

Running Head: Comparison of Three Weighting Procedures

A Comparison of Three Weighting Procedures for High- and Low-Stakes Examinations
with Mixed Item Formats in Different Subject Areas

W. Todd Rogers
Denise M. Nowicki
University of Alberta

Corresponding Author

W. Todd Rogers

Centre for Research in Applied Measurement and Evaluation

6-110 Education North

University of Alberta

Edmonton, Alberta

T6G 2G5

Phone: 780 492 3763

Fax: 780 492 0001

E-mail: todd.rogers@ualberta.ca

Abstract

The interchangeability of scores yielded by three weighting procedures applied to low-stakes achievement tests and to high-stakes examinations containing both selected response (SR) items and constructed response (CR) items in Language Arts and Mathematics was examined. The three scoring procedures included an unweighted procedure in which scores from the set of SR items and the set of CR items/tasks were added; a weighted procedure in which the CR items were weighted so that the CR and SR items contributed equally; and pattern scoring in which each item was individually weighted.

While the different weighting procedures yielded similar score distributions for all four tests at the group level, they were sufficiently dissimilar at the student level to warrant using them interchangeably. Pattern scoring provided the smallest standard errors, particularly at the lower end of the ability distribution. Whereas test stakes was not a factor, subject area may be a factor. Further, differences between the three score distributions suggest that care must be taken in choosing one weighting procedure over the others in a criterion-referenced situation, especially when a cut-score is set in the tail of the score distribution.

A Comparison of Three Procedures for Weighting Selected and Constructed Items on High- and Low-Stakes Examinations in Different Subject Areas

Most large-scale assessments today contain both selected response (SR) and constructed response (CR) items so as to obtain a more comprehensive assessment of the knowledge and skills possessed by students. An issue that arises with the inclusion of the two item types is how they should be represented or weighted in the total score (Haertel, 2006; Kolen, 2006). For example, a CR item, especially one that calls for an extended response, requires a longer response time than a single SR item. Consequently, given fixed administration times and the need to cover a broad domain, the number of CR items is less than the number of SR items in large-scale assessment instruments. Therefore, the number of score points derived from the CR items is generally less than the number of score points derived from the SR items. Without correcting for the unequal number of score points from each item type, the content and cognitive processes assessed by CR items may be de-emphasized in favor of the content and processes assessed by SR items. This is undesirable given the content and cognitive processes assessed by CR items are important attributes of the domain about which inferences are to be drