (3.0%). In contrast, the WCRX2 and WN/M scores for 1,258 (62.9%) students were within one DTM. If the DTM is relaxed to 1.00, the corresponding numbers, taken in the same order, are 699 (35.0%), 204 (10.2%), 187 (9.4%), and 1881 (94.0%). The gain for UNW and PTRN procedures is attributable to students receiving one more score point using the UNW scoring procedure than using pattern scoring procedure. The results are similar in Sample 2.

Pure Mathematics 30

Comparability of Samples

Both the means and standard deviations between the two samples were similar (see Table 35). The differences between the means and standard deviations (sd) were less than one DTM for the SR and CR items and exam total for the two samples (0.05 and 0.02 for the means and sd of the SR items, 0.04 and 0.02 for the means and sd of the CR (NR) items, 0.00 for both the means and sd of the CR (OE) items, and 0.01 and 0.00 for the means and sd of the total exam). On average, the students earned about 70% of the maximum SR points possible (23 out of 33) and about 67% of the maximum CR points (14.1/21). These results combined with the negative skewness and kurtosis for both samples indicate that Pure Math 30 was a relatively easy exam. The internal consistency of the selection items for the total population from which the samples were drawn was 0.85 (Ping Yang, Personal Communication, October 18,

Table 35

Summary Classical Test Score Statistics: Pure Math 30

	Sample 1				Sample 2			
	Mean	SD	Skew	Kurtosis	Mean	SD	Skew	Kurtosis
SR	22.90	5.64	-0.38	-0.40	22.95	5.62	-0.35	-0.43
CR (NR)	4.25	1.44	-0.64	-0.23	4.21	1.46	-0.60	-0.32
CR (OE)	9.84	3.03	-0.24	-0.61	9.84	3.03	-0.20	-0.63
Exam	36.99	9.23	-0.33	-0.46	37.00	9.23	-0.30	-0.43

Note: SR = selected response; CR (NR) = numerical response constructed response; CR (OE) = open ended constructed response; EXAM = Total number of points (SR PTS + CR PTS).

2007). Lastly, the correlations between the selection and constructed responses scores, 0.66 for Sample 1 and 0.65 for Sample 2, were moderate, suggesting that each item type was measuring, in part, something different (see Table 36). Like English 30, the two samples were randomly equivalent and non-overlapping information was yielded by the selection and constructed response items.

Assumptions of IRT

Unidimensionality. The assumption of unidimensionality was assessed separately for the subset of selection items and subset of constructed response items given each was analyzed separately using item response theory. Principal component analysis yielded 10 components with eigenvalues greater than 1.0 for the SR items in Pure Math 30 for Sample 1. The eigenvalue for the first component, 4.96, was 4.00 times greater than the eigenvalue of the second component 1.24. Further, the successive differences between remaining components were small (0.09, 0.05, 0.04, 0.00, 0.02, and 0.02).

Table 36

Correlations Classical Test Score Statistics: Pure Math 30

	SR Pts	CR Pts (NR)	CR Pts (OE)	Exam	SR Pts	CR Pts (NR)	CR Pts (OE)	Exam
SR Pts	1.00	066	0.76	0.96	1.00	0.65	0.76	0.96
CR Pts (NR)	0.66	1.00	0.64	0.77	0.65	1.00	0.65	0.77
CR Pts (OE)	0.76	0.64	1.00	0.89	0.76	0.65	1.00	0.89
Exam	0.96	0.77	0.89	1.00	0.96	0.77	0.89	1.00

Note: SR = selected response; CR (NR) = numerical response constructed response; CR (OE) = open ended constructed response; EXAM = Total number of points (SR PTS + CR PTS).

Principal component analysis for the CR numerical response (NR) items in Pure Math 30 yielded one component with an eigenvalue greater than 1.0 for Sample 1. The eigenvalue for the first component, 1.76, was 1.86 times greater than the eigenvalue of the second component 0.95. Further, the successive differences between remaining components were small (0.06, 0.07, 0.02, and 0.02). For the CR open ended (OE) items, principal component analysis yielded one component with an eigenvalue greater than 1.0 for Sample 1. The eigenvalue for the first component, 2.00, was 3.63 times greater than the eigenvalue of the second component 0.55. Further, the successive differences between the remaining component were small (0.10).

In Sample 2, the principal component analysis yielded seven components with eigenvalues greater than 1.0 for the SR items in Pure Math 30. The eigenvalue for the first component, 4.93, was 4.04 times greater than the eigenvalue of the second component 1.22; the successive differences between the remaining components were small (0.07, 0.02, 0.06, 0.01, 0.04, and 0.02).

Principal component analysis for the CR (NR) items in Sample 2 yielded one component with an eigenvalue greater than 1.0. The eigenvalue for the first component 1.79, was 1.84 times greater than the eigenvalue of the second component 0.97. Further, the successive differences between remaining components were small (0.10, 0.03, 0.05, and 0.04). For the CR (OE) items, principal component yielded one component with an eigenvalue greater than 1.0. The eigenvalue for the first component 2.03, was 3.80 times greater than the eigenvalue of the second component

0.53. Further, the successive difference between the remaining component was small (0.10).

The scree plots confirm the dominance of the first principal component for the SR items, CR (NR) items, and CR (OE) items in both Sample1 and Sample 2 (see Appendix C13 to Appendix C18).

Non-linear factor analysis. The fit indices for the Pure Math 30 selection items for both samples are presented in Table 37. For both samples, the unidimensional model fit the data well: the changes in the fit statistics were marginal when the number of factors was increased from 1 to 2. The Tanaka values went up by 0.003 in Sample 1 and 0.002 in Sample 2, and RMSR values went down 0.001 in Sample 1 and Sample 2.

The results of the principal component analysis, the scree plots, and NOHARM suggested that there was a dominant component underlying the student responses to the CR and SR items on the Pure Math 30 examination. Consequently, the assumption of essential dimensionality was met for both sets of items.

Local independence. Given that the assumption of essential unidimensionality was met for both the SR and CR items, the assumption of essential

Table 37
NOHARM Fit Indices for Pure Math 30

	Sample 1		Sample 2	
No. of Factors	Tanaka	RMSR	Tanaka	RMSR
1	0.988	0.005	0.989	0.005
2	0.991	0.004	0.991	0.004

item independence was obtained for both the selected response and constructed items in both samples.

Speededness. The percentage of students who did not complete the last three items was calculated. Less than 1% of the students did not complete the last three questions. Thus, it was concluded that speededness was not a factor.

Taken together, the three sets of results presented above indicate that the assumptions for the use of IRT were met for Pure Math 30.

Fit among Weighted, Unweighted, and Pattern Scores at the Group Level

Inspection of the means in Table 39 reveals that the four scale means for both samples were all less than 500, a finding that is attributable to the easiness of the items. As shown in Table 38 the mean for UNW was less than each of the other means, yet no significant differences are claimed among the other three scoring procedures due to the lack of transitivity. The same held true for Sample 2. The four score distributions were positively skewed and leptokurtic. The standard deviations of the four distributions all exceeded 50 for both samples. While the standard deviation for WCRX2 was smaller in Sample 1 and larger in Sample 2 than each of the other means, no significant differences are claimed among the other three scoring procedures due to the lack of transitivity. Lastly, the correlations (see Table 39) among the four sets of scores were all 1.00 for both samples.

Table 38

Measures of Central Tendency for the Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Pure Math 30

	Sample 1				Sample 2			
	Mean	SD	Skew	Kurtosis	Mean	SD	Skew	Kurtosis
UNW	495.37	59.66	0.43	1.15	497.08	59.28	0.29	1.00
WCRX2	496.62	60.47	0.35	1.28	498.51	57.56	0.36	0.93
WN/M	496.34	58.97	0.44	1.23	497.98	58.13	0.33	0.96
PTRN	495.92	59.45	0.40	1.27	497.56	58.76	0.34	0.89

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two; WN/M – Weighted SR/CR.

Table 39

Correlations of the Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Pure Math 30

	Sample 1				Sample 2			
	UNW	WCRX2	WN/M	PTRN	UNW	WCRX2	WN/M	PTRN
UNW	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
WCRX2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
WN/M	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PTRN	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two; WN/M – Weighted SR/CR.

Standard error. The plots of conditional standard errors of measurement for each scoring procedure are shown in Figure 21 and Figure 22 for Samples 1 and 2, respectively. The distributions are parabolic in shape, with low to moderate standard errors around the means (between 400 and 550), and increasing values on either side of this interval. The magnitudes of the standard errors of measurement varied from

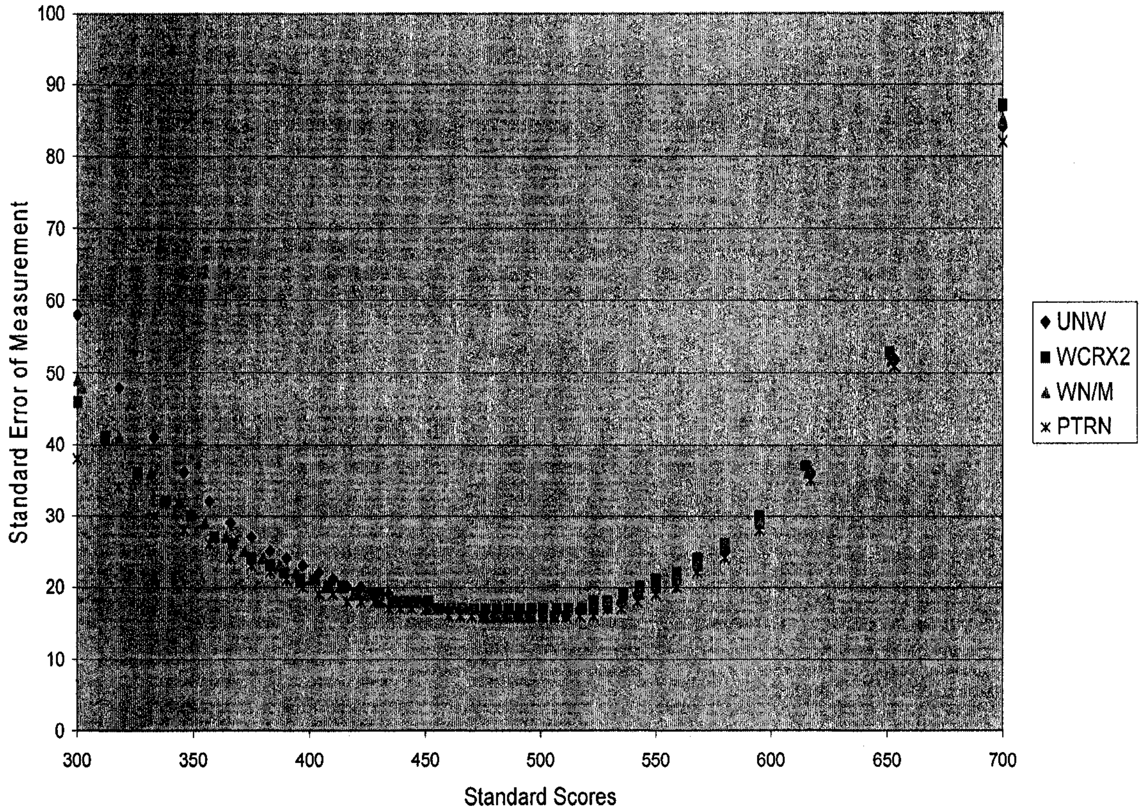

Figure 21. Standard Error of Measurement for Pure Math 30 Sample 1

five scale points (around the mean) to 20 scale points (around the low end of the distribution) for the majority of score points. The standard errors then increased more rapidly for scores above 550 than the standard errors for scores below 400. At a scale score of 300 the standard errors are widely spread especially between UNW and PTRN with a difference of about 20 SE points. At a scale score of 700 the standard errors are closely grouped with a range of about 5 SE points. The UNW scoring procedure resulted in the highest amount of error than the three remaining score procedures until standard score of 450 where UNW crossed into the three remaining scoring procedures. As such, all four scoring procedures had similar SEs from

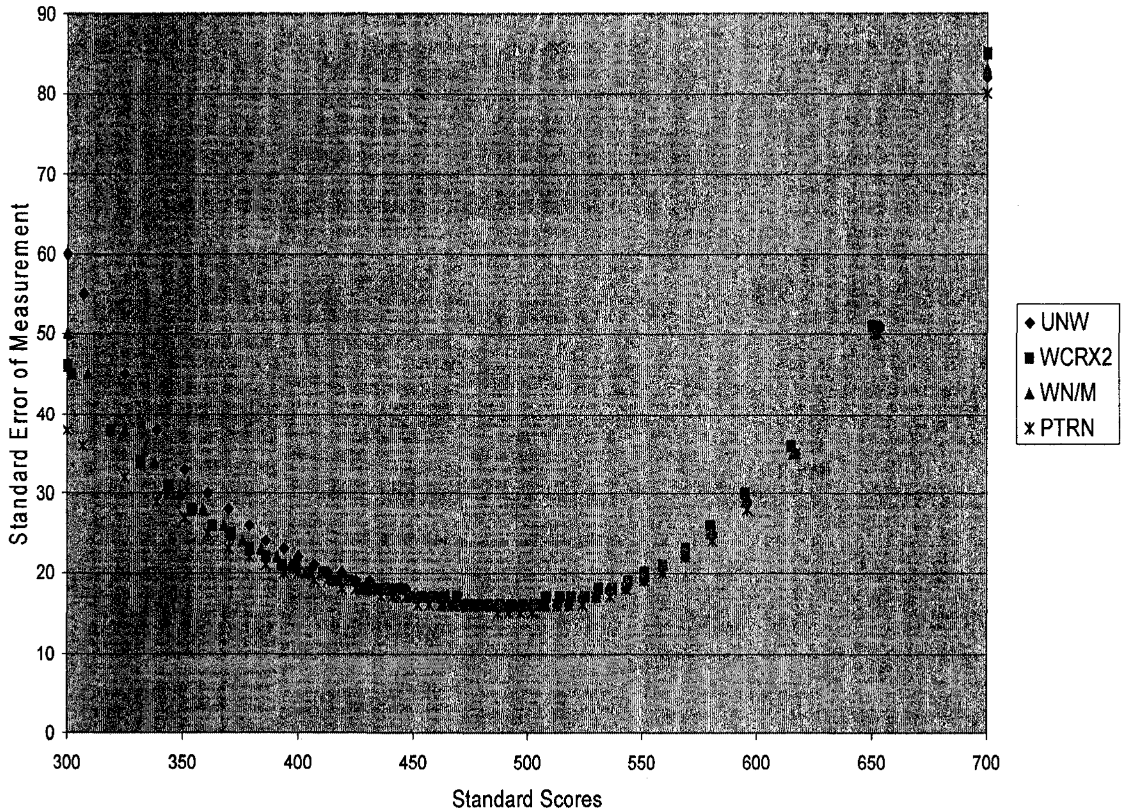

Figure 22. Standard Error of Measurement for Pure Math 30 Sample 2

around 425 to 525. Across the scale scores, PTRN scoring resulted in marginally lower amounts of error than the remaining score procedures, particularly at the low end of the scale score distribution. Taken together, these findings suggest that there is less precise measurement of student ability at the higher end and the lower end of the scale score for all four scoring procedures.

Scale score differences at the 10th, 50th and 90th percentiles. The pattern of percentile scale score differences was similar in both samples (see Table 40). Further, the magnitudes of the differences between pairs of scoring procedures were comparable between the two samples. The WCRX2 and WN/M scoring procedures were similar at the 10th and 50th percentile points but not at the 90th percentile point in both samples, while the UNW and PTRN scoring procedures were similar at the 50th percentile point in both samples. The WCRX2 and PTRN scoring procedures were similar at the 50th percentile in Sample 1 while the WN/M and PTRN procedures were similar at the 50th percentile in both samples. However, there was a lack of transitivity at the 10th, 50th, and 90th percentiles. Hence, no significant differences among the scores are claimed. In the case of the 10th percentile, the differences were negative, ranging from -2.00 to -7.00; in contrast these same differences were positive at the 90th percentile point, ranging from 1.00 to 7.00. That is, the pattern scores at the 10th percentile point were less than the scores yielded by the other three scoring procedures but greater at the 90th percentile point and the UNW scores at the 10th percentile were less than the WCRX2 and WN/M scores.

Root mean square. As shown in Table 41 the UNW and WN/M scoring procedures produced low RMS values, 1.76 for Sample 1 and 1.78 for Sample 2, indicating closer agreement between the two sets of scores. The agreement between UNW and WCRX2 and WN/M scoring procedures were less: the RMS values were, respectively, 2.49 for both in Sample 1 and 2.66 for both in Sample 2. Lastly, the RMS values for the PTRN scoring procedure versus the UNW, WCRX2, and WN/M scoring procedures were larger, ranging from 5.93 to 6.20 for Sample 1 and 5.58 to 5.73 for Sample 2. The lack of agreement between the PTRN scoring procedure and the other scoring procedures corresponds with the findings presented above at the 10th and 90th percentiles.

Table 40

Scale Score Differences at the 10th, 50th, and 90th Percentiles: Pure Math 30

	Sample 1				Sample 2		
	10^{th} Percentile	50^{th} Percentile	90^{th} Percentile		10^{th} Percentile	50^{th} Percentile	90^{th} Percentile
UNW vs. WCRX2	-2.00	-1.00	0.00	UNW vs. WCRX2	-3.00	-1.00	0.00
UNW vs. WN/M	-2.00	-1.00	0.00	UNW vs. WN/M	-2.00	-1.00	0.00
WCRX2 vs. WN/M	0.00	0.00	1.00	WCRX2 vs. WN/M	0.00	0.00	1.00
UNW vs. PTRN	-7.00	0.00	6.00	UNW vs. PTRN	-7.00	0.00	6.00
WCRX2 vs. PTRN	-6.00	0.00	7.00	WCRX2 vs. PTRN	-5.00	1.00	7.00
WN/M vs. PTRN	-6.00	0.00	7.00	WN/M vs. PTRN	-6.00	0.00	6.00

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two; WN/M – Weighted SR/CR

Table 41

Root Mean Squares of the Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Pure Math 30

	Sample 1	Sample 2
UNW vs. WCRX2	2.49	2.66
UNW vs. WN/M	1.76	1.78
WCRX2 vs. WN/M	2.49	2.66
UNW vs. PTRN	6.20	5.70
WCRX2 vs. PTRN	5.93	5.73
WN/M vs. PTRN	5.97	5.58

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two; WN/M – Weighted SR/CR.

Proficiency levels. The observed score to UNW scale score conversion table indicated that the corresponding cut scores were 493, 495, and 542 for Sample 1 and 441, 497, and 543 for Sample 2 in the UNW scale score distribution. These same cut points were used in each of the other three score distributions. The number and percentage of students in each of the four performance levels for each of the four scoring procedures are shown in Table 42 for both Sample 1 and Sample 2.

The UNW, WCRX2 and WN/M scoring procedures classified the same number of students at each level. This was expected as previous results have suggested that the UNW, WCRX2 and WN/M scoring procedures yielded scores that were similarly distributed. The PTRN scoring procedure resulted in 17 fewer students placed in the first level, 1 fewer in the second level, 8 more at the third level, and 24 fewer in the fourth level than the two weighted scoring procedures in Sample1; the comparable numbers in Sample 2 are slightly different: 15, 3, and 1 greater, and 19 fewer in

Table 42

Proficiency Levels of the Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Pure Math 30

	Sample 1									Sample 2							
	UNW		WCRX2		WN/M		PTRN			UNW		WCRX2		WN/M		PTRN	
	#	%	#	%	#	%	#	%		#	%	#	%	#	%	#	%
Level 1 (300 – 438)	283	14.2	283	14.2	283	14.2	300	15.0	Level 1 (300 – 440)	274	13.7	274	13.7	274	13.7	289	14.5
Level 2 (439 – 494)	700	35.0	700	35.0	700	35.0	699	35.0	Level 2 (441 – 496)	742	37.1	742	37.1	742	37.1	745	37.3
Level 3 (495 – 541)	611	65.6	611	65.6	611	65.6	619	31.0	Level 3 (497 – 542)	584	29.2	584	29.2	584	29.2	585	29.3
Level 4 (542 – 700)	406	85.9	406	85.9	406	85.9	382	19.1	Level 4 (543 – 700)	400	20.0	400	20.0	400	20.0	381	19.1

Note: PTRN = Pattern; UNW = Unweighted; WCRX2 = Weighted CR factor of two; WN/M = Weighted SR/CR; # = number of students; % = percentage of Students.

Levels 1 through 4 respectively. In the case of the Pure Math 30, only if the PTRN scoring procedure was used, did the student placing in levels result in changes.

Difference in individual student scaled scores. A graphical representation of the distribution of differences between scaled scores yielded by the UNW, WCRX2, WN/M, and PTRN scoring procedures is provided in Figure 23 for Sample 1 and Figure 24 for Sample 2 (see Appendix E for the corresponding tables). As shown, the greatest differences occurred when the PTRN scoring procedure was involved. In these cases, the differences ranged from -53 to 38 in Sample 1 and -46 to 51 in

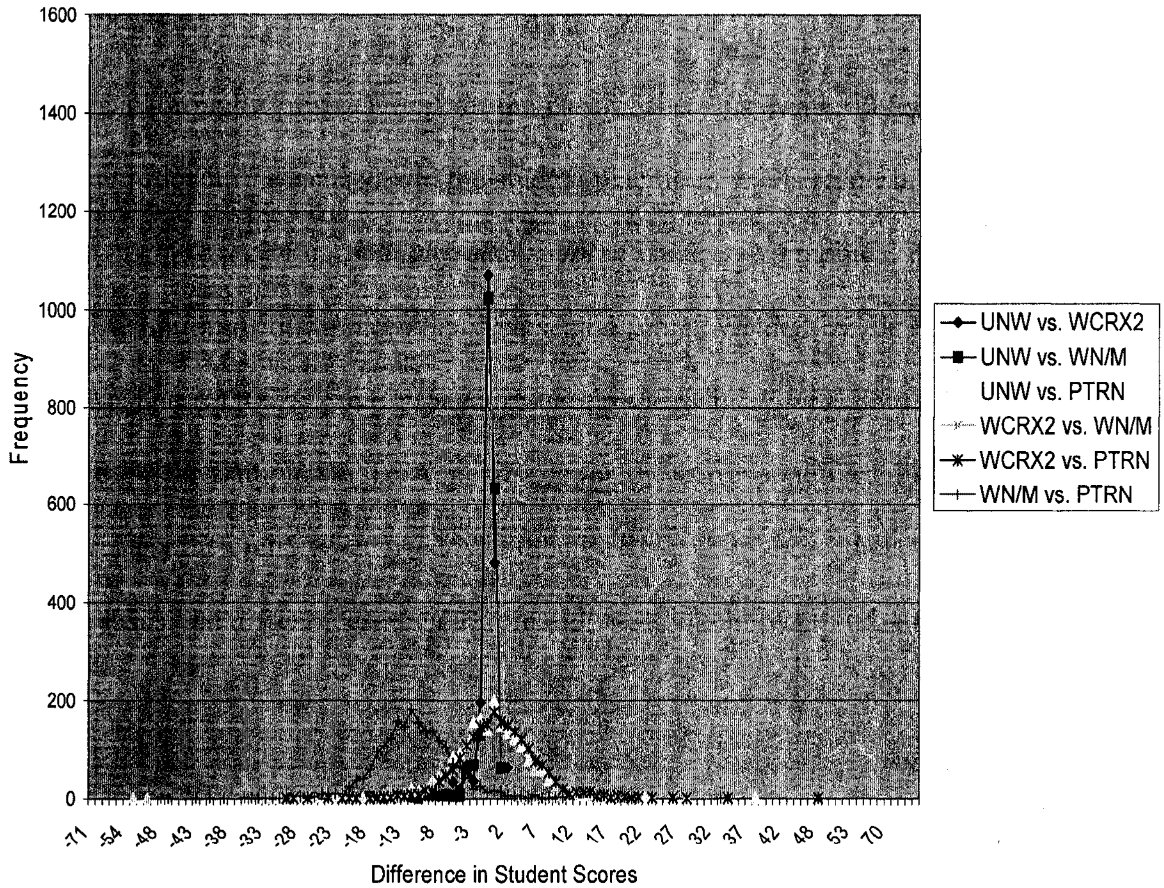

Figure 23. Differences Between Unweighted, Weighted CRx2, Weighted N/M and Pattern Scores: Sample 1

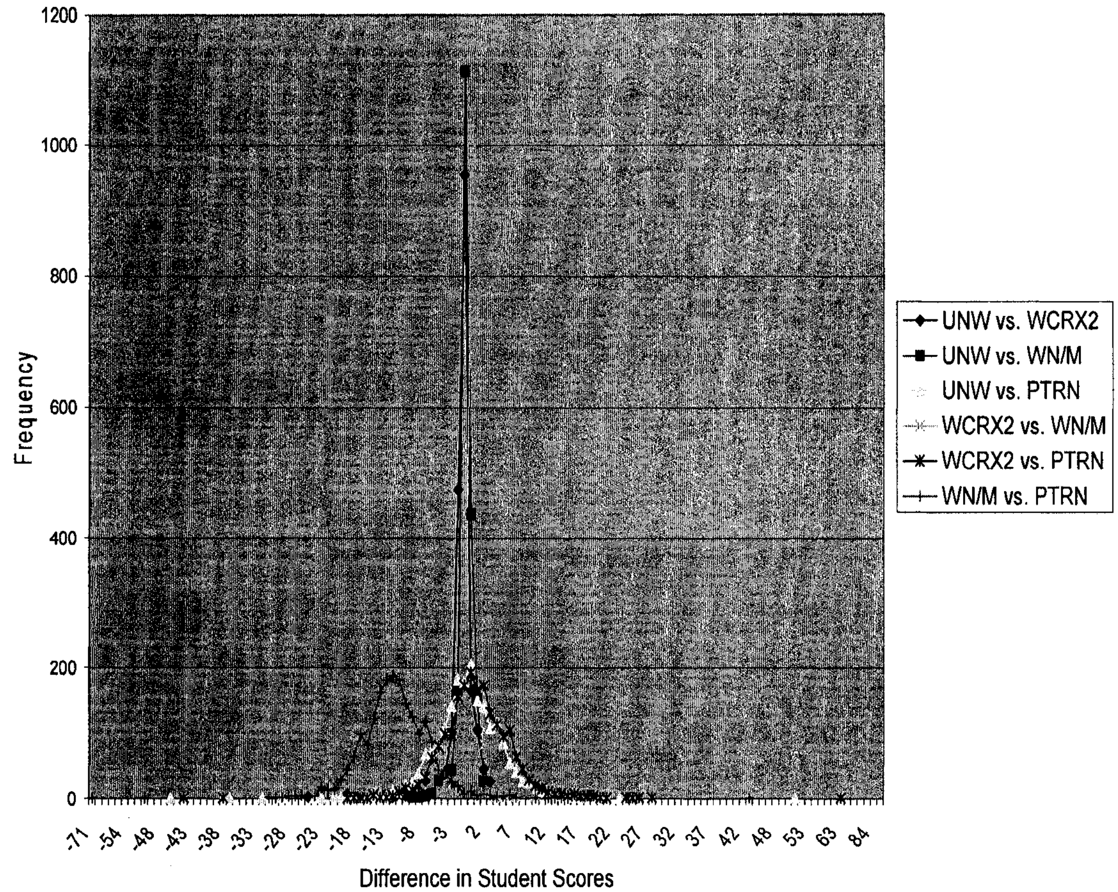

Figure 24. Differences Between Unweighted, Weighted CRx2, Weighted N/M and Pattern Scores: Sample 2

Sample 2 for the UNW procedure, -30 to 48 in Sample 1 and -44 to 63 in Sample 2 for the WCRX2 procedure, and -46 to 36 in Sample 1 and -46 to 59 in Sample 2 for the WN/M procedure. Individual student scores did vary somewhat between the UNW and WN/M and the UNW and WCRX2 procedures, ranging from -25 to 3. The differences between the WCRX2 and WN/M scoring procedures were smaller, varying from -1 to 10.

When using a DTM of 0.50, the differences yielded by the four scoring procedures were significantly different for the vast majority of students for all pairs except the WCRX2 and WN/M pair. For example, in Sample 1 the scores yielded by the UNW procedure and the PTRN scoring procedure were within one DTM for 200 (10%)students. The corresponding numbers for each of the weighted procedures and the PTRN procedure were 177 (8.9%) and 182 (9.1%) for the WCRX2 and WN/M procedures, respectively. In contrast, the UNW and WCRX2 procedures for 481 (24.1%) students, the UNW and WN/M procedures for 632 (31.6%) students and the WCRX2 and WN/M procedures for 1,488 (74.4%) students were within one DTM. If the DTM was to be relaxed to 1.00, the corresponding numbers, taken in the same order, would have been 487 (24.4%), 484 (24.2%), 484 (24.4%), 1550 (77.5%) 1718 (86.0%), and 1928 (96.4%). The results were similar in Sample 2.

Summary

Classical test score statistics indicated that the two samples for both English 30 and Pure Math 30were randomly equivalent and that non-overlapping information was yielded by the SR and constructed CR items. Through analysis with principal component analysis, scree plots and, in the case of the selection items, NOHARM and the accompanying fit statistics, the assumptions of unidimensionality and item independence were met for both samples on both examinations. The assumption of speededness was also met. Thus the assumptions of IRT were met.

At the group level in English 30, the means of the score distributions revealed that the scoring procedures could be placed into two sets for both samples: UNW (483.02) with PTRN (483.40) and WCRX2 (480.82) with WN/M (480.40). The members in each pair differed by less than one DTM and the difference between the two members in one pair and the two members in the second pair differed by more than one DTM. The means of the scores in Pure Math 30 were all less than 500, varying from 495.37 (UNW) to 496.62 (WCRX2). While the means differed by less than one DTM between the WCRX2 and WN/M scoring procedures and between the WN/M and PTRN scoring procedures there was lack of transitivity in that the remaining pair-wise differences between the other scoring procedures. The standard deviations of the four distributions all exceeded 50 for both samples on both examinations. Further, for both English 30 samples the standard deviation of the distribution of UNW scale scores exceeded the standard deviations of the distributions of the remaining three scale scores by more than one DTM; the differences among the standard deviations of the remaining scale scores were within one DTM. For Pure Math 30, there was lack of transitivity between the standard deviations of the distribution. Lastly, the correlations among the four sets of scores were all above 0.98 for English 30 and 1.00 for Pure Math 30; the differences among the six pairs of correlations were all less than one DTM. Taken together, the results first reveal that the scoring procedures at the group level tended to rank the students the same, but differed in their central tendency and variability with the UNW and PTRN scoring procedures and the WCRX2 and WN/M scoring procedures yielding comparable means in English 30.

The magnitudes of the standard errors of measurement varied from 5 to 10 transformed score points for the majority of score points in English 30 and 5 to 20 transformed score points for Pure Math 30. The distributions of the standard errors were parabolic in shape with the English 30 SEs much wider across the center of the distribution. The minimum SE for all procedures occurred around the means with a sharp increase for scores above 550. The UNW scoring resulted in the highest amount of error, and PTRN scoring resulting in marginally lower amounts of error than the remaining score procedures for both examinations. However, for Pure Math 30, at a standard score of about 425 the UNW SEs crossed into the SEs for the three remaining scoring procedures. At that point and until around the scale score of 525, all four scoring procedures had similar SEs. Taken together, these findings suggest that there is less precise measurement of student ability at the higher end of the scale score than at the lower end for all four scoring procedures.

The pattern of percentile scale score differences was similar for both samples in English 30. The WCRX2 and WN/M scoring procedures were similar at the 10th and 50th percentile points but not at the 90th percentile point in both samples, while the UNW and PTRN scoring procedures were similar at the 50th percentile point in both samples. The remaining differences exceeded 0.50 in absolute value. However, there was a lack of transitivity at the 10th, 50th, and 90th percentiles in Pure Math 30, hence, no significant differences among the scores are claimed. The pattern scores at the 10th percentile point were less than the scores yielded by the other three scoring procedures but greater at the 90th percentile point on both examinations. The RMS are consistent with the percentile results with the RMS involving PTRN scores markedly higher than the RMS values for the other three scoring procedures on both examinations.

Proficiency levels results suggest that, although the differences in corresponding percentages were small (all less than two percent) in English 30 and did not fluctuate between UNW, WCRX2, and WN/M scoring procedures in Pure

Math 30, the number of students placed in the levels did vary. Up to 100 students (English 30) and 24 students (Pure Math 30) were placed at a different level if the WCRX2 or WN/M procedures were used in place of the PTRN procedure and up to 66 students if the UNW procedure was used in place of the PTRN procedure in English 30.

Lastly, at the student level, the differences between scores yielded by the four scoring procedures were significantly different (exceeded one DTM) for the vast majority of students for all pairs except the scores yielded by the WCRX2 and WN/M procedures.

Thus, perhaps with the exception of the WCRX2 and WN/M scoring procedures, the scores yielded by the UNW, WCRX2, WN/M and PTRN scoring procedures cannot be used interchangeably for English 30 or Pure Math 30.

CHAPTER 6

A brief summary of the purpose of the study and the procedures followed to address this purpose are presented in the beginning of this chapter. This summary is followed by a discussion of results. The limitations of the study are presented next, followed by the conclusion. The chapter concludes with implications for practice and recommendations for future research.

Purpose and Procedures of the Study

The purpose of this study was to examine the interchangeability of scores yielded by four scoring procedures advanced in the literature when applied to low-stakes achievement tests and to high-stakes school leaving examinations containing both selected response (SR) items and constructed response (CR) items. The four scoring procedures were the unweighted procedure in which scores from the set of SR items and the set of CR items were simply added (UNW); the weighted procedure in which the CR items were given a weight of two while the SR items were weighted one (WCRX2), the weighted procedure in which the CR items were weighted so that they contributed as much to the total scores as the SR items (WN/M), and pattern scores yielded by an Item Response Analysis of the full test. Schaeffer et al. (2002) and Sykes and Hou (2003) examined the comparability of the UNW, WCRX2, PTRN, and other scoring procedures to determine the degree to which the scores were interchangeable. Using low-stakes Grade 8 and Grade 9 examinations, both sets of researchers found that the scoring procedures yielded similar results at the group level. However, they did not present results at the student level. Thus it may be that the procedures considered do not lead to interchangeable scores at the individual student level. Further, it may that the procedures will not lead to interchangeable scores at the group and student level when the tests are high-stakes high-school school leaving examinations.

Two low-stakes tests – the Alberta English and Math 9 provincial achievement tests – and two high-stakes examinations - the English 30 and Pure Math 30 provincial school leaving diploma examinations – were analyzed to provide a replication across two different subject areas. Two random samples of 2,000 students were selected without replacement from the population of students who took each test to allow examination of the stability of the results across samples. A difference that matters (DTM) of 0.50 (Kolen & Brennan, 2004; Dorans, 2004; Dorans & Feigenbaum, 1994) was used to examine differences in scores. The two samples were randomly equivalent for each test and examination. The assumptions underlying the use of item response theory were met for both the selected response (SR) and constructed response (CR) items included in each test and examination.

The SR and CR items were simultaneously calibrated on the same theta scale (Ercikan et al., 1998; Fitzpatrick et al., 1996; Schaeffer et al., 2002; Sykes & Hou, 2003; Sykes & Yen, 2000). The three-parameter logistic model (3PL: Lord 1980) was used for the SR items and a generalization of Masters' (1982) Partial Credit Model was used for the CR items/tasks The parameters were estimated using PARDUX (Burket, 1998), a proprietary computer program developed at CTB-McGraw Hill. PARDUX, as described in Schaeffer et al. (2002), uses a marginal maximum likelihood procedure implemented with the EM algorithm (Bock &

Aitken, 1981). WINFLUX (Burket, 1999), also proprietary and developed at CTB-McGraw Hill, was used to place the item parameters onto a common score scale with a mean of 500 and a standard deviation of 50.

The interchangeability of scores yielded by the four scoring procedures was evaluated at the group and student level. A comparison of means, standard deviations, and standard error was conducted at the group level. At the student level, a) the differences among the four scores were compared at the 10th, 50th, and 90th percentile points, b) the differences between pairs of scores were summarized using the Root Mean Square, c) the comparability of criterion-referenced decisions was assessed with respect to three cut-score points in the distributions of scores, and d) individual student score differences were examined.

Results and Discussion

The results for the four scoring procedures at the group and student levels were quite stable across samples. Consequently, the summary of the results presented here are for Sample 1. First, the results for each test and examination are summarized. These summaries are then followed by a summary of similarities and differences at the group level with the results reported by Schaeffer et al. (2002) and Sykes and Hou (2003) for the low-stakes tests each set of authors considered. This is then followed with a summary that points out similarities and differences among the results for the test and examinations considered.

Low-Stakes Achievement Tests

English 9. The means of the scores were all less than 500, varying from 493.10 (WCRX2) to 494.71 (PTRN). While the means differed by less than one

DTM between the UNW and WN/M scoring procedures and between the UNW and PTRN scoring procedures, there was a lack of transitivity in that the remaining pair-wise differences between the other scoring procedures exceeded one DTM. The standard deviations of the four distributions all exceeded 50 for both samples. Further, the standard deviation of the distribution of UNW scale scores exceeded the standard deviations of the distributions of the remaining three scale scores by more than one DTM; the differences among the standard deviations of the remaining scale scores were within one DTM. The four score distributions were negatively skewed with the exception of PTRN scores. Sample 1 was slightly platykurtic and Sample 2 was slightly leptokurtic. The negative skewness reflects the easiness of the test, and explains why the means of the transformed scores were less than 500. Lastly, the correlations among the four sets of scores were all above 0.96; the differences among the six pairs of correlations were all less than one DTM. Taken together, the results reveal that the scoring procedures at the group level tended to rank the students the same but differed in their central tendency and variability.

The magnitudes of the standard errors of measurement varied from 5 to 10 transformed score points for the majority of score points. The minimum standard error for all the procedures occurred around the means (between 475 and 575). The standard errors then increased, but much more rapidly for scores above 575 than for scores below 475; at a scale score of 700, the SE was around 30 SE points higher than at a scale score of 300. Across the scale scores, the four scoring procedures were similar with UNW scoring resulting in marginally higher amounts of error and PTRN scoring resulting in marginally lower amounts of error than the remaining score procedures, particularly for low scores.

The scores yielded at the 10^{th} and 50^{th} percentile points exhibited a lack of transitivity and are therefore not reported. The remaining differences at the 90^{th} percentile exceeded 0.50 in absolute value. The pattern scores at the 10^{th} percentile point were less than the scores yielded by the other three scoring procedures but greater at the 90^{th} percentile point. The root mean square differences (RMS) were consistent with the percentile results, with the RMS involving PTRN scores markedly higher than the RMS values for the other three scoring procedures while the RMS values for UNW and WN/M were more comparable.

The proficiency level results suggested that the classification percentages did not fluctuate between the UNW and WN/M scoring procedures at the first two proficiency levels. The WCRX2 and WN/M scoring procedures classified the same number of students at Levels 3 and 4. However, the number of students placed in the levels did vary, with up to 118 students being placed at a different level if the PTRN procedure was used in place of the other three scoring procedures.

Lastly at the student level, the differences between scores yielded by the four scoring procedures exceeded one DTM for the vast majority of students for all pairs of scores except the scores yielded by UNW and WN/M procedures. For the latter pair, nearly three in four were within one DTM, while for the other pairs the percentages of scores within one DTM varied between approximately 3% and 37%.

Thus, perhaps with the exception of the UNW and WN/M scoring procedures, the scores yielded by the UWN, WCRX2, WN/M. and PTRN scoring procedures cannot be used interchangeably for English 9.

Math 9. The means of the scores were all less than 500, varying from 492.23 (UNW) to 496.33 (WN/M). The means for the UNW and PTRN scoring procedures differed by less than one DTM; the differences between the means of the other members in each pair were greater than one DTM. The standard deviations of the four distributions all exceeded 50. Further, for both samples the standard deviation of the distributions of UNW and PTRN scale scores were within one DTM; the differences among the standard deviations of the remaining scale scores exceeded one DTM. With the exception of the WN/M procedure, the score distributions were negatively skewed and all four distributions were leptokurtic. Lastly, the correlations among the four sets of scores were all above 0.98 for both samples; the differences among the six pairs of correlations were all less than one DTM. Taken together, the results reveal that the scoring procedures at the group level tended to rank the students the same but differed in their central tendency and variability with the exception of the UNW and PTRN scale scores.

The magnitudes of the standard errors of measurement varied from 10 to 40 transformed score points for the majority of score points. The SEs were parabolic in shape with the overall magnitudes low for the scores around the center of the distribution, and a sharp increase for scores below 400 and above 575. For scale scores less than about 375, the unweighted scoring resulted in the highest amount of error with the differences among the four score procedures being similar. At the scale score of about 400, the UNW and WN/M standard errors crossed. The distribution of the SEs of the UNW standard errors then closely followed the WCRX2 and PTRN distributions, while the WN/M standard errors increased markedly. Across the scale scores, PTRN scoring resulted in marginally lower amounts of error than the remaining score procedures, particularly at the low end of the scale score distributions.

With the exception of the PTRN with the WN/M procedures, the scores of all of the scoring procedures were within one DTM at the 50th percentile. In contrast, the differences between the scores corresponding to the 10th percentile and corresponding to the 90th percentile exceeded one DTM. The PTRN scoring procedure was the lowest, followed in turn by the UNW procedure at the 10th percentile. In contrast, the pattern was reversed at the 90th percentile with the PTRN procedure resulting in the highest scores followed again by the UNW procedure. The RMS were consistent with the percentile results with the RMS involving PTRN scores markedly higher than the RMS values for the other three scoring procedures.

The proficiency level results suggested that classification percentages did not fluctuate between the UNW and WCRX2 scoring procedures at Level 2. The WCRX2 and WN/M scoring procedures classified the same number of students at Levels 3 and 4. However, the number of students placed in the levels did vary, with up to 66 students being placed at a different level if the UNW or WCRX2 procedures were used in place of PTRN at the second level.

Lastly, at the student level the differences yielded by the four scoring procedures were significantly different for the vast majority of students for all pairs.

The percentage of students within one DTM ranged from 8% (WN/M and PTRN) to 30% (UNW and WCRX2). Thus, the scores yielded by the UNW, WCRX2, WN/M and PTRN scoring procedures cannot be used interchangeably for Math 9.

High-Stakes Examinations

English 30. In English 30, the four means suggested that the scoring procedures can be placed in two sets: UNW (483.02) with PTRN (483.40) and WCRX2 (480.82) with WN/M (480.40). The members in each pair differed by less than one DTM and the difference between the two members in one pair and the two members in the second pair differed by more than one DTM. With the exception of WN/M and PTRN in Sample 2, the scoring distributions were slightly positively skewed. The kurtosis suggests a somewhat leptokurtic distribution with the exception of PTRN in Sample 1 which was slightly platykurtic.

The standard deviations of the four distributions all exceeded 50 for both samples. Further, for both samples the standard deviation of the distribution of UNW scale scores exceeded the standard deviations of the distributions of the remaining three scale scores by more than one DTM; the differences among the standard deviations of the remaining scale scores were within one DTM. Lastly, the correlations among the four sets of scores were all above 0.98 were all within one DTM of each other. Taken together, the results first reveal that the scoring procedures at the group level tended to rank the students the same, but differed in their central tendency and variability with the UNW and PTN scoring procedures and the WCRX2 and WN/M scoring procedures yielding comparable means.

The overall magnitudes of the standard errors of measurement varied from 5 to 10 transformed score points for the majority of score points. The minimum SE for all procedures occurred around the means with a sharp increase for scores above 550. The UNW scoring resulted in the highest amount of error, with the three remaining score procedures more similar and smaller. Across the scale scores, the four scoring procedures were similar with UNW scoring resulting in marginally higher amounts of error and PTRN scoring resulting in marginally lower amounts of error than the remaining score procedures, particularly at the low and high end of the scale score distribution.

The WCRX2 and WN/M scoring procedures were similar at the 10th and 50th percentile points but not at the 90th percentile point, while the UNW and PTRN scoring procedures were similar at the 50th percentile point in both samples. The remaining differences exceeded one DTM. The pattern score corresponding to the 10th percentile point was less than the scores yielded by the other three scoring procedures. In contrast, the pattern score corresponding to the 90th percentile point exceeded the scores yielded by the other procedures. Lastly, the two weighted procedures yielded scores greater than the scores yielded by the UNW procedure at the 90th percentile point, but not to as great an extent as observed when pattern scoring was considered. The RMS were consistent with the percentile results with the RMS involving PTRN scores markedly higher than the RMS values for the other three scoring procedures.

The proficiency level results suggested that, although the differences in corresponding classification percentages were small (all less than two percent), the number of students placed in the levels did vary, with up to 100 students being placed at a different level if the WCRX2 or WN/M procedures were used in place of the PTRN procedure, and up to 66 students if the UNW procedure was used in place of the PTRN procedure.

Lastly, at the student level, the differences between scores yielded by the four scoring procedures exceeded one DTM for the vast majority of students for all pairs of scores except the scores yielded by the WCRX2 and WN/M pair. For the latter pair, over three in five were within one DTM, while the percentages for the other pairs of scoring procedures varied between approximately 1.0% (UNW and WN/M) to 3.9% (UNW and PTRN).

Thus, perhaps with the exception of the WCRX2 and WN/M scoring procedures, the scores yielded by the UNW, WCRX2, WN/M and PTRN scoring procedures cannot be used interchangeably for English 30.

Pure Math 30. The means of the scores were all less than 500, varying from 495.37 (UNW) to 496.62 (WCRX2). While the means differed by less than one DTM between the WCRX2 and WN/M scoring procedures and between the WN/M and PTRN scoring procedures there was lack of transitivity in that the remaining pair-wise differences between the other scoring procedures differed by more than one DTM. The standard deviations of the four distributions all exceeded 50 for both samples. Further, while the differences among the standard deviations between the WN/M and PTRN scoring procedures and the UNW and PTRN procedures were within one DTM there was lack of transitivity in that the between the standard deviations of the distribution of the remaining scale scores exceeded one DTM. The correlations among the four sets of scores were all 1.00. Taken together, the results first reveal that the scoring procedures at the group level tended to rank the students the same but differed in their central tendency and variability.

The magnitudes of the standard errors of measurement varied from 5 to 20 transformed score points for the majority of score points. The SE distributions were parabolic in shape, with low to moderate standard errors around the means (between 400 and 550), and increasing values on either side of this interval. The standard errors then increased more rapidly for scores above 550 than the standard errors for scores below 400. At a scale score of 300 the standard errors are widely spread especially between UNW and PTRN with a difference of about 20 SE points. At a scale score of 700 the standard errors are closely grouped with a range of about 5 SE points. The UNW scoring procedure resulted in the highest amount of error than the three remaining score procedures until standard score of 450 where UNW crossed into the three remaining scoring procedures. As such, all four scoring procedures had similar SEs from around 425 to 525. Across the scale scores, PTRN scoring resulted in marginally lower amounts of error than the remaining score procedures, particularly at the low end of the scale score distribution.

The pattern of percentile scale score differences was similar in both samples. However, there was a lack of transitivity at the 10^{th}, 50^{th}, and 90^{th} percentiles. Hence, no significant differences among the scores are claimed. The pattern scores at the 10^{th} percentile point were less than the scores yielded by the other three scoring procedures but greater at the 90^{th} percentile point. The RMS were consistent with the percentile results, with the RMS involving the PTRN scores markedly higher than the

RMS values for the other three scoring procedures and with the RMS values for UNW and WN/M were more comparable.

The proficiency level results suggested that classification results did not fluctuate between the UNW, WCRX2 and WN/M scoring procedures at each level. The PTRN scoring procedure resulted in 17 fewer students placed in the first level, 1 fewer in the second level, 8 more at the third level, and 24 fewer in the fourth level than the two weighted scoring procedures.

Lastly, at the student level, the differences between scores yielded by the four scoring procedures were exceeded one DTM for the vast majority of students for all pairs except the scores yielded by the WCRX2 and WN/M procedures. For the latter pair, three in four were within one DTM, while the percentages for the other pairs of scoring procedures varied between approximately 8.9% (WCRX2 and PTRN) to 31.6% (UNW and WN/M).

Thus, perhaps with the exception of the WCRX2 and WN/M scoring procedures, the scores yielded by the UNW, WCRX2, WN/M and PTRN scoring procedures cannot be used interchangeably for Pure Math 30.

Summary of Results and Discussion

Taken together, the results presented above reveal that 1) as pointed out earlier, the descriptive analyses were stable across samples thus no notable differences were noted between the four scoring procedures at the group level, 2) differences were noted at the student level: pattern scoring generally had the lowest SEs and resulted in the greatest differences of scores across all four tests; pattern scoring generally had the greatest differences at the 10^{th} and 90^{th} percentiles; pattern scoring resulted with the greatest number of students affected at the four proficiency levels; differences in individual student scaled scores were most pronounced when pattern scoring was involved, and 3) results appear to be a function of the raw score weight of the SR and CR items.

1. The four scale means for the four exams ranged from 480.40 to 498.51. Although the same transformation was used by Schaeffer et al. (2002) and Sykes and Hou (2003), their means ranged from slightly below 500 (493.1) to slightly above 500 (504.33). The difference between those findings and the findings in the present study occurred because the mean ability estimates on the IRT theta scale in the present study were all less than zero (-0.04 to -0.20). Consequently, when the scores were transformed, the means were less than 500. The standard deviations, which ranged from 54.25 to 64.21, were consistent with those found by Sykes and Hou (2003) and Schaeffer et al. (2002). The skewness and kurtosis values revealed the low-stakes tests were generally negatively skewed while the high-stakes examinations were positively skewed both with a relatively flat mode.

2. Differences were noted at the student level:

    2.1. As in Schaeffer et al. (2002) and Sykes and Hou (2003), the SEs were higher at the lower and higher ends of the scale score distributions, ranging from 40 to 115 at the lower end of the scale score distribution and 48 to 115 at the upper end of the scale score distribution; all of the SE distributions were parabolic in shape, with the bottom of the parabolas in English 9 and English 30 being wider than that of Math 9 and Pure Math 30. The PTRN scoring procedure resulted in SEs that were consistently less than the UNW and two weighted procedures across all four tests with the greatest differences occurring in the lowest portion of the scale (300 through 400). Unlike Schaeffer et al. (2002) and Sykes and Hou (2003), the UNW SEs crossed the SEs for the three remaining scoring procedures on both Math exams while it stayed above the SEs for the other three scoring procedures throughout the SE distribution on the two English exams. The resulting lower standard errors for the PTRN scoring procedure assures that the scores are more accurate estimates of each student's true score than the other three scoring procedures particularly for students at the lower end of the ability distribution.

2.2. The percentile scale score differences for all four tests resulted in PTRN scoring with the greatest variability of scores at the 10^{th} and 90^{th} percentiles. Across all four tests, the PTRN scores were consistently lower than the UNW, WCRX2, and WN/M scoring procedures at the 10^{th} percentile (ranging from -6.00 to -21.00) and higher than the other three scoring procedures at the 90^{th} percentile (6.00 to 19.00).

2.3. The proficiency level results suggested that the classification percentages fluctuated somewhat (from 0% to 2%) between the

UNW, WCRX2, and WN/M scoring procedures at the four proficiency levels across the four tests. However, the number of students placed in the levels did vary when the PTRN procedure was compared with the other three scoring procedures with up to 118 students being placed at a different level.

2.4. The differences in individual student scaled scores were most pronounced between the PTRN and UNW, WCRX2, and WN/M procedures on all four tests. The differences ranged from -46 to +112 for PTRN and UNW; -60 to +117 for PTRN and WCRX2; and -59 to +85 for PTRN and WN/M.

2.5. When considering all of the analyses conducted, the two low-stakes exams were similar only in proficiency levels, where the WCRX2 and WN/M scoring procedures had the same results in Levels 3 and 4 for English 9 and Math 9.

3. The means of the WCRX2 and WN/M scoring procedures for Pure Math 30 and English 30 were within one DTM; differences between the WCRX2 and WN/M scoring procedures on both English 30 and Pure Math 30 were the same at the 10^{th} and 50^{th} percentiles; the results for both the WCRX2 and WN/M scoring procedures were the same across all four proficiency levels on both exams; and the results at the individual level were within one DTM for almost three out of four of the students for the WCRX2 and WN/M scoring procedures for both English 30 and Pure Math 30. This was likely due to the similarity between the weight

for the CR items on WCRX2 and the WN/M for both examinations. For example, the WN/M on English 30 at 2.33 and the WN/M for Pure Math 30 at 1.7. Thus, if the high-stakes exam being investigated does not maintain the pattern of close to twice as many SR as CR items, this finding may not hold true.

Limitations of the Study

Although research was provided to support the claim that the PATs are low-stakes and the DIPs are high-stakes examinations (see Chapter 2), a more thorough assessment of the stakes of the achievement tests and diploma examinations would be beneficial. The motivation of the students writing PATs and DIPs in Alberta needs to be assessed to determine better whether the Grade 9 PATs are indeed low-stakes in light of the inclusion of these tests as part of the accountability pillar of the Government of Alberta's four pillars of education, and the reported inclusion of the scores from these exams being included in the final grade of the year.

Conclusions

1. At the group level, the four scoring procedures yielded similar results on all four tests. The scale score distributions and correlational patterns were comparable.
2. At the student level, the four scoring procedures did not yield scale score distributions that were sufficiently similar to warrant using the procedures interchangeably.

3. Pattern scoring provided the smallest standard errors of the four scoring procedures, particularly at the lower end of the ability distribution.
4. Stakes was not a factor affecting the four scoring methods.
5. Differences noted between the two English and the two Math tests suggest subject is a factor affecting the scale score distributions.
6. The four scoring methods can be used for norm referenced tests without bias. However, the four scoring methods result in different student scale scores and thus would not be appropriate for criterion referenced situations like those used by Alberta Education.

Implications for Practice

At the group level, the scores from the four scoring procedures were stable, thus scores at this level may be interpreted interchangeably. However, at the student level, it was found that the four scoring procedures did not yield scale score distributions that were sufficiently similar to warrant using the procedures interchangeably especially on criterion-referenced tests like those used by Alberta Education. As a result, student scores and ultimately decisions made based on those scores may be affected. This can potentially harm students in that their opportunity for graduation and scholarship may be altered depending on which scoring procedure is used.

As such, researchers and government officials should carefully consider the implications of which scoring procedure is chosen for each particular test and examination. For example, in Alberta three cut-scores are set to distinguish between those students who demonstrate a standard of excellence, those who demonstrate acceptable standard, and those who do not. Pattern scoring provided the lowest standard error with the unweighted scoring procedure resulting in the highest standard error. When a cut-score is set, the amount of error at the location of the cut-score is an important consideration. It follows then that the standard error resulting from each scoring procedure is one issue that must be carefully considered. Once a procedure is chosen, a detailed justification and procedure for use should be provided to the stake holders, including the education community, students and parents.

Recommendations for Future Research

- Further examination of the motivation of the students writing PATs in Alberta would be beneficial to determine better whether the Grade 9 PATs are indeed low-stakes.
- An expansion to other provinces, territories, and states in which low-stakes tests and/or high-stakes examinations are administered will determine if the findings noted in Alberta in the present study are consistent across the country.
- Further examination of the effects of the four scoring procedures in other subject areas with tests and examinations with SR and CR items is warranted.
- Student performance was high on the tests and examinations used in the current study. An examination of low-stakes and high-stakes examinations with normal distributions will address the possible effects of the negative skewness and kurtosis found in this study.

- The raw score weighting on both high-stakes examinations resulted in the selection items worth about twice as much as the construction items. Examination of high-stakes examinations without this weighting scheme may result in differences at the student level that were not noted in this study.
- Finally, a simulation study that addresses the use of the four scoring procedures on simulated data with both SR and CR items would
    - provide a benchmark for distributions that are both normally distributed and distributions that are negatively skewed, and
    - provide an opportunity to explore the impact of the increased variability due to the multiple score levels of the constructed response items over the selected response items.

References

Alberta Education. (2004a). Achievement general information bulletin. [On-line]. Available: http://www.education.gov.ab.ca/k%5F12/testing/achievement/ach%5Fgib/default.asp

Alberta Education. (2004b). Assessment highlights Pure Mathematics 30 2003–2004 school year. [On-line] Available: http://www.education.gov.ab.ca/k%5F12/ testing/ diploma/ highlights/ MA30Pure.pdf

Alberta Education. (2004c). Diploma examinations program. [On-line]. Available: www.education.gov.ab.ca/k%5F12/testing/diploma/ dip_gib/ examinationprogram.asp

Alberta Education. (2004d). English Language Arts 30 information bulletin. [On-line]. Available: http://www.education.gov.ab.ca/k_12/testing/diploma/bulletins/humanities/eng301/default.asp

Alberta Education. (2004e). Grade 9 English Language Arts information bulletin. [On-line]. Available: http://www.education.gov.ab.ca/k_12/testing/achievement/bulletins/ Gr9_ELA/gr9_ela_toc.asp

Alberta Education. (2004f). Grade 9 Mathematics information bulletin. [On-line]. Available: http://www.education.gov.ab.ca/k_12/testing/achievement/bulletins/Gr9_Math/gr9_math_toc.asp

Alberta Education. (2004g). Multiyear report 1999/2000 to 2003/2004. [On-line]. Available: http://www.education.gov.ab.ca/k_12/testing/results_2004/dip/multi_reports.asp

Alberta Education. (2004h). Multiyear report 2000 to 2004. [On-line]. Available: http://www.education.gov.ab.ca/k_12/testing/results_2004/ach/multiyr.asp

Alberta Education. (2004i). Pure Mathematics 30 information bulletin. [On-line]. Available: http://www.education.gov.ab.ca/k%5F12/testing/diploma/bulletins/math-science/pure_ma30/default.asp

Alberta Education. (2005). Guidelines for interpreting diploma examination multiyear reports. [On-line]. Available: http://www.education.gov.ab.ca/k_12/testing/multipublic/dip/dipguide_multi.htm

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

Bennett, R.E., Rock, D.A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. Journal of Educational Measurement, 28, 77-92.

Bock, R.D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 37, 29-51.

Bock, R.D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. Applied Psychological Measurement, 12, 261-280.

Bridgeman, B., & Rock, D.A. (1993). Relationships among multiple-choice and open-ended analytical questions. Journal of Educational Measurement, 30, 313-329.

British Columbia Education. (2003). Interpreting and communicating British Columbia Foundation Skills Assessment results. [On-line]. Available: http://www.bced.gov.bc.ca/ assessment/ fsa/fsa_interpretation_2003.pdf

Brown, S.M., & Walberg, H.J. (1993). Motivational effects on test scores of elementary students. Journal of Educational Research, 86, 133-136.

Bryk, A., Raudenbush, S., & Congdon, R. (1996). HLM4: Hierarchical linear & nonlinear modeling [Computer software]. Chicago: Scientific Software.

Burket, G.R. (1998). PARDUX for Windows (Version 1.17). [Computer software]. Monterey, CA: CTB/McGraw-Hill.

Burket, G.R. (1999). WINFLUX (Version 1.01) [Computer software]. Monterey, CA: CTB/McGraw-Hill.

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

DeMars, C.E. (2000). Test stakes and item format interactions. Applied Measurement in Education, 13, 55-77.

Dorans, N.J. (2004). Using subpopulation invariance to assess test score equity. Journal of Educational Measurement, 41, 43-68.

Dorans, N.J. & Feigenbaum, M.D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I.M. Lawrence, N.J. Dorans, M.D. Feigenbaum, N.J. Feryok, A.P. Schmitt, & N.K. Wright (eds.), Technical issues related to the introduction of the new SAT and PSAT/NMSQT, (RM-94-10). Princeton, NJ. Educational Testing Service, 91-122.

Engelhard, G. (2002). Monitoring raters in performance assessment. In G. Tindal. & T.M. Haladyna (Eds.), Large-scale assessment programs for all students: Validity, technical adequacy, and implementation (pp.261-288). Mahwah, NJ: Lawrence Erlbaum Associates.

Ercikan, K. (2002). Scoring examinee responses for multiple inferences: Multiple scoring in assessments. Educational Measurement, Issues and Practice, 21, 8-15.

Ercikan, K., Schwarz, R.D., Julian, M.W., Burket, G.R., Weber, M.M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. Journal of Educational Measurement, 35, 137-154.

Fitzpatrick, A.R., Link, V.B., Yen, W.M., Burket, G.R., Ito, K., & Sykes, R.C. (1996). Scaling performance assessments: A comparison of one-parameter and two-parameter partial credit models. Journal of Educational Measurement, 33, 291-314.

Fraser, C. (1988). NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory. Armidale, Australia: The University of New England.

Glass, G.V., & Hopkins, K.D. (1996). Statistical Procedures in Education and Psychology, 3rd Ed. Needham Heights, MA: Allyn & Bacon.

Goldberg, G.L., & Roswell, B.S. (2001). Are multiple measures meaningful?: Lessons from a statewide performance assessment. Applied Measurement in Education, 14, 125-150.

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

Kiplinger, V.L., & Linn, R.L. (1992). Raising the stakes of test administration: The impact on student performance on NAEP. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles. (ERIC Document Reproduction Service No. ED378221)

Kolen, M.J., & Brennan, R.L. (2004). Test equating, scaling, and linking: Procedures and practices, 2nd Ed. Springer.

Lewis, D.M., Mitzel, H.C., & Green, D.R. (1996, June). Standard setting: A Bookmark approach. In D.R. Green (Chair), IRT-based standard-setting procedures utilizing behavioral anchoring. Symposium conducted at the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.

Li, H., & Stout, W. (1995). A version of Dimtest to assess latent trait unidimensionality for mixed polytomous and dichotomous item response data. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Lord, F.M. (1950). Notes on comparable scales for test scores. (ETS RB-50-48) Princeton, NJ. Educational Testing Service.

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

Marascuilo, L. A. (1966). Large-sample multiple comparisons. Psychological Bulletin, 65, 280-290.

Masters, G.N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

McDonald, R.P. (1967). Nonlinear factor analysis. Psychometric Monographs, No. 15.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23, 13-23.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159-176.

Paris, S.G., Lawton, T.A., & Turner, J.C. (1992). Reforming achievement testing to promote students' learning. In C. Collins & J.M. Mangieri (Eds.), Teaching thinking: An agenda for the 21st century (pp.223-241). Hillside, NJ: Lawrence Erlbaum Associates, Inc.

Phelps, R.P. (1998). The demand for standardized student testing. Educational Measurement: Issues and Practice, 17, 5-23.

Phelps, R.P. (2000). Trends in large-scale testing outside the United States. Educational Measurement: Issues and Practice, 19, 11-21.

Principles for fair student assessment practices for education in Canada. (1993). Edmonton Alberta: Joint Advisory Committee.

Rodriguez, M.C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. Journal of Educational Measurement, 40, 163-178.

Rudner, L.M. (2001). Informed test component weighting. Educational Measurement: Issues and Practice, 20, 16-19.

Schaeffer, G.A., Henderson-Montero, D., Julian, M., & Bene, N.H. (2002). A comparison of three scoring procedures for tests with selected-response and constructed-response items, Educational Assessment, 8, 317-340.

Sykes, R.C., & Hou, L. (2003). Weighting constructed-response items in IRT-based exams. Applied Measurement in Education, 16, 257-275.

Sykes, R.C., & Yen, W.M. (2000). The scaling of mixed-item format test with the one-parameter and two-parameter partial credit models. Journal of Educational Measurement, 37, 221-244.

Thissen, D., Nelson, L., Rosa, K., & McLeod, L.D. (2001). Item response theory for items scored in more than two categories. In D. Thissen & H. Wainer (Eds.), Test scoring (pp.141-186). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), Test scoring (pp.73-140). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., Wainer, H., & Wang, X. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. Journal of Educational Measurement, 31, 113-123.

Tindal, G. (2002). Large-scale assessments for all students: Issues and options. In G. Tindal. & T.M. Haladyna (Eds.), Large-scale assessment programs for all students: Validity, technical adequacy, and implementation (pp.1-24). Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of construction. Applied Measurement in Education, 6, 103-118.

Van den Bergh, H. (1990). On the construct validity of multiple-choice items for reading comprehension. Applied Psychological Measurement, 14, 1-12.

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

Wolf, L.F., & Smith, J.K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. Applied Measurement in Education, 8, 227-242.

Wolf, L.F., Smith, J.K., & Birnbaum, M.E. (1995). Consequence of performance, test motivation, and mentally taxing items. Applied Measurement in Education, 8, 341-351.

Yen, W.M. (1981). Using simulation results to choose a latent trait mode. Applied Psychological Measurement, 24, 185-201.

Yen, W.M. (1984). Obtaining maximum likelihood trait estimates for number-correct scores for the three-parameter logistic model. Journal of Educational Measurement, 21, 93-111.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30, 187-213.

Yen, W.M. (1999). Selected item response theory scoring options for estimating trait values (Internal memorandum). Monterey, CA: CTB/McGraw Hill.

Yen, W.M., & Candell, G.L. (1991). Increasing score reliability with item-pattern scoring: An empirical study in five score metrics. Applied Measurement in Education, 4, 209-228.

Zieky, M.J., & Livingston, S.A. (1977). Manuals for setting standards on the Basic Skills Assessment tests. Princeton, NJ: Educational Testing Service.

Appendix A: Low-Stakes Test Specifications

Table A1

Grade 9 Language Arts PAT Specifications

Dimensions Of Three Main Topic Areas	
Narrative/Essay Writing	
Reporting Category	Description of Writing Assignment
Content (selecting ideas and details to achieve a purpose) Students respond to a given prompt by writing a narrative or essay. Students establish their purpose, select ideas and supporting details to achieve the purpose, and communicate in a manner appropriate to their audience.	The writing assignment requires students to respond to a prompt that consists of a topic and a collection of materials that students may use, if they wish. These materials include graphics, quotes, and short literary excerpts. Students may use ideas from previous experience and/or reading. Students are to respond by writing a narrative or essay.
Organization (organizing ideas and details into a coherent whole) Students organize their ideas to produce a unified and coherent narrative/essay that links events and details, sentences, and paragraphs to support the purpose.	
Sentence Structure (structuring sentences effectively) Students control sentence structure and use a variety of sentence types, beginnings, and sentence lengths to enhance communication.	
Vocabulary (selecting and using words and expressions correctly and effectively) Students choose specific words and expressions that are appropriate for their audience and effective in establishing a voice/tone that will help achieve their purpose.	
Conventions (using the conventions of written language correctly and effectively) Students use conventions accurately and effectively to communicate.	

Table A1 (Continued)

Functional Essay	
Reporting Category	Description of Writing Assignment
Content (thought and detail) Students develop, organize, and evaluate ideas for a specified purpose and audience.. Content Management (using the conventions of written language correctly and effectively) Students communicate clearly and effectively by selecting words and phrases appropriate to their purpose. Students demonstrate control of sentence structure, usage, mechanics, and format.	The functional writing assignment requires students to write to a specified audience in the context of a business letter. They are also expected to correctly address a blank envelope.

Reading			
Reporting Category	Language Function		Total
	Informational	Narrative/Poetic	Questions %
Identifying and Interpreting Ideas and Details Students recognize explicit or implicit ideas and details and make inferences about the relationships between ideas and details.	6	11	17 (31%)
Interpreting Text Organization Students identify and analyze genre. Students identify and analyze the author's choice of form, organizational structure, style, literary techniques, text features, and conventions.	5	6	11 (20%)
Associating Meaning Students use contextual clues to determine the connotative meaning of words, phrases, and figurative language.	5	6	11 (20%)
Synthesizing Ideas Students make generalizations by integrating information from an entire selection in order to identify the purpose, theme, main idea, or mood of the selection.	6	10	16 (29%)
Number (Percentage)of Questions	22 (40%)	33 (60%)	55 (100%)

(Alberta Education, 2004e)

Table A2

Grade 9 Mathematics PAT Specifications

	General Outcomes	Reporting Category: Knowledge	Reporting Category: Skills	Total Questions %
Dimensions of Four Main Topic Areas	Number • Explain and illustrate the structure and the interrelationship of the sets of numbers within the rational number system • Develop a number sense of powers with integral exponents and rational bases • Use a scientific calculator or a computer to solve problems involving rational numbers • Explain how exponents can be used to bring meaning to large and small numbers, and use calculators or computers to perform calculations involving these numbers	4	9	13 (26%)
	Patterns and Relations •Generalize, design, and justify mathematical procedures by using appropriate patterns, models, and technology • Solve and verify linear equations and inequalities in one variable • Generalize arithmetic operations from the set of rational numbers to the set of polynomials	4	11	15 (30%)
	Space and Shape • Use trigonometric ratios to solve problems involving a right triangle • Describe the effects of dimension changes in related 2-D shapes and 3-D objects in solving problems involving area, perimeter, surface area, and volume • Specify conditions under which triangles may be similar or congruent, and use these conditions to solve problems • Use spatial problem solving in building, describing, and analyzing geometric shapes • Apply coordinate geometry and pattern recognition to predict the effects of translations, rotations, reflections, and dilatations on 1-D lines and 2-D shapes	5	9	14 (28%)
	Statistics and Probability • Collect and analyze experimental results expressed in two variables; use technology, as required	3	5	8 (16%)

Table A2 Continued

• Explain the **use of probability and statistics in the** solution of complex problems			
Number (Percentage) of Questions	16 (32%)	34 (68%)	50 (100%)

(Alberta Education, 2004f)

Appendix B: High-Stakes Examination Specifications

Table B1

English Language Arts 30 Examination Specifications

English Language Arts 30 Diploma Examination Part A: Written Response

Description of Writing Assignment	Reporting Category (Scoring Criteria)	Cross-Reference to Program of Studies	Proportion of Total Examination Mark	
			Report Cat.	Sect.
The Personal Response to Texts Assignment requires the student to respond personally, critically, and/or creatively to the content and contexts of a variety of texts while exploring ideas and impressions that the student may also consider in the Critical / Analytical Response to Literary Texts Assignment.	Ideas and Impressions The student is required to reflect on and explore ideas and impressions prompted by the texts and the topic.	2.1 2.2 2.3 4.1	10%	20%
	Presentation The student is required to select an appropriate and effective prose form to convey impressions, to explore ideas, and to create a unifying effect and effective voice. The student is required to communicate clearly.	3.1 3.2 4.1 4.2	10%	
The Critical / Analytical Response to Literary Texts Assignment sets a specific writing topic but allows the student to choose relevant literary text(s) and a procedure of development, and to select supporting details from the chosen literary text(s). The Critical / Analytical Response to Literary	Thought and Understanding The student is required to address the topic by demonstrating an understanding of the ideas developed by the text creator(s) and by analyzing and explaining the personality traits, roles, relationships, motivations, attitudes, and values of characters developed and presented in literary text(s).	2.1 2.2 4.1 4.2	7.5%	30%

Texts Assignment requires the student to understand literal and implied meanings in the chosen text(s) and to synthesize thoughts clearly and express ideas effectively and correctly in writing.	Supporting Evidence The student is required to present relevant support and evidence from a literary text (or texts) to support ideas. Significant appropriate evidence skillfully used is required to create an effective and convincing response.	2.3 3.2 4.1 4.2	7.5%	
	Form and Structure The student is required to develop a coherent, unified composition by choosing an appropriate procedure to create a unified effect. A controlling idea may be implicit or explicit within the composition.	2.2 3.1 4.1 4.2	5%	
	Matters of Choice The student is required to demonstrate a repertoire of stylistic choices and vocabulary in a deliberate, precise, and controlled manner.	4.2	5%	
	Matters of Correctness The student is required to write clearly and correctly, appropriately applying the conventions for written language.	4.2	5%	
Proportion of Total Examination Mark			50%	50%

English Language Arts 30 Diploma Examination Part B: Reading

Reporting Category	A. Form Literal Under-standing	B. Infer, Apply, and Analyze	C. Assess and Form Generaliza-tions	Total Items
1. Construct meaning from content and context, and engage contextual knowledge				30–40 items
2. Relate textual forms, elements, and techniques to content, purpose, and effect				15–25 items
3. Connect self, culture, and milieu to text and text creators				5–15 items
Total Items	5–15 items	30–40 items	15–25 items	70 items (50%)

(Alberta Education, 2004d)

Table B2

Pure Mathematics 30 Examination Specifications

Content	Percent Emphasis
Transformations of Functions	15
Exponents, Logarithms, and Geometric Series	20
Trigonometry	24
Conic Sections	12
Permutations and Combinations	19
Statistics	10

Explanation of Cognitive Levels

Procedural, conceptual, and problem-solving cognitive levels are addressed throughout the examination. The emphasis of each cognitive level was approximately equal.

Procedures

The assessment of students' knowledge of mathematical procedures should involve recognition, defense, execution, and verification of appropriate procedures and the steps contained within them. The use of technology can allow for conceptual understanding prior to specific skill development or vice versa. Students must appreciate that procedures are created or generated to meet specific needs in an efficient manner and thus can be modified or extended to fit new situations. Assessment of students' procedural knowledge will not be limited to an evaluation of their proficiency in performing procedures, but were extended to reflect the skills presented above.

Table B2 (Continued)

Concepts
An understanding of mathematical concepts goes beyond a mere recall of definitions and recognition of common examples. Assessment of students' knowledge and understanding of mathematical concepts should provide evidence that they can compare, contrast, label, verbalize and define concepts, identify and generate examples and counter-examples as well as properties of a given concept, and recognize the various meanings and interpretations of concepts. Students who have developed a conceptual understanding of mathematics can also use models, symbols and diagrams to represent concepts. Appropriate assessment will also provide evidence of the extent to which students have integrated their knowledge of various concepts.
Problem Solving
Appropriate assessment of problem-solving skills is achieved by allowing students to adapt and extend the mathematics they know and encourage the use of strategies to solve unique and unfamiliar problems. Assessment of problem solving involves measuring the extent to which students use these strategies and knowledge, and their ability to verify and interpret results. Students' ability to solve problems develops over time as a result of their experiences with relevant situations that present opportunities to solve various types of problems. Evidence of problem-solving skills is often linked to clarity of communication. Students demonstrating strong problem-solving skills should be able to clearly explain the process they have chosen, using clear language and appropriate mathematical notation and conventions.

(Alberta Education, 2004i)

APPENDIX C
Scree Plots

Low-Stakes Examinations

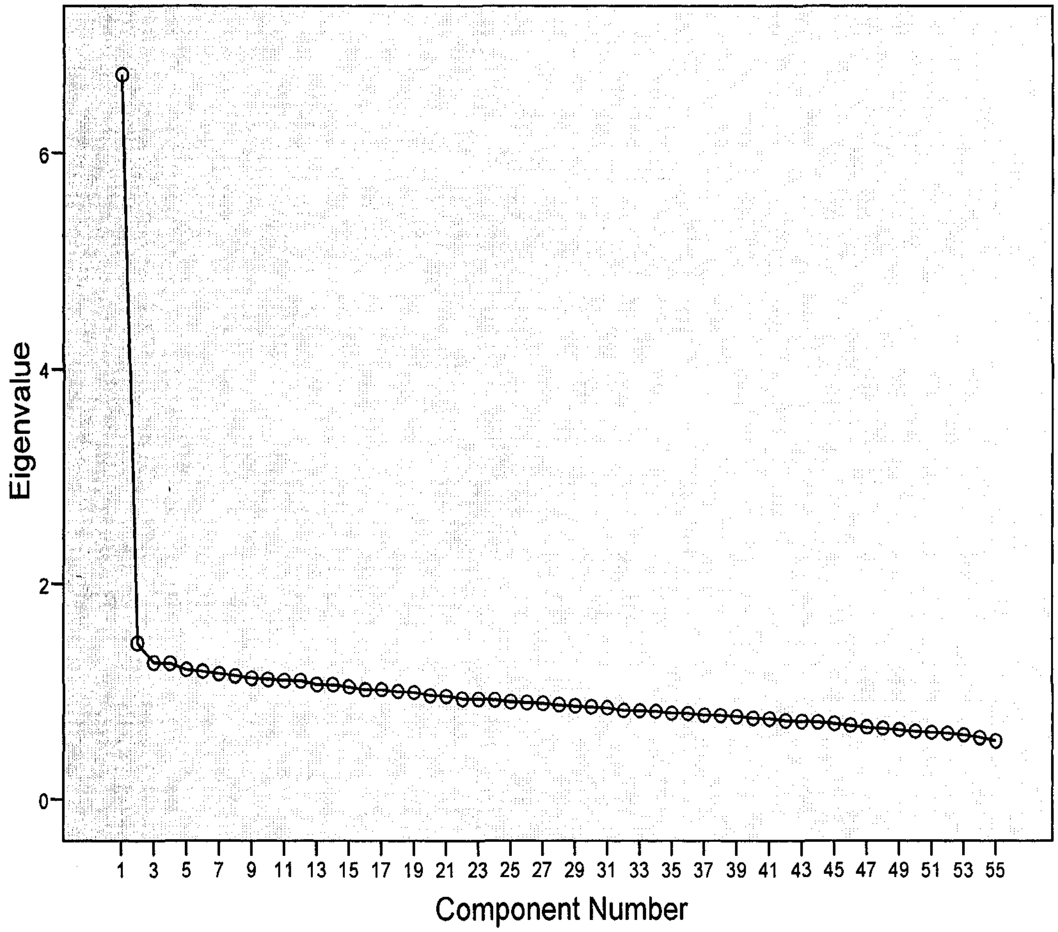

Figure C1 Scree Plot for SR English 9 Sample 1

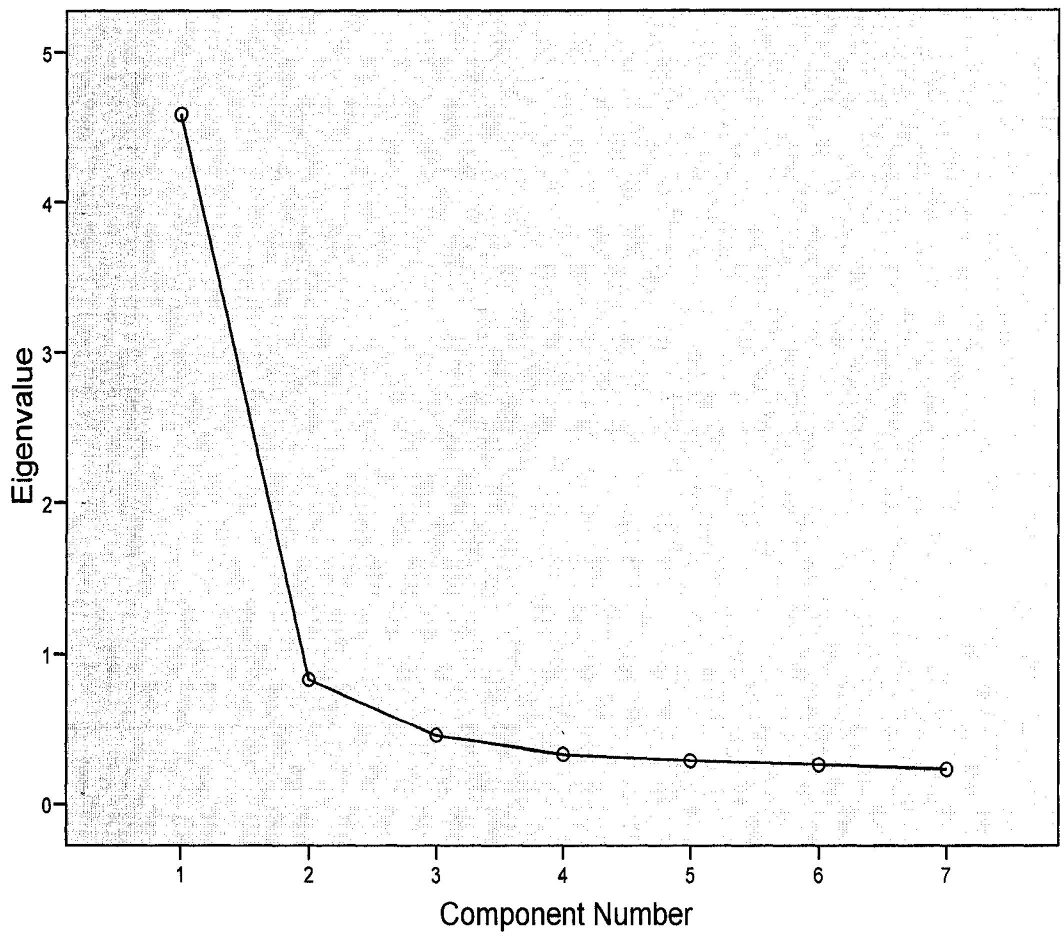

Figure C2 Scree Plot for CR English 9 Sample 1

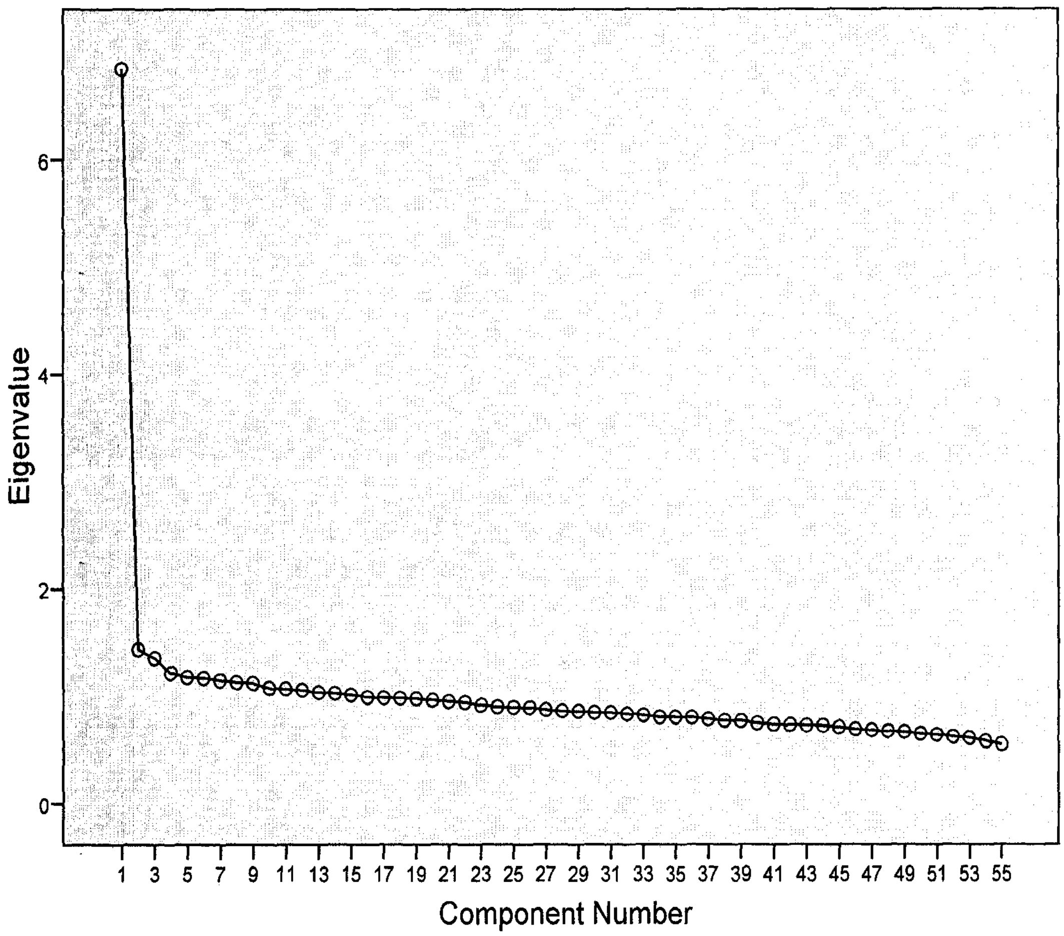

Figure C3 Scree Plot for SR English 9 Sample 2

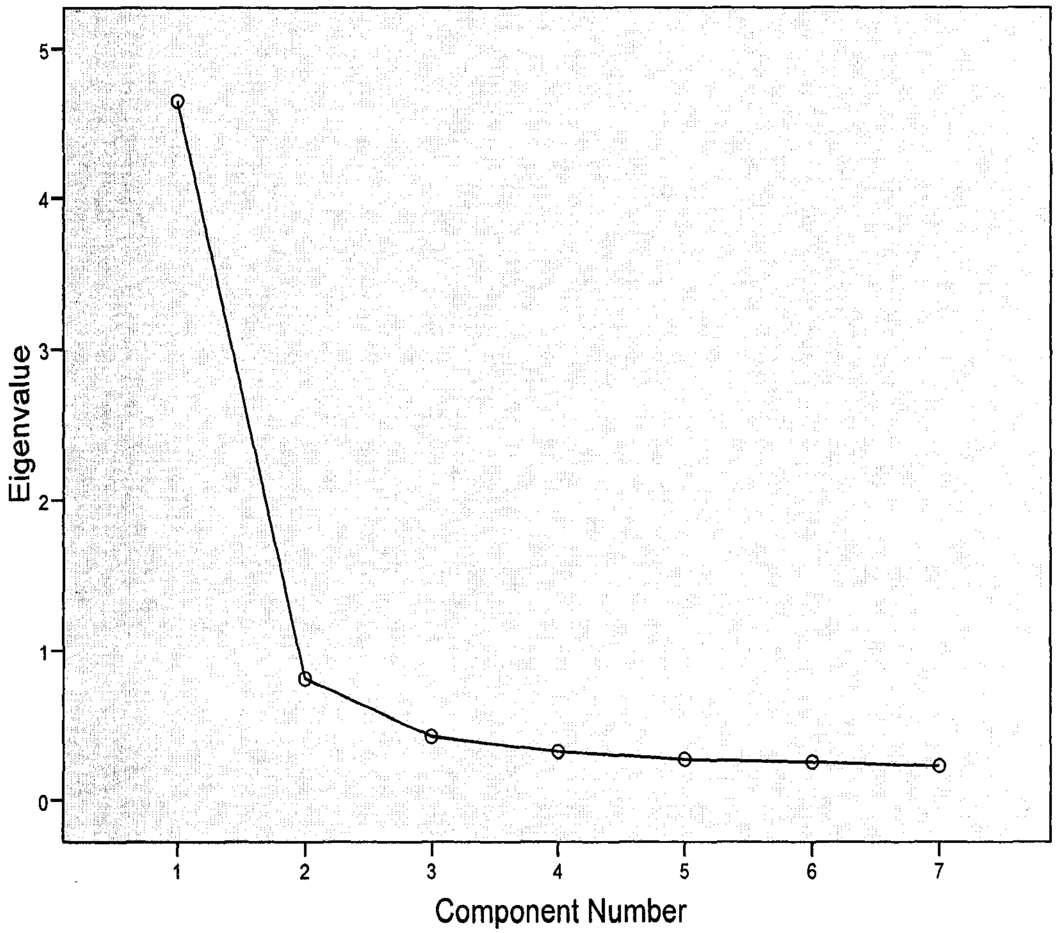

Figure C4 Scree Plot for CR English 9 Sample 2

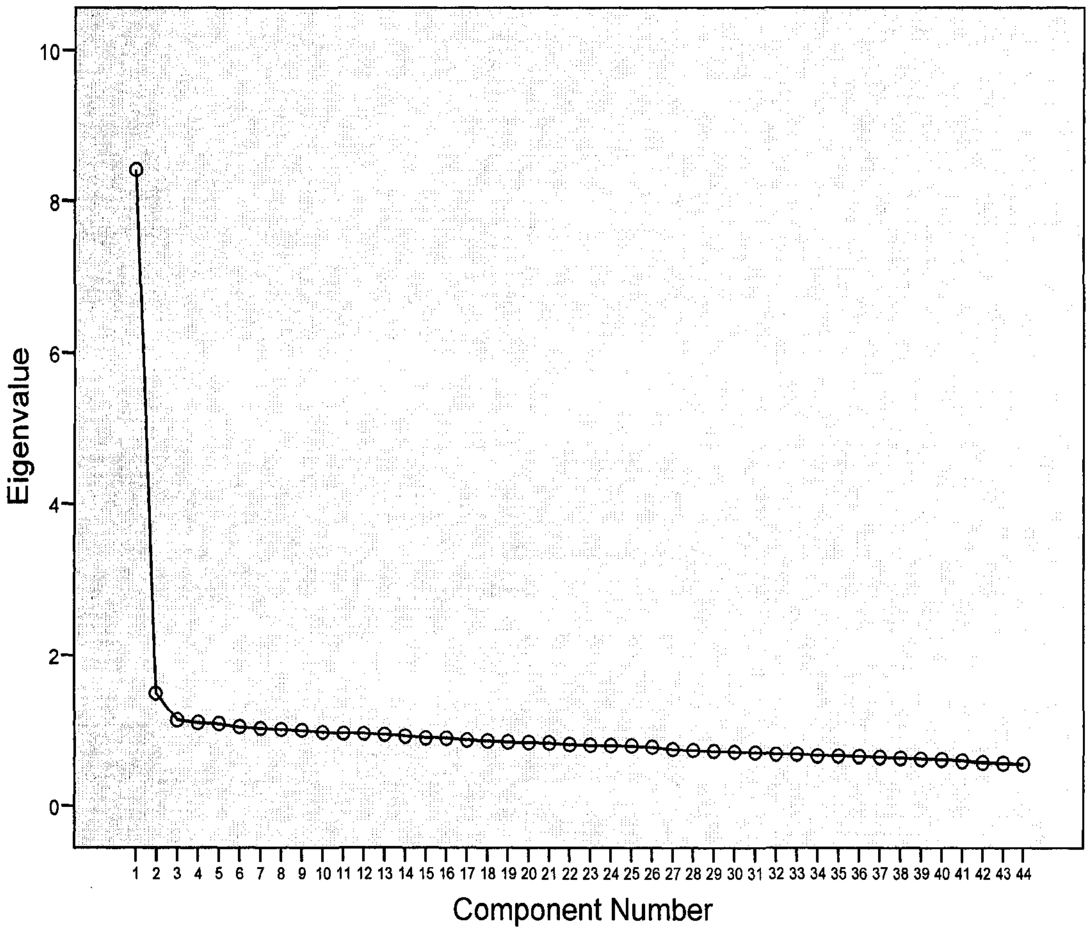

Figure C5 Scree Plot for SR Math 9 Sample 1

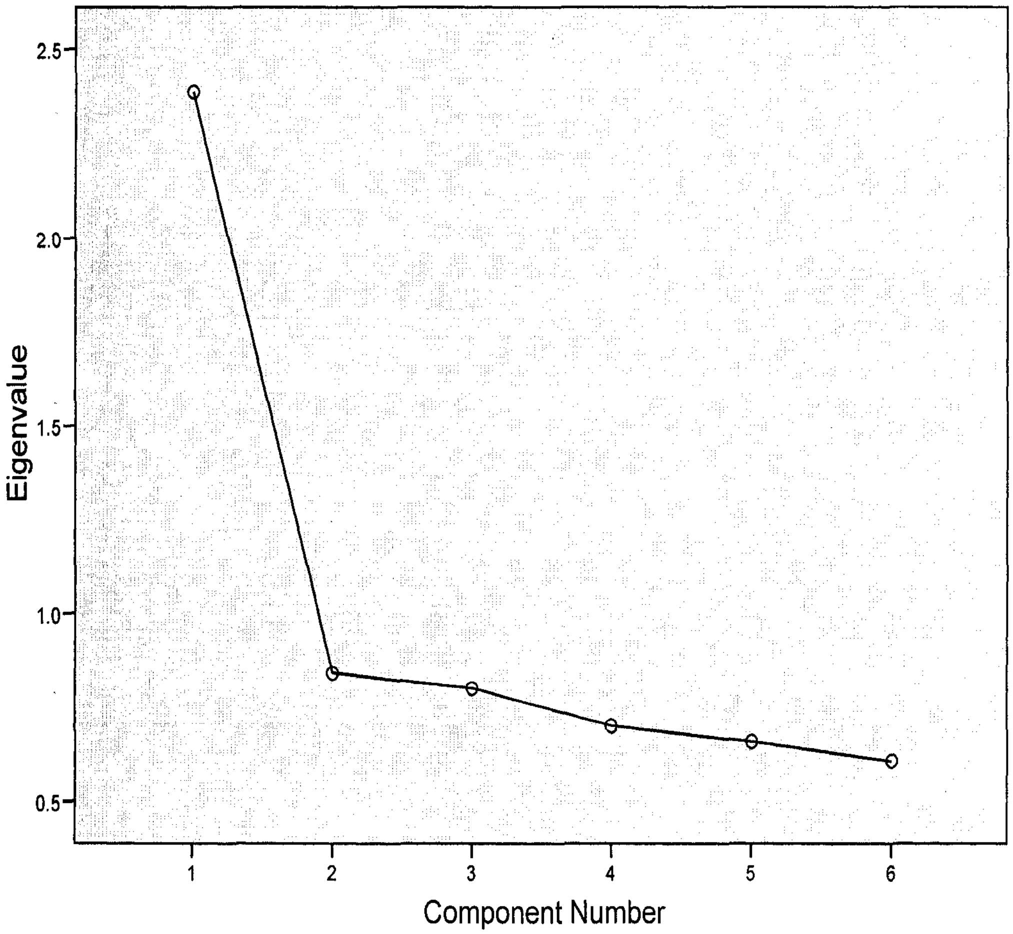

Figure C6 Scree Plot for CR Math 9 Sample 1

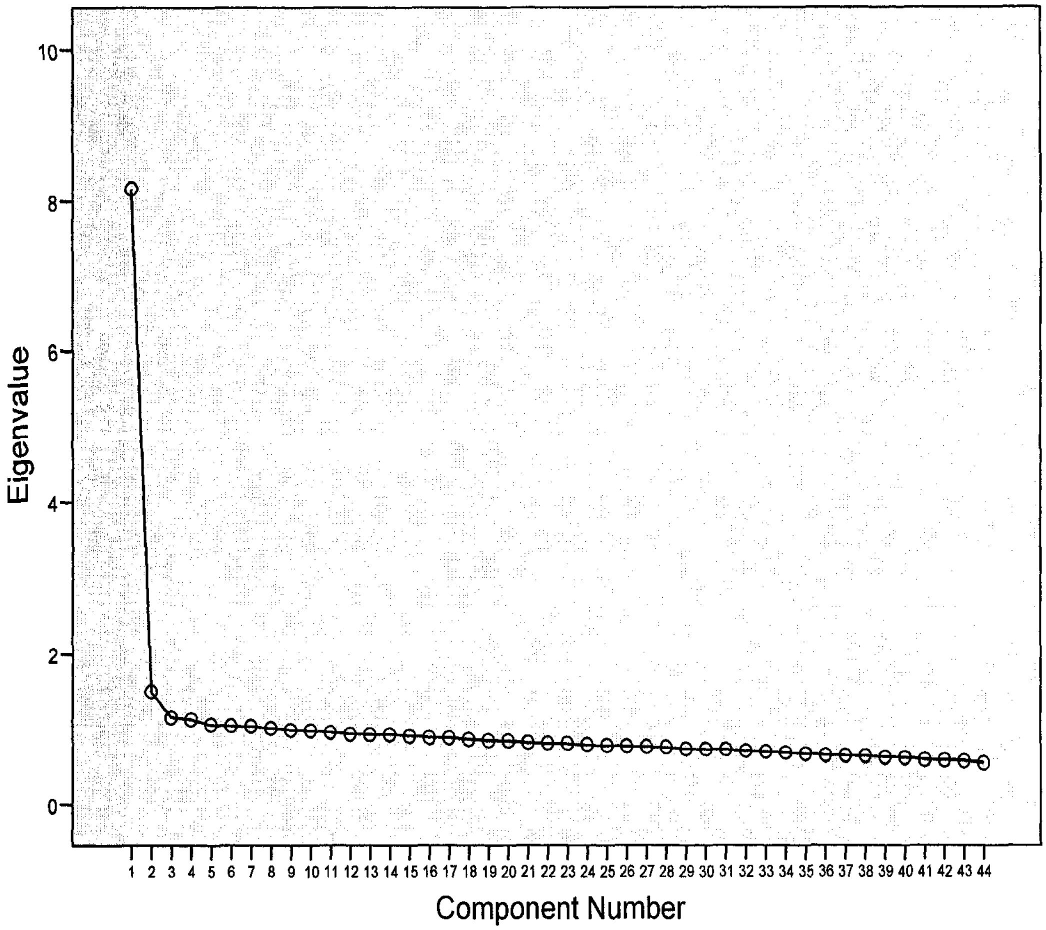

Figure C7 Scree Plot for SR Math 9 Sample 2

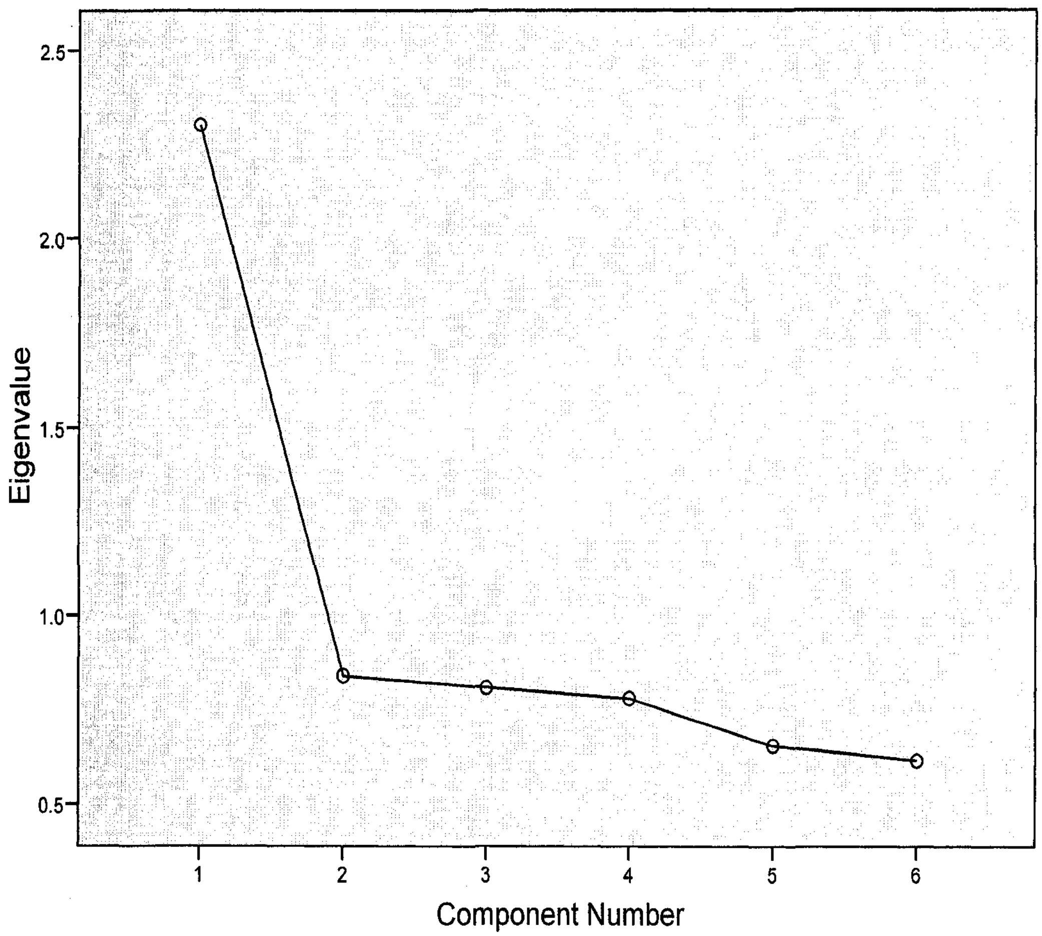

Figure C8 Scree Plot for CR Math 9 Sample 2

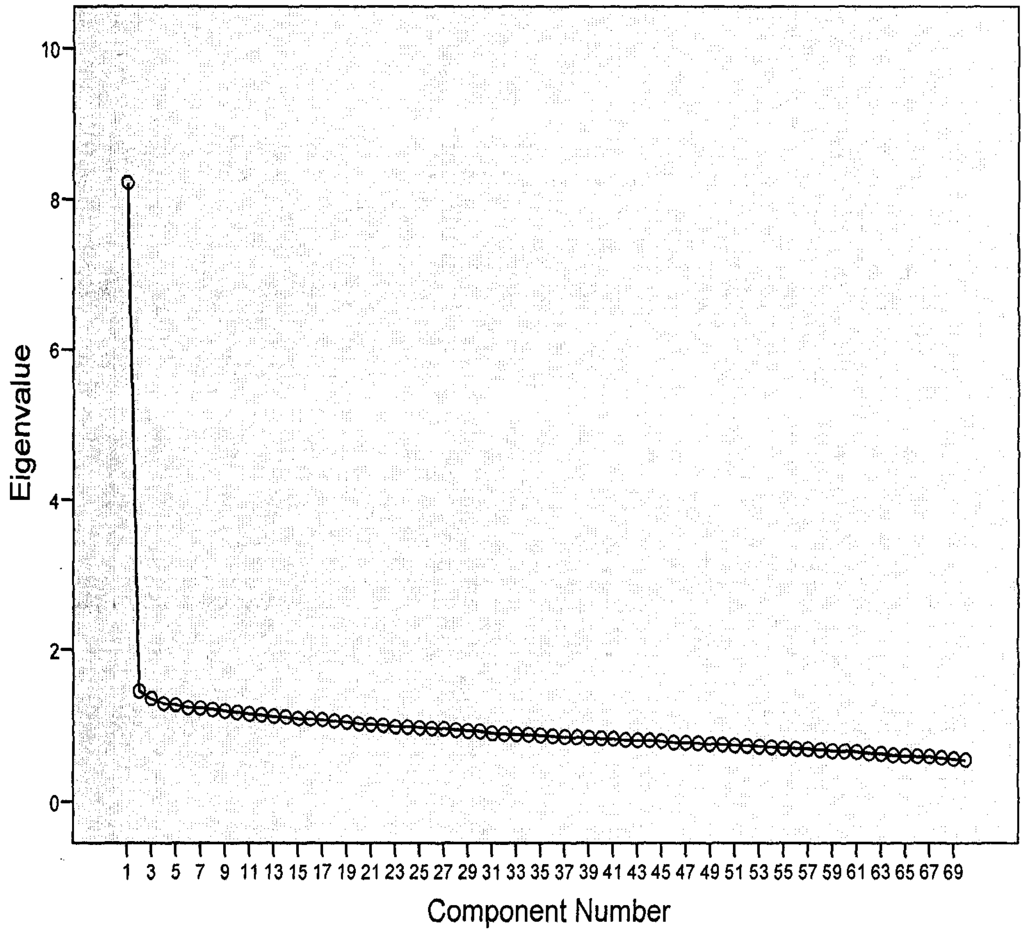

Figure C9 Scree Plot for SR English 30 Sample 1

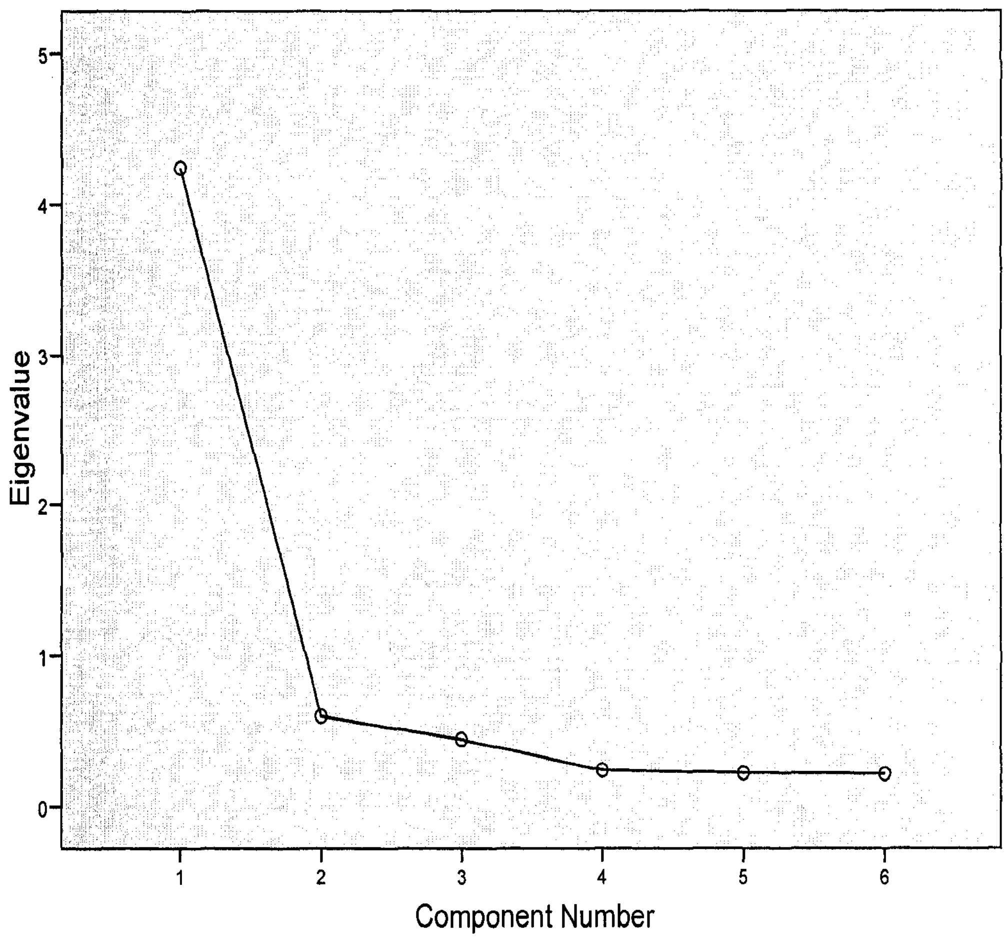

Figure C10 Scree Plot for CR English 30 Sample 1

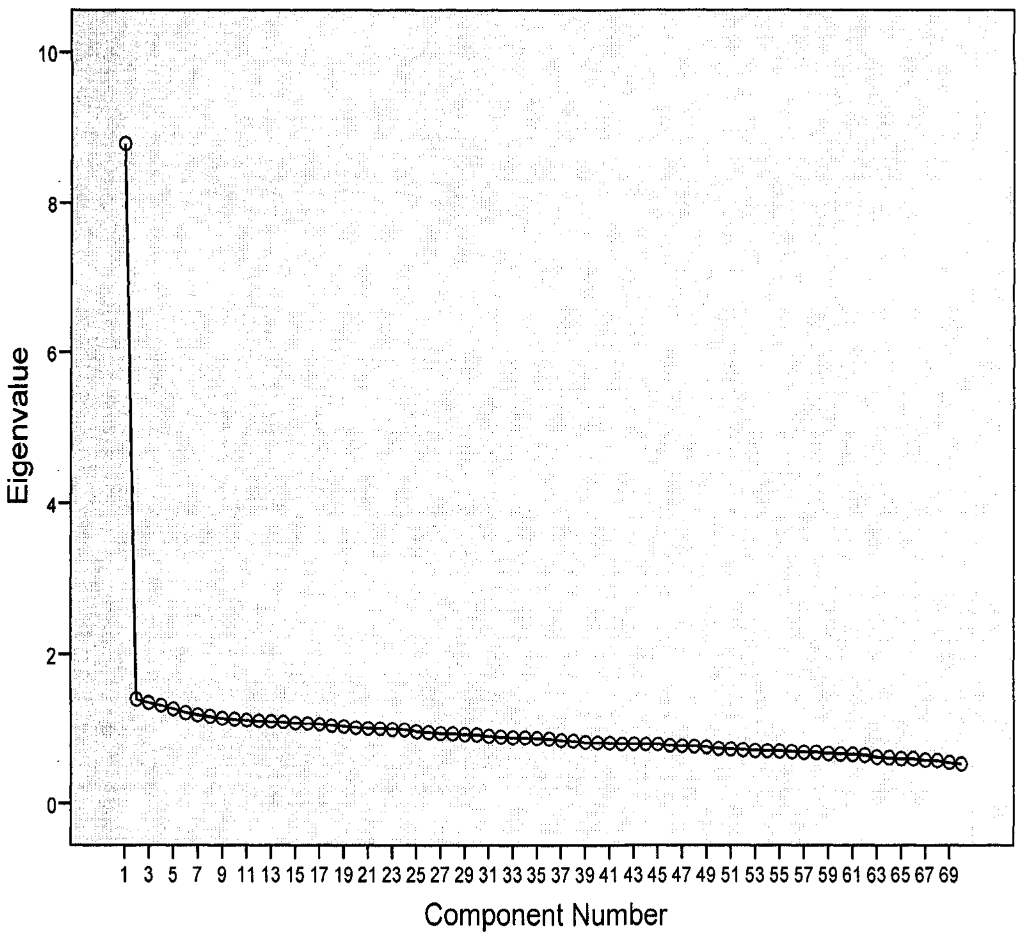

Figure C11 Scree Plot for SR English 30 Sample 2

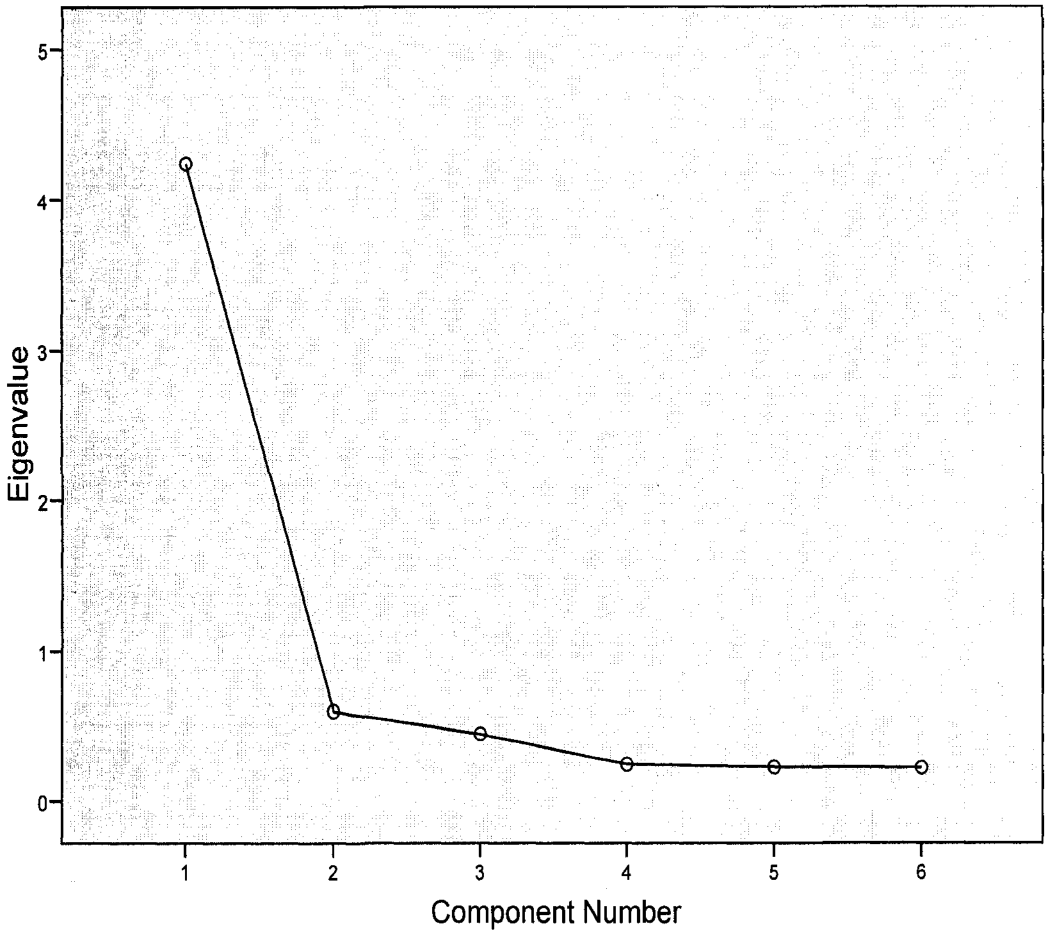

Figure C12 Scree Plot for CR English 30 Sample 2

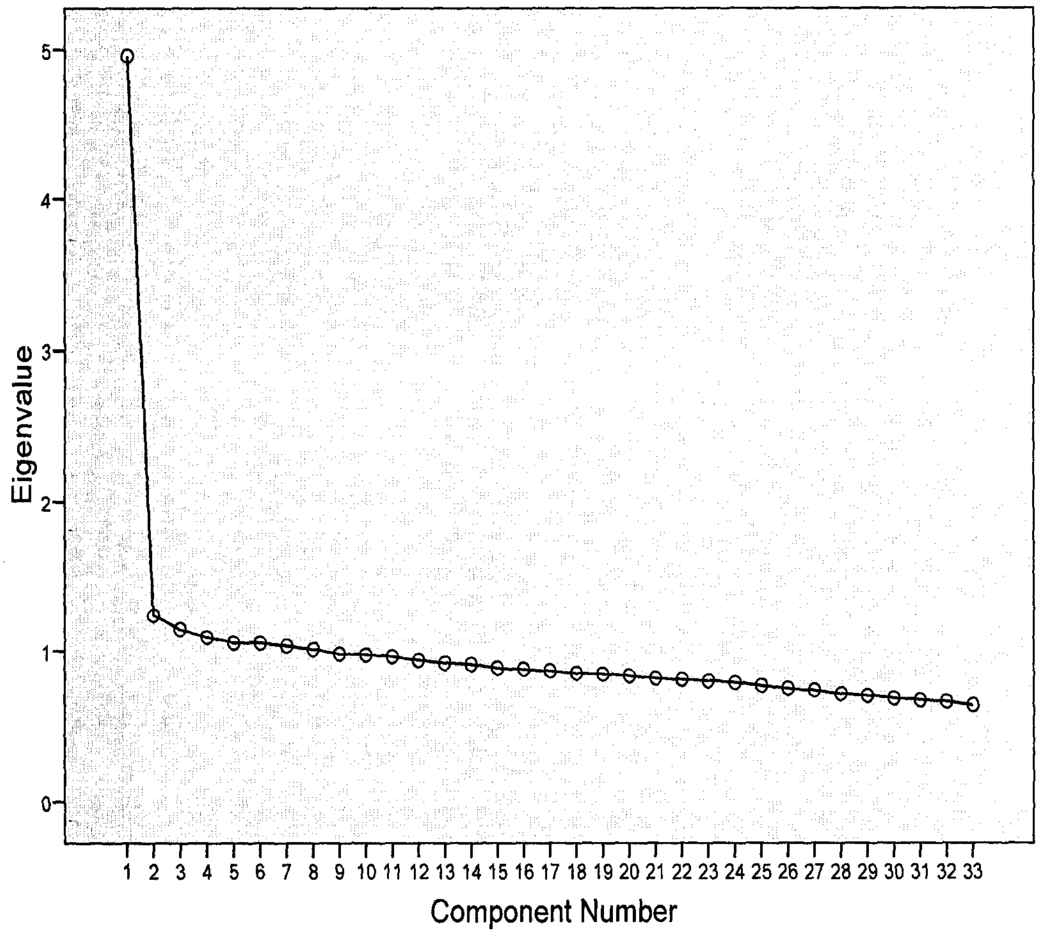

Figure C13 Scree Plot for SR Pure Math 30 Sample 1

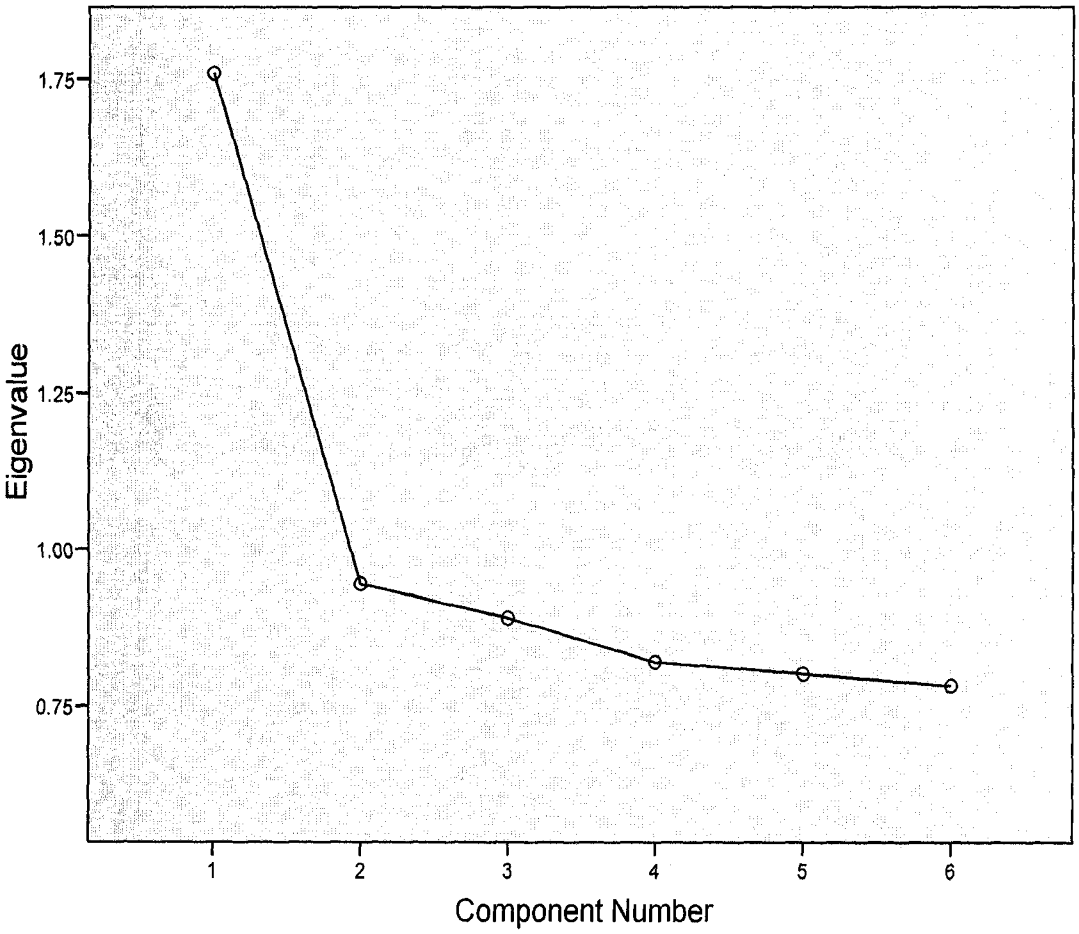

Figure C14 Scree Plot for CR (NR) Pure Math 30 Sample 1

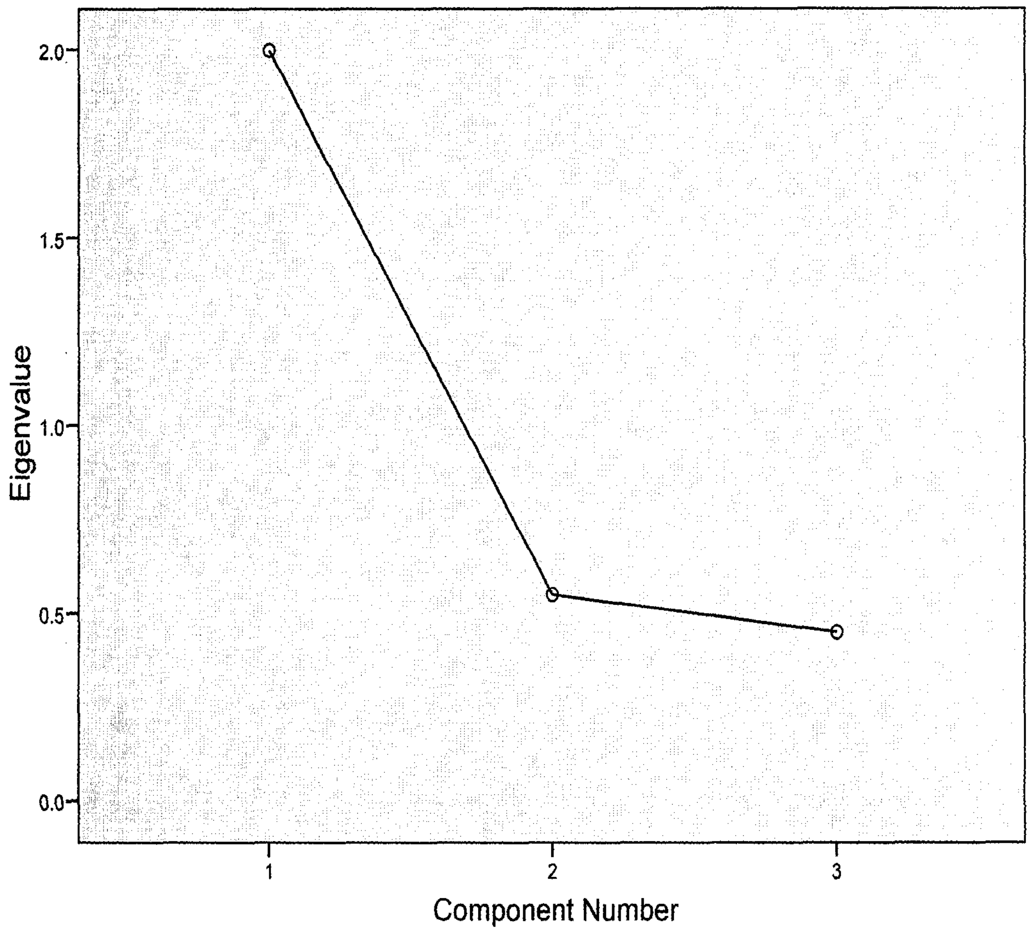

Figure C15 Scree Plot for CR (OE) Pure Math 30 Sample 1

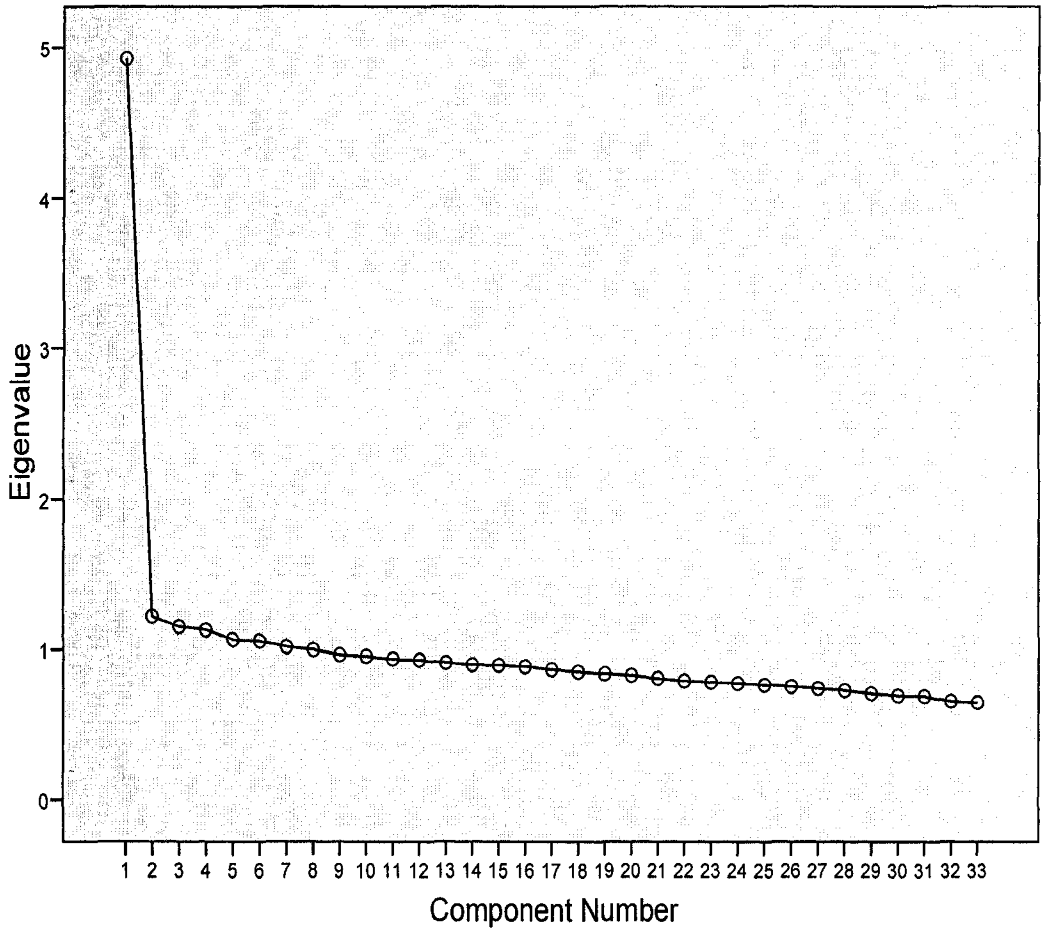

Figure C16 Scree Plot for SR Pure Math 30 Sample 2

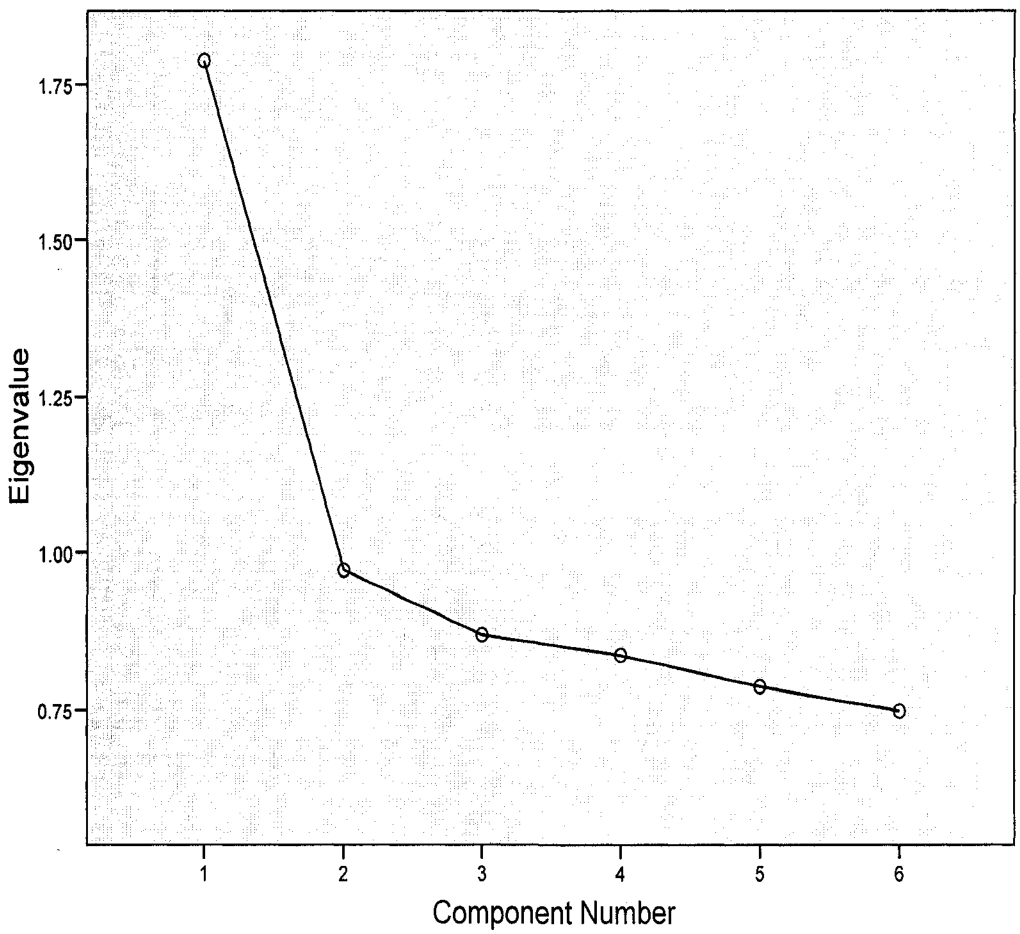

Figure C17 Scree Plot for CR (NR) Pure Math 30 Sample 2

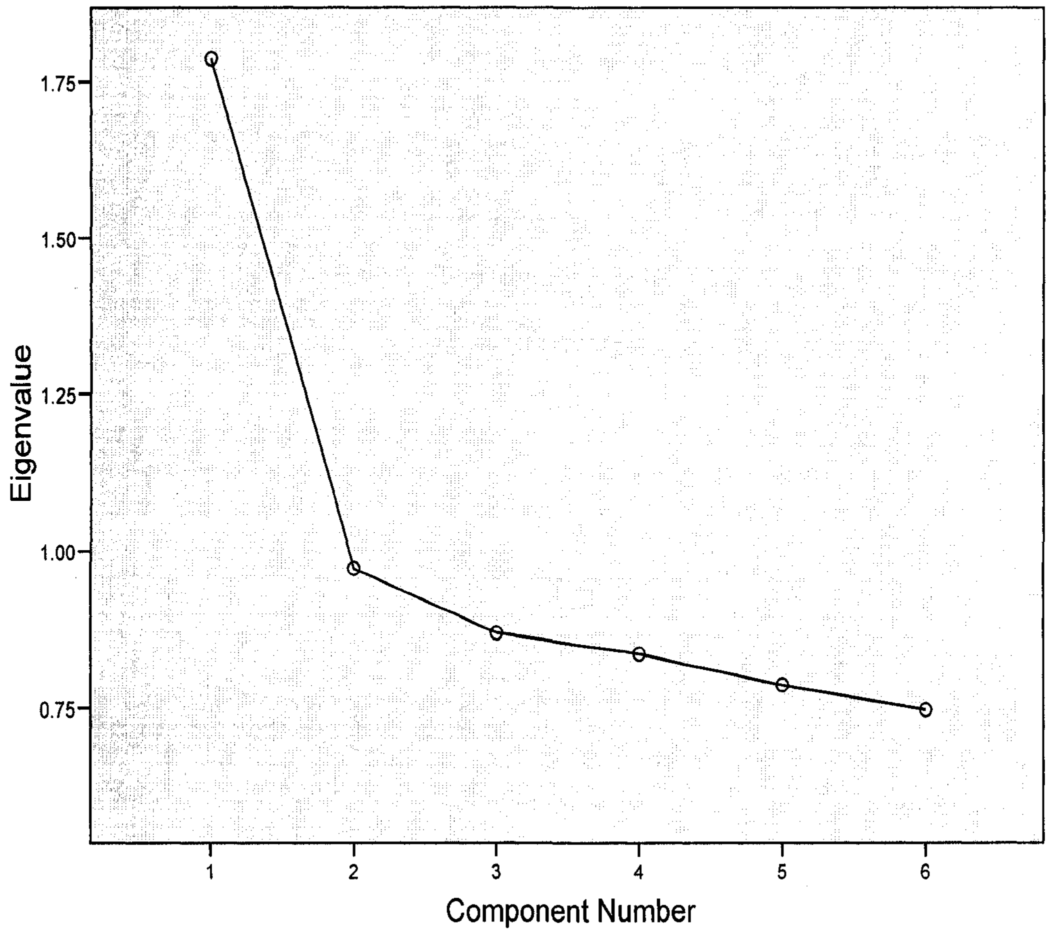

Figure C18 Scree Plot for CR (OE) Pure Math 30 Sample 2

APPENDIX D

Repeated Measures

Low-Stake Examinations

Table D1

Repeated Measures Multivariate and Analysis of Variance English 9 Sample 1

Multivariate Tests(c)

Effect	Value	F	Hypothesis df	Error df	Sig.	Noncent. Parameter	Observed Power(a)
Pillai's Trace	.260	233.707(b)	3.000	1997.000	.000	701.121	1.000
Wilks' Lambda	.740	233.707(b)	3.000	1997.000	.000	701.121	1.000
Hotelling's Trace	.351	233.707(b)	3.000	1997.000	.000	701.121	1.000
Roy's Largest Root	.351	233.707(b)	3.000	1997.000	.000	701.121	1.000

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power(a)
method	Sphericity Assumed	2817.426	3	939.142	16.564	.000	49.693	1.000
	Greenhouse-Geisser	2817.426	1.027	2743.363	16.564	.000	17.012	.998
	Huynh-Feldt	2817.426	1.027	2743.253	16.564	.000	17.012	.998
	Lower-bound	2817.426	1.000	2817.426	16.564	.000	16.564	.997

Note: Computed using alpha = 0.20

Table D2

Repeated Measures Multivariate and Analysis of Variance English 9 Sample 2

Multivariate Tests(c)

Effect	Value	F	Hypothesis df	Error df	Sig.	Noncent. Parameter	Observed Power(a)
Pillai's Trace	.246	217.382(b)	3.000	1997.000	.000	652.146	1.000
Wilks' Lambda	.754	217.382(b)	3.000	1997.000	.000	652.146	1.000
Hotelling's Trace	.327	217.382(b)	3.000	1997.000	.000	652.146	1.000
Roy's Largest Root	.327	217.382(b)	3.000	1997.000	.000	652.146	1.000

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power(a)
method	Sphericity Assumed	2871.143	3	957.048	15.439	.000	46.316	1.000
	Greenhouse-Geisser	2871.143	1.026	2799.104	15.439	.000	15.836	.996
	Huynh-Feldt	2871.143	1.026	2798.998	15.439	.000	15.837	.996
	Lower-bound	2871.143	1.000	2871.143	15.439	.000	15.439	.996

Note: Computed using alpha = .20

Table D3

Repeated Measures Multivariate and Analysis of Variance Math 9 Sample 1

Multivariate Tests(c)

Effect	Value	F	Hypothesis df	Error df	Sig.	Noncent. Parameter	Observed Power(a)
Pillai's Trace	.092	67.729(b)	3.000	1997.000	.000	203.188	1.000
Wilks' Lambda	.908	67.729(b)	3.000	1997.000	.000	203.188	1.000
Hotelling's Trace	.102	67.729(b)	3.000	1997.000	.000	203.188	1.000
Roy's Largest Root	.102	67.729(b)	3.000	1997.000	.000	203.188	1.000

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power(a)
method	Sphericity Assumed	22132.175	3	7377.392	124.571	.000	373.712	1.000
	Greenhouse-Geisser	22132.175	1.754	12620.871	124.571	.000	218.449	1.000
	Huynh-Feldt	22132.175	1.755	12610.686	124.571	.000	218.626	1.000
	Lower-bound	22132.175	1.000	22132.175	124.571	.000	124.571	1.000

Note: Computed using alpha = .20

Table D4

Repeated Measures Multivariate and Analysis of Variance Math 9 Sample 2

Multivariate Tests(c)

Effect	Value	F	Hypothesis df	Error df	Sig.	Noncent. Parameter	Observed Power(a)
Pillai's Trace	.112	83.671(b)	3.000	1997.000	.000	251.013	1.000
Wilks' Lambda	.888	83.671(b)	3.000	1997.000	.000	251.013	1.000
Hotelling's Trace	.126	83.671(b)	3.000	1997.000	.000	251.013	1.000
Roy's Largest Root	.126	83.671(b)	3.000	1997.000	.000	251.013	1.000

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power(a)
method	Sphericity Assumed	20435.598	3	6811.866	142.329	.000	426.988	1.000
	Greenhouse-Geisser	20435.598	1.880	10872.206	142.329	.000	267.525	1.000
	Huynh-Feldt	20435.598	1.881	10862.332	142.329	.000	267.768	1.000
	Lower-bound	20435.598	1.000	20435.598	142.329	.000	142.329	1.000

Note: Computed using alpha = .20

Table D5

Repeated Measures Multivariate and Analysis of Variance English 30 Sample 1

Multivariate Tests(c)

Effect	Value	F	Hypothesis df	Error df	Sig.	Noncent. Parameter	Observed Power(a)
Pillai's Trace	.538	773.616(b)	3.000	1997.000	.000	2320.848	1.000
Wilks' Lambda	.462	773.616(b)	3.000	1997.000	.000	2320.848	1.000
Hotelling's Trace	1.162	773.616(b)	3.000	1997.000	.000	2320.848	1.000
Roy's Largest Root	1.162	773.616(b)	3.000	1997.000	.000	2320.848	1.000

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power(a)
method	Sphericity Assumed	13823.731	3	4607.910	127.546	.000	382.637	1.000
	Greenhouse-Geisser	13823.731	1.074	12877.207	127.546	.000	136.921	1.000
	Huynh-Feldt	13823.731	1.074	12875.851	127.546	.000	136.935	1.000
	Lower-bound	13823.731	1.000	13823.731	127.546	.000	127.546	1.000

Computed using alpha = .20

Table D6

Repeated Measures Multivariate and Analysis of Variance English 30 Sample 2

Multivariate Tests(c)

Effect	Value	F	Hypothesis df	Error df	Sig.	Noncent. Parameter	Observed Power(a)
Pillai's Trace	.510	692.434	3.000	1997.000	.000	2077.302	1.000
Wilks' Lambda	.490	692.434	3.000	1997.000	.000	2077.302	1.000
Hotelling's Trace	1.040	692.434	3.000	1997.000	.000	2077.302	1.000
Roy's Largest Root	1.040	692.434	3.000	1997.000	.000	2077.302	1.000

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power(a)
method	Sphericity Assumed	11955.722	3	3985.241	136.332	.000	408.996	1.000
	Greenhouse-Geisser	11955.722	1.081	11062.293	136.332	.000	147.343	1.000
	Huynh-Feldt	11955.722	1.081	11061.018	136.332	.000	147.360	1.000
	Lower-bound	11955.722	1.000	11955.722	136.332	.000	136.332	1.000

Note: Computed using alpha = .20

Table D7

Repeated Measures Multivariate and Analysis of Variance Pure Math 30 Sample 1

Multivariate Tests(c)

Effect	Value	F	Hypothesis df	Error df	Sig.	Noncent. Parameter	Observed Power(a)
Pillai's Trace	.317	308.393	3.000	1997.000	.000	925.179	1.000
Wilks' Lambda	.683	308.393	3.000	1997.000	.000	925.179	1.000
Hotelling's Trace	.463	308.393	3.000	1997.000	.000	925.179	1.000
Roy's Largest Root	.463	308.393	3.000	1997.000	.000	925.179	1.000

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power(a)
method	Sphericity Assumed	1760.434	3	586.811	60.801	.000	182.402	1.000
	Greenhouse-Geisser	1760.434	1.186	1483.746	60.801	.000	72.139	1.000
	Huynh-Feldt	1760.434	1.187	1483.374	60.801	.000	72.157	1.000
	Lower-bound	1760.434	1.000	1760.434	60.801	.000	60.801	1.000

Note: Computed using alpha = .20

Table D8

Repeated Measures Multivariate and Analysis of Variance Pure Math 30 Sample 2

Multivariate Tests(c)

Effect	Value	F	Hypothesis df	Error df	Sig.	Noncent. Parameter	Observed Power(a)
Pillai's Trace	.297	281.439	3.000	1997.000	.000	844.317	1.000
Wilks' Lambda	.703	281.439	3.000	1997.000	.000	844.317	1.000
Hotelling's Trace	.423	281.439	3.000	1997.000	.000	844.317	1.000
Roy's Largest Root	.423	281.439	3.000	1997.000	.000	844.317	1.000

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power(a)
method	Sphericity Assumed	2230.355	3	743.452	86.375	.000	259.124	1.000
	Greenhouse-Geisser	2230.355	1.232	1810.170	86.375	.000	106.424	1.000
	Huynh-Feldt	2230.355	1.233	1809.618	86.375	.000	106.457	1.000
	Lower-bound	2230.355	1.000	2230.355	86.375	.000	86.375	1.000

Note: Computed using alpha = .20

APPENDIX E

Differences Between Scoring Procedures

Low-Stakes Examinations

Table E1

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 9 Sample 1

	UNW vs. WCRX2		UNW vs. WN/M		UNW vs. PTRN		WCRX2 vs. WN/M		WCRX2 vs. PTRN		WN/M vs. PTRN	
	F	CF	F	CF	F	CF	F	CF	F	CF	F	CF
-71												
-60									1	1		
-59					1	1				1	1	1
-58						1				1		1
-57						1				1		1
-54						1			1	2		1
-53						1				2	1	2
-52					1	2				2		2
-50						2				2		2
-49					1	3				2	1	3
-48					1	4			1	3		3
-47						4				3	1	4
-46						4				3		4
-45						4			1	4		4

Table E1 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 9 Sample 1

-44	1	5			2	6		4
-43	1	6			1	7	3	7
-42	2	8				7	1	8
-41	1	9			1	8	1	9
-40		9			2	10		9
-39	1	10			2	12	1	10
-38	2	12			1	13	2	12
-37	1	13			5	18	1	13
-36	4	17			2	20	4	17
-35	5	22			9	29	5	22
-34	4	26			9	38	8	30
-33	8	34			9	47	9	39
-32	10	44			9	56	7	46
-31	8	52			7	63	7	53
-30	6	58			4	67	8	61
-29	9	67			12	79	6	67
-28	6	73			10	89	10	77
-27	14	87			11	100	11	88
-26	6	93			12	112	11	99
-25	13	106			17	129	16	115
-24	15	121			19	148	14	129
-23	17	138			21	169	14	143
-22	22	160			19	188	22	165

Table E1 (continued)
Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 9 Sample 1

-21					17	177			17	205	22	187
-20					22	199			28	233	21	208
-19					23	222			32	265	24	232
-18					26	248			31	296	31	263
-17					30	278			20	316	26	289
-16					22	300			25	341	25	314
-15					31	331			36	377	27	341
-14					33	364			36	413	38	379
-13					44	408	4	4	58	471	44	423
-12					44	452		4	42	513	41	464
-11					37	489	3	7	39	552	39	503
-10					36	525		7	38	590	36	539
-9					44	569	4	11	50	640	52	591
-8	1	1			49	618	9	20	46	686	49	640
-7		1			44	662	11	31	50	736	44	684
-6		1			49	711	6	37	48	784	48	732
-5	2	3			60	771	15	52	56	840	63	795
-4	2	5			63	834	41	93	56	896	57	852
-3		5			60	894	19	112	42	938	57	909
-2	2	7	2	2	39	933	257	369	46	984	36	945
-1	125	132	5	7	45	978	765	1134	52	1036	41	986
0	684	816	1439	1446	52	1030	734	1868	53	1089	64	1050
1	590	1406	467	1913	56	1086	127	1995	58	1147	51	1101

Table E1 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 9 Sample 1

2	315	1721	56	1969	66	1152		1995	64	1211	65	1166
3	167	1888	20	1989	47	1199	3	1998	57	1268	48	1214
4		1888	7	1996	54	1253	1	1999	45	1313	70	1284
5	44	1932	4	2000	60	1313		1999	53	1366	48	1332
6	16	1948			64	1377	1	2000	54	1420	60	1392
7	15	1963			45	1422			46	1466	40	1432
8	6	1969			46	1468			43	1509	50	1482
9		1969			38	1506			36	1545	38	1520
10	11	1980			43	1549			36	1581	39	1559
11	9	1989			36	1585			39	1620	34	1593
12		1989			30	1615			41	1661	28	1621
13	4	1993			39	1654			25	1686	49	1670
14		1993			33	1687			33	1719	25	1695
15	3	1996			32	1719			30	1749	31	1726
16		1996			30	1749			24	1773	28	1754
17		1996			29	1778			33	1806	33	1787
18	4	2000			28	1806			31	1837	27	1814
19					26	1832			17	1854	27	1841
20					23	1855			22	1876	19	1860
21					17	1872			15	1891	16	1876
22					15	1887			10	1901	18	1894
23					16	1903			20	1921	15	1909
24					16	1919			16	1937	16	1925

Table E1 (continued)
Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 9 Sample 1

25	13	1932	5	1942	13	1938
26	12	1944	8	1950	7	1945
27	4	1948	8	1958	5	1950
28	10	1958	9	1967	9	1959
29	3	1961	3	1970	3	1962
30	4	1965	4	1974	6	1968
31	5	1970	2	1976	6	1974
32	5	1975	3	1979	3	1977
33	3	1978	2	1981	3	1980
34	2	1980	4	1985	1	1981
35	2	1982	1	1986	3	1984
36	2	1984		1986	1	1985
37	3	1987	2	1988	2	1987
38	4	1991	4	1992	4	1991
39		1991		1992		1991
40	1	1992	1	1993	2	1993
41		1992		1993		1993
42		1992		1993		1993
43	2	1994	1	1994	1	1994
44		1994	1	1995		1994
45		1994		1995	1	1995
46	1	1995		1995		1995
48		1995		1995		1995

Table E1 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 9 Sample 1

49	1	1996	2	1997	1	1996
50	1	1997	1	1998	1	1997
51		1997		1998	1	1998
52	1	1998		1998		1998
53		1998		1998		1998
54		1998		1998		1998
55		1998		1998		1998
61		1998		1998		1998
69		1998	1	1999		1998
70	1	1999		1999	1	1999
72		1999		1999		1999
74		1999		1999		1999
84		1999	1	2000		1999
85	1	2000			1	2000

Table E2

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 9 Sample2

	UNW vs. WCRX2		UNW vs. WN/M		UNW vs. PTRN		WCRX2 vs. WN/M		WCRX2 vs. PTRN		WN/M vs. PTRN	
	F	CF	F	CF	F	CF	F	CF	F	CF	F	CF
-72					1	1						
-71											1	1
-70									1	1		
-60												
-59												
-58												
-57									1	2		
-56					1	2					1	2
-54												
-53												
-52					1	3			1	3	1	3
-50												
-49												
-48					1	4			1	4	1	4
-47									1	5		
-46									1	6	1	5
-45					1	5			1	7		
-44									1	8	2	7
-43					2	7					1	8

Table E2 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 9 Sample2

-42	4	11	2	10	2	10
-41			3	13	2	12
-40	1	12	3	16	1	13
-39	2	14	4	20		
-38	4	18	4	24	6	19
-37	6	24	4	28	5	24
-36	9	33	5	33	7	31
-35	4	37	4	37	5	36
-34	2	39	11	48	3	39
-33	2	41	6	54	5	44
-32	6	47	5	59	7	51
-31	9	56	5	64	6	57
-30	6	62	11	75	9	66
-29	8	70	10	85	5	71
-28	9	79	19	104	15	86
-27	14	93	10	114	14	100
-26	18	111	19	133	18	118
-25	14	125	12	145	12	130
-24	16	141	17	162	17	147
-23	15	156	26	188	15	162
-22	14	170	18	206	17	179
-21	19	189	25	231	16	195
-20	21	210	25	256	24	219

Table E2 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 9 Sample2

-19					21	231	1	1	23	279	23	242
-18					20	251			26	305	26	268
-17					34	285			31	336	37	305
-16					35	320	1	2	42	378	31	336
-15					30	350		2	37	415	38	374
-14					47	397		2	35	450	43	417
-13					35	432		2	37	487	32	449
-12					39	471	2	4	38	525	40	489
-11					41	512			44	569	40	529
-10	1	1			50	562	7	11	39	608	46	575
-9					33	595	5	16	43	651	37	612
-8					49	644			49	700	47	659
-7					58	702			52	752	59	718
-6					38	740	19	35	47	799	46	764
-5	1	2			56	796			56	855	45	809
-4	3	5			48	844	60	95	51	906	57	866
-3	7	12	1	1	53	897	45	140	47	953	42	908
-2	9	21	1	2	50	947	243	383	46	999	58	966
-1	140	161	19	21	45	992	683	1066	48	1047	46	1012
0	710	871	1388	1409	55	1047	773	1839	47	1094	48	1060
1	450	1321	477	1886	45	1092	149	1988	42	1136	44	1104
2	447	1768	90	1976	41	1133	7	1995	53	1189	44	1148
3	92	1860	14	1990	45	1178	4	1999	48	1237	50	1198

Table E2 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 9 Sample2

4	26	1886	7	1997	52	1230			50	1287	51	1249
5	19	1905	2	1999	46	1276			54	1341	43	1292
6	60	1965	1	2000	52	1328			40	1381	58	1350
7	10	1975			57	1385	1	2000	43	1424	52	1402
8	9	1984			49	1434			55	1479	52	1454
9					54	1488			42	1521	43	1497
10					40	1528			41	1562	43	1540
11					46	1574			44	1606	39	1579
12	5	1989			30	1604			27	1633	40	1619
13					38	1642			33	1666	33	1652
14	7	1996			34	1676			38	1704	29	1681
15					34	1710			30	1734	33	1714
16					24	1734			31	1765	27	1741
17	2	1998			29	1763			25	1790	29	1770
18					24	1787			21	1811	26	1796
19					27	1814			26	1837	26	1822
20					26	1840			26	1863	26	1848
21	1	1999			22	1862			18	1881	20	1868
22	1	2000			14	1876			11	1892	13	1881
23					22	1898			21	1913	23	1904
24					12	1910			16	1929	14	1918
25					10	1920			5	1934	9	1927
26					10	1930			8	1942	8	1935

Table E2 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 9 Sample2

27		4	1934		4	1946	6	1941
28		8	1942		7	1953	5	1946
29		12	1954		8	1961	13	1959
30		5	1959		6	1967	2	1961
31		6	1965		5	1972	5	1966
32		8	1973		5	1977	8	1974
33		5	1978		3	1980	4	1978
34		2	1980		1	1981	3	1981
35		1	1981		2	1983	1	1982
36		3	1984		1	1984	2	1984
37		1	1985		3	1987	1	1985
38		2	1987		1	1988	2	1987
39		1	1988		1	1989	1	1988
40								
41		1	1989				1	1989
42								
43		1	1990		4	1993	1	1990
44		2	1992		1	1994	2	1992
45		2	1994			1994	2	1994
46		2	1996		2	1996	2	1996
48								
49								
50								

Table E2 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 9 Sample2

51	1	1997	2	1998	1	1997
52	1	1998			1	1998
53						
54						
55						
61						
69						
70						
72						
74						
78	1	1999	1	1999	1	1999
84						
85						
94			1	2000		
95					1	2000
96	1	2000				

Table E3

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Math 9 Sample 1

	UNW vs. WCRX2	UNW vs. WN/M		UNW vs. PTRN		WCRX2 vs. WN/M		WCRX2 vs. PTRN		WN/M vs. PTRN
-80		13	13							
-71										
-69		10	23			10	10			
-64		20	43							
-58				1	1					
-57				1	2					
-55		4	47			4	14			
-54										
-52				2	4	13	27	2	2	
-51				1	5					
-50										
-49				1	6					
-48				2	8					
-47				1	9					
-46										
-45		19	66							
-44				1	10					
-43				1	11					
-42				1	12					
-41										

Table E3 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Math 9 Sample 1

-40					1	13			1	3		
-39					1	14			1	4		
-38					1	15						
-37						15	20	47				
-36					1	16						
-35			1	67	1	17	1	48	1	5		
-34			21	88	1	18			1	6		
-33												
-32					1	19			1	7		
-31					1	20						
-30					3	23			1	8		
-29					2	25						
-28	13	13	18	106	1	26	19	67	1	9		
-27	20	33			1	27					1	1
-26					1	28						
-25					1	29			1	10		
-24												
-23			27	133	1	30			3	13		
-22					2	32	21	88	1	14		
-21					2	34			1	15	1	2
-20					5	39			3	18		
-19			33	166	3	42					2	4
-18					5	47	18	106	3	21	5	9

Table E3 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Math 9 Sample 1

-17	19	52			1	48			4	25	1	10
-16					5	53			4	29	3	13
-15			36	202	9	62	27	133				
-14					7	69			10	39	3	16
-13			29	231	8	77			8	47	6	22
-12	21	73	32	263	15	92	33	166	5	52	11	33
-11					10	102			14	66	13	46
-10	18	91	35	298	22	124	36	202	26	92	26	72
-9					32	156	29	231	20	112	26	98
-8	27	118	74	372	32	188	32	263	39	151	40	138
-7	33	151	35	407	57	245	35	298	57	208	55	193
-6					68	313	74	372	68	276	68	261
-5	36	187	40	447	79	392	35	407	96	372	93	354
-4	61	248	108	555	121	513	148	555	106	478	102	456
-3	35	283	113	668	129	642	113	668	132	610	122	578
-2	149	432	49	717	142	784	313	981	164	774	148	726
-1	270	702	264	981	174	958	248	1229	161	935	136	862
0	596	1298	180	1161	184	1142	228	1457	164	1099	159	1021
1	702	2000	133	1294	152	1294	387	1844	149	1248	115	1136
2			235	1529	136	1430	125	1969	137	1385	105	1241
3			315	1844	104	1534	31	2000	112	1497	107	1348
4			125	1969	101	1635			89	1586	95	1443
5			31	2000	91	1726			80	1666	67	1510

Table E3 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Math 9 Sample 1

6	54	1780	62	1728	69	1579
7	37	1817	38	1766	43	1622
8	32	1849	30	1796	35	1657
9	28	1877	26	1822	25	1682
10	18	1895	32	1854	29	1711
11	21	1916	9	1863	27	1738
12	7	1923	17	1880	21	1759
13	9	1932	9	1889	11	1770
14	5	1937	12	1901	13	1783
15	4	1941	8	1909	13	1796
16	1	1942	7	1916	11	1807
17	3	1945	3	1919	10	1817
18	5	1950	5	1924	15	1832
19	1	1951	3	1927	9	1841
20	6	1957	1	1928	13	1854
21	5	1962			8	1862
22	4	1966	5	1933	8	1870
23			3	1936	3	1873
24	3	1969	6	1942	8	1881
25	9	1978	6	1948	6	1887
26			5	1953	1	1888
27			4	1957	3	1891
28		1978	4	1961	5	1896

Table E3 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Math 9 Sample 1

29	1	1979	2	1963	3	1899
30	1	1980	1	1964	2	1901
31					3	1904
32			2	1966	7	1911
33					2	1913
34	1	1981			1	1914
35	1	1982	1	1967	3	1917
36			1	1968	2	1919
37	1	1983	1	1969	3	1922
38			1	1970		
39	1	1984			1	1923
40	1	1985			2	1925
41			2	1972	5	1930
42			2	1974	2	1932
43	1	1986			1	1933
44					2	1935
45					2	1937
46					1	1938
47			2	1976	2	1940
48	3	1989	1	1977		
49			2	1979	1	1941
50	1	1990				
51					2	1943

Table E3 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Math 9 Sample 1

52	3	1993	7	1986	5	1948
53			1	1987	4	1952
55	1	1994			3	1955
57					1	1956
58					1	1957
60			1	1988	3	1960
61					1	1961
62			1	1989	1	1962
63	1	1995	1	1990	1	1963
64					1	1964
65			1	1991		1964
69			3	1994	7	1971
70	1	1996				
71					1	1972
72						
73	1	1997	1	1995		
74						
75	1	1998				
76					2	1974
77					1	1975
78			1	1996	1	1976
80					3	1979
82			1	1997	1	1980

Table E3 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Math 9 Sample 1

83	1	1999				
84					2	1982
85			1	1998		
86					2	1984
88			1	1999	1	1985
89					6	1991
91					1	1992
93					1	1993
97					3	1996
98					1	1997
103					1	1998
104					1	1999
112	1	2000				
117			1	2000		
127					1	2000

Table E4

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Math 9 Sample 2

	UNW vs. WCRX2	UNW vs. WN/M		UNW vs. PTRN		WCRX2 vs. WN/M		WCRX2 vs. PTRN	WN/M vs. PTRN
-80			0						
-74		15	15						
-73				1	1				
-71			15	1	2				
-69			15		2				
-66		12	27		2				
-65				1	3				
-64			27		3				
-62				1	4				
-60					4	12	12		
-59			27	1	5		12		
-58			27	1	6		12		
-57			27	1	7		12		
-56				1	8		12		
-55			27		8		12		
-54			27		8		12		
-53		5	32		8	5	17		
-52			32	1	9		17	0	
-51		21	53		9		17	0	

Table E4 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Math 9 Sample 2

-50				53		9		17	1	1
-49				53		9		17		1
-48				53		9		17		1
-47				53		9		17		1
-46				53	2	11		17		1
-45				53		11		17	1	2
-44				53		11		17		2
-43				53		11		17		2
-42				53		11	15	32	1	3
-41				53		11		32	1	4
-40				53		11		32		4
-39				53		11		32	1	5
-38			13	66	1	12		32	1	6
-37				66	1	13		32		6
-36			6	72	1	14	6	38		6
-35				72	1	15		38		6
-34				72	2	17		38		6
-33				72		17		38		6
-32	15	15		72		17		38	1	7
-31		15	16	88	1	18	21	59		7
-30		15		88	3	21		59		7
-29		15		88		21		59		7
-28		15		88	1	22		59		7

Table E4 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Math 9 Sample 2

-27		15		88	1	23		59	1	8		
-26		15		88	4	27		59	3	11		
-25		15		88	2	29		59	1	12		
-24		15	29	117	2	31	13	72	2	14	2	2
-23		15		117	1	32		72		14	2	4
-22		15		117		32		72	3	17		4
-21		15		117	4	36		72	2	19	1	5
-20	21	36	27	144	6	42	16	88	1	20	2	7
-19		36		144	2	44		88	1	21	1	8
-18		36		144	6	50		88	2	23		8
-17		36	24	168	7	57		88	2	25	1	9
-16		36		168	2	59	29	117	1	26	3	12
-15		36		168	6	65		117	5	31	1	13
-14	13	49	36	204	7	72	2	119	5	36	3	16
-13		49	26	230	14	86	27	146	5	41	5	21
-12		49		230	7	93	24	170	8	49	13	34
-11	16	65		230	12	105		170	14	63	11	45
-10		65	40	270	29	134	34	204	29	92	14	59
-9		65	52	322	26	160	26	230	23	115	35	94
-8	29	94		322	49	209		230	39	154	43	137
-7	27	121	77	399	69	278	92	322	61	215	44	181
-6	12	133	40	439	81	359		322	80	295	62	243
-5	24	157	67	506	96	455	77	399	96	391	76	319

Table E4 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Math 9 Sample 2

-4	60	217	46	552	124	579	40	439	114	505	125	444
-3	40	257	106	658	106	685	113	552	122	627	105	549
-2	236	493	112	770	145	830	155	707	138	765	129	678
-1	288	781	197	967	147	977	187	894	133	898	136	814
0	894	1675	233	1200	181	1158	306	1200	197	1095	147	961
1	325	2000	146	1346	120	1278	294	1494	140	1235	152	1113
2			477	1823	131	1409	506	2000	145	1380	136	1249
3			177	2000	126	1535		2000	101	1481	100	1349
4				2000	103	1638			86	1567	89	1438
5				2000	75	1713			84	1651	74	1512
6					52	1765			67	1718	64	1576
7					46	1811			48	1766	42	1618
8					40	1851			37	1803	43	1661
9					23	1874			23	1826	27	1688
10					11	1885			22	1848	27	1715
11					23	1908			14	1862	19	1734
12					18	1926			19	1881	23	1757
13					6	1932			9	1890	16	1773
14					7	1939			13	1903	17	1790
15					10	1949			8	1911	11	1801
16					3	1952			6	1917	12	1813
17					5	1957			8	1925	10	1823
18					4	1961			6	1931	10	1833

Table E4 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Math 9 Sample 2

19	2	1963	9	1940	12	1845
20		1963	2	1942	6	1851
21	4	1967	4	1946	11	1862
22	2	1969	4	1950	9	1871
23	1	1970	5	1955	2	1873
24		1970	2	1957	7	1880
25	1	1971	3	1960	3	1883
26	5	1976	1	1961	8	1891
27		1976		1961	5	1896
28	1	1977		1961	5	1901
29	2	1979	3	1964	7	1908
30	1	1980	2	1966	1	1909
31	1	1981	2	1968	4	1913
32	3	1984	1	1969	5	1918
33		1984	1	1970		1918
34		1984	1	1971	2	1920
35	3	1987	6	1977	4	1924
36	2	1989	2	1979	8	1932
37	2	1991	2	1981	1	1933
38	1	1992		1981	5	1938
39	1	1993	1	1982	2	1940
40		1993	2	1984	2	1942
41		1993		1984		1942

Table E4 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Math 9 Sample 2

42	1	1994		1984	3	1945
43		1994		1984	2	1947
44		1994	2	1986		1947
45	2	1996	2	1988	3	1950
46		1996	1	1989	1	1951
47		1996	1	1990	2	1953
48		1996	1	1991	1	1954
49	1	1997		1991	5	1959
50		1997	1	1992	2	1961
51		1997		1992	1	1962
52	1	1998		1992	1	1963
53		1998	1	1993	6	1969
55		1998	3	1996	3	1972
57		1998	1	1997	1	1973
59		1998		1997	1	1974
60		1998		1997	2	1976
61		1998		1997	2	1978
62		1998		1997	1	1979
63		1998	1	1998	1	1980
64		1998		1998	1	1981
66		1998		1998	5	1986
69		1998		1998	1	1987
70		1998		1998		1987

Table E4 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Math 9 Sample 2

71	1	1999		1998		1987
72		1999		1998		1987
73		1999		1998	1	1988
74		1999		1998		1988
75		1999		1998		1988
76		1999		1998		1988
77		1999		1998	5	1993
78		1999		1998		1993
79		1999		1998		1993
80		1999		1998		1993
81		1999		1998	1	1994
82		1999	1	1999		1994
83		1999		1999	1	1995
86		1999		1999	3	1998
94	1	2000		1999		1998
101			1	2000		1998
102					1	1999
112						1999
114					1	2000

Table E5

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 30 Sample 1

	Sample 1											
	UNW vs. WCRX2		UNW vs. WN/M		UNW vs. PTRN		WCRX2 vs. WN/M		WCRX2 vs. PTRN		WN/M vs. PTRN	
	F	CF	F	CF	F	CF	F	CF	F	CF	F	CF
-46											1	1
-45												1
-44									1	1		1
-40											1	2
-39											2	4
-38					1	1			2	3	1	5
-37						1			1	4	2	7
-36					2	3			3	7	1	8
-35					1	4			3	10	2	10
-34					2	6			3	13	3	13
-33					4	10			1	14	1	14
-32					4	14			2	16	3	17
-31					2	16			4	20	3	20
-30					2	18			2	22	3	23
-29					3	21			2	24	6	29
-28					6	27			7	31	7	36
-27					3	30			11	42	10	46

Table E5 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 30 Sample 1

-26					7	37	4	46	6	52
-24					9	46	18	64	21	73
-25					5	51	10	74	10	83
-23					10	61	14	88	14	97
-22					9	70	17	105	18	115
-21					18	88	18	123	21	136
-20					16	104	28	151	28	164
-19					16	120	25	176	30	194
-18					27	147	35	211	30	224
-17					22	169	32	243	36	260
-16					28	197	24	267	32	292
-15					27	224	48	315	40	332
-14					30	254	33	348	42	374
-13			1	1	45	299	51	399	47	421
-12					42	341	44	443	48	469
-11	1	1	1	2	38	379	48	491	48	517
-10					38	417	61	552	59	576
-9	1	2			50	467	44	596	43	619
-8					75	542	53	649	54	673
-7			4	6	36	578	75	724	77	750
-6	4	6			67	645	61	785	60	810
-5			4	10	67	712	70	855	76	886
-4	6	12	2	12	60	772	69	924	68	954

Table E5 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 30 Sample 1

-3	6	18	6	18	71	843			67	991	65	1019
-2	11	29	21	39	67	910	2	2	74	1065	76	1095
-1	15	44	5	44	68	978	18	20	79	1144	74	1169
0	23	67	15	59	77	1055	1258	1278	66	1210	58	1227
1	661	728	444	503	70	1125	605	1883	59	1269	55	1282
2	800	1528	865	1368	81	1206	98	1981	63	1332	72	1354
3	188	1716	303	1671	77	1283	15	1996	64	1396	71	1425
4	139	1855	45	1716	60	1343	3	1999	70	1466	67	1492
5	28	1883	139	1855	55	1398	1	2000	60	1526	55	1547
6	29	1912	28	1883	65	1463			52	1578	51	1598
7	35	1947			54	1517			51	1629	53	1651
8			29	1912	49	1566			48	1677	42	1693
9	16	1963	35	1947	50	1616			37	1714	34	1727
10	8	1971			50	1666			34	1748	30	1757
11	10	1981	16	1963	46	1712			35	1783	36	1793
12	10	1991	8	1971	38	1750			35	1818	37	1830
13	5	1996	10	1981	38	1788			35	1853	32	1862
14					27	1815			22	1875	18	1880
15	3	1999	10	1991	28	1843			27	1902	28	1908
16			5	1996	20	1863			10	1912	13	1921
17					24	1887			12	1924	10	1931
18	1	2000			17	1904			12	1936	10	1941
19			3	1999	13	1917			9	1945	6	1947

Table E5 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 30 Sample 1

20			10	1927	13	1958	16	1963
21			13	1940	8	1966	6	1969
22			11	1951	5	1971	3	1972
23	1	2000	7	1958	2	1973	2	1974
24			7	1965	5	1978	5	1979
25			4	1969	4	1982	4	1983
26			5	1974	1	1983	1	1984
27			6	1980	1	1984	1	1985
28			1	1981	1	1985	1	1986
29					2	1987	1	1987
30			2	1983				
31					1	1988	1	1988
32			1	1984	1	1989	2	1990
33			2	1986	1	1990	1	1991
34			2	1988	1	1991		
35			1	1989			1	1992
36					1	1992		
37			1	1990	1	1993	1	1993
38			2	1992	3	1996	3	1996
39								
40			2	1994	1	1997	1	1997
41			2	1996				
42			1	1997				

Table E5 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 30 Sample 1

49			1	1998	1	1998
50						
51	1	1998	1	1999	1	1999
52						
53	1	1999				
72			1	2000	1	2000
74	1	2000				

Table E6

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 30 Sample 2

	UNW vs. WCRX2		UNW vs. WN/M		UNW vs. PTRN		WCRX2 vs. WN/M		WCRX2 vs. PTRN		WN/M vs. PTRN	
	F	CF	F	CF	F	CF	F	CF	F	CF	F	CF
-49											1	1
-48									1	1		
-45					1	1						
-41											1	2
-40									1	2	1	3
-39											1	4
-38					1	2			2	4		
-37					1	3					1	5
-36					1	4						
-34									1	5	2	7
-33									3	8	1	8
-32					2	6					2	10
-31					1	7			1	9	1	11
-30					1	8			4	13	5	16
-29									2	15	5	21
-28					3	11			8	23	5	26
-27					4	15			5	28	9	35
-26					4	19			10	38	8	43
-25					8	27			10	48	12	55
-24					9	36			12	60	11	66

Table E6 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 30 Sample 2

-23					6	42			8	68	8	74
-22					9	51			10	78	12	86
-21					12	63			18	96	18	104
-20					8	71			13	109	14	118
-19					13	84			21	130	22	140
-18					17	101			20	150	27	167
-17					23	124			33	183	30	197
-16					22	146			27	210	29	226
-15					27	173			30	240	28	254
-14					24	197			36	276	46	300
-13			2	2	31	228			50	326	43	343
-12	2	2		2	44	272			42	368	41	384
-11		2	1	3	41	313			50	418	63	447
-10	1	3		3	46	359			60	478	54	501
-9		3		3	53	412			58	536	57	558
-8		3	2	5	65	477			73	609	67	625
-7	2	5		5	70	547			67	676	78	703
-6		5		5	61	608			75	751	68	771
-5	2	7	2	7	73	681			65	816	71	842
-4	1	8	6	13	69	750			79	895	75	917
-3	5	13	1	14	83	833			70	965	64	981
-2	5	18	4	18	68	901			83	1048	85	1066
-1	12	30	12	30	74	975	11	11	83	1131	90	1156

Table E6 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 30 Sample 2

	UNW vs. WCRX2		UNW vs. WN/M		UNW vs. PTRN		WCRX2 vs. WN/M		WCRX2 vs. PTRN		WN/M vs. PTRN	
	F	CF	F	CF	F	CF	F	CF	F	CF	F	CF
0	53	83	2	32	85	1060	1394	1405	77	1208	67	1223
1	898	981	903	935	82	1142	547	1952	67	1275	70	1293
2	519	1500	347	1282	77	1219	27	1979	87	1362	89	1382
3	241	1741	351	1633	80	1299	16	1995	74	1436	71	1453
4	145	1886	195	1828	80	1379	5	2000	79	1515	87	1540
5	28	1914	58	1886	70	1449			57	1572	50	1590
6	19	1933	28	1914	64	1513			77	1649	70	1660
7	19	1952	19	1933	76	1589			44	1693	46	1706
8	12	1964	19	1952	39	1628			57	1750	56	1762
9					67	1695			46	1796	44	1806
10	15	1979	12	1964	50	1745			33	1829	29	1835
11	10	1989			53	1798			26	1855	24	1859
12			15	1979	26	1824			22	1877	28	1887
13	5	1994			31	1855			22	1899	16	1903
14			10	1989	19	1874			16	1915	13	1916
15		1994			18	1892			8	1923	9	1925
16	6	2000			18	1910			11	1934	12	1937
17			5	1994	14	1924			17	1951	17	1954
18					14	1938			9	1960	9	1963
19			6	2000	11	1949			6	1966	4	1967

Table E6 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: English 30 Sample 2

20	9	1958	6	1972	6	1973
21	8	1966	4	1976	6	1979
22	6	1972	7	1983	4	1983
23	5	1977				
24	2	1979	3	1986	4	1987
25	4	1983	4	1990	3	1990
26	4	1987	2	1992	1	1991
27	2	1989	1	1993	2	1993
28	1	1990				
29	1	1991				
30	2	1993	1	1994	1	1994
31	1	1994	2	1996	3	1997
32	1	1995				
33	1	1996				
34			1	1997		
50	1	1997			1	1998
51			1	1998		
53	1	1998				
74			1	1999	1	1999
75	1	1999				
76			1	2000	1	2000
78	1	2000				

Table E7

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Pure Math 30 Sample 1

	UNW vs. WCRX2	UNW vs. WN/M	UNW vs. PTRN	WCRX2 vs. WN/M	WCRX2 vs. PTRN	WN/M vs. PTRN
-71						
-60						
-59						
-58						
-57						
-54						
-53		1	1			
-52						
-50		1	2			
-49						
-48						
-47						
-46						
-45						
-44						
-43						
-42						
-41						
-40						

Table E7 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Pure Math 30 Sample 1

-39												
-38												
-37												
-36											1	1
-35											1	2
-34												
-33					1	3						
-32												
-31					1	4					1	3
-30									1	1		
-29					2	6			1	2	1	4
-28					1	7						
-27					1	8			1	3	1	5
-26	3	3			3	11						
-25												
-24									2	5	2	7
-23					1	12					2	9
-22					4	16			1	6	1	10
-21					1	17			2	8	2	12
-20	4	7			5	22			1	9	2	14
-19					1	23					1	15
-18			3	3	1	24			1	10		
-17									2	12		

Table E7 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Pure Math 30 Sample 1

-16	1	8			8	32			3	15	5	20
-15					3	35			2	17	1	21
-14			4	7	5	40			7	24	4	25
-13	7	15		7	9	49			5	29	10	35
-12	4	19		7	20	69			8	37	10	45
-11			1	8	14	83			7	44	10	55
-10	7	26			18	101			16	60	10	65
-9	6	32	7	15	34	135			14	74	24	89
-8	6	38			43	178			36	110	38	127
-7			7	22	51	229			47	157	42	169
-6	34	72	6	28	85	314			59	216	59	228
-5	18	90	6	34	92	406			81	297	94	322
-4	67	157	52	86	107	513			108	405	108	430
-3	36	193	67	153	157	670			125	530	122	552
-2	195	388	129	282	163	833			147	677	158	710
-1	1069	1457	1024	1306	139	972	62	62	151	828	147	857
0	481	1938	632	1938	200	1172	1488	1550	177	1005	182	1039
1			62	2000	148	1320	378	1928	156	1161	155	1194
2	62	2000			133	1453	34	1962	149	1310	138	1332
3					121	1574	19	1981	136	1446	136	1468
4					105	1679	7	1988	120	1566	118	1586
5					78	1757	1	1989	102	1668	105	1691
6					58	1815	4	1993	76	1744	71	1762

Table E7 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Pure Math 30 Sample 1

7	57	1872			68	1812	64	1826
8	37	1909	3	1996	53	1865	52	1878
9	18	1927			34	1899	19	1897
10	10	1937	4	2000	18	1917	25	1922
11	14	1951			9	1926	12	1934
12	11	1962			13	1939	16	1950
13	8	1970			13	1952	13	1963
14	9	1979			13	1965	6	1969
15	4	1983			5	1970	5	1974
16	3	1986			7	1977	5	1979
17	3	1989			2	1979	2	1981
18	3	1992			5	1984	7	1988
19	1	1993			4	1988	2	1990
20	2	1995			2	1990	2	1992
21	1	1996			3	1993	3	1995
22	1	1997					3	1998
23	1	1998			2	1995		
24								
25								
26					2	1997		
27								
28	1	1999			1	1998		
29								

Table E7 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Pure Math 30 Sample 1

30						
31						
32					1	1999
33						
34			1	1999		
35						
36						
37						
38	1	2000				
39						
40						
41						
42						
43						
44						
45					1	2000
46						
48			1	2000		

Table E8

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Pure Math 30 Sample 2

	UNW vs. WCRX2	UNW vs. WN/M	UNW vs. PTRN	WCRX2 vs. WN/M	WCRX2 vs. PTRN	WN/M vs. PTRN
-71						
-60						
-59						
-58						
-57						
-54						
-53						
-52						
-50						
-49						
-48						
-47						
-46			1 1			1 1
-45						
-44					1 1	
-43						
-42						
-41						

Table E8 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Pure Math 30 Sample 2

-40												
-39												
-38					1	2			1	2	1	2
-37					1	3						
-36												
-35												
-34												
-33												
-32					1	4						
-31												
-30												
-29					1	5						
-28												
-27												
-26												
-25	3	3										
-24					1	6						
-23					1	7						
-22												
-21					4	11						
-20					3	14						

Table E8 (continued)
Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Pure Math 30 Sample 2

-19	7	10			1	15			1	3	2	4
-18			3	3	1	16			1	4	1	5
-17						16			1	5	2	7
-16					2	18			1	6	4	11
-15	6	16			8	26			6	12	4	15
-14					7	33			2	14	5	20
-13			5	8	4	37			2	16	1	21
-12	2	18		8	13	50			4	20	4	25
-11				8	10	60			10	30	18	43
-10	4	22	6	14	22	82			16	46	14	57
-9	8	30	2	16	25	107			13	59	15	72
-8			2	18	40	147			20	79	30	102
-7	28	58	4	22	70	217			33	112	45	147
-6			8	30	75	292			64	176	66	213
-5	30	88	28	58	83	375			81	257	100	313
-4	39	127			109	484			99	356	83	396
-3	102	229	45	103	141	625			99	455	123	519
-2	474	703	166	269	182	807			157	612	162	681
-1	953	1656	1115	1384	178	985	73	73	173	785	181	862
0	165	1821	437	1821	206	1191	994	1067	192	977	187	1049
1	106	1927	152	1973	150	1341	819	1886	166	1143	181	1230

Table E8 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Pure Math 30 Sample 2

2	46	1973	27	2000	141	1482	84	1970	173	1316	145	1375
3	27	2000			109	1591	12	1982	131	1447	126	1501
4					116	1707	2	1984	114	1561	100	1601
5					86	1793	6	1990	96	1657	118	1719
6					55	1848	5	1995	105	1762	87	1806
7					42	1890	3		65	1827	44	1850
8					27	1917			46	1873	43	1893
9					25	1942			29	1902	26	1919
10					8	1950	2	2000	21	1923	20	1939
11					10	1960			16	1939	8	1947
12					9	1969			10	1949	12	1959
13					8	1977			7	1956	6	1965
14					6	1983			7	1963	8	1973
15					5	1988			5	1968	6	1979
16					3	1991			7	1975	3	1982
17					3	1994			4	1979	4	1986
18									5	1984	6	1992
19									5	1989	2	1994
20					1	1995			2	1991		
21					2	1997			1	1992		
22											1	1995

Table E8 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Pure Math 30 Sample 2

23			1	1998				
24					2	1994	1	1996
25			1	1999	1	1995	1	1997
26					2	1997	1	1998
27							1	1999
28					2	1999		
29								
30								
31								
32								
33								
34								
35								
36								
37								
38								
39								
40								
41								
42								
43								

Table E8 (continued)

Differences Between Unweighted, Weighted CRx2, Weighted N/M, and Pattern Scores: Pure Math 30 Sample 2

44												
45												
46												
48												
49												
50												
51					1	2000						
52												
53												
54												
55												
59											1	2000
61												
63									1	2000		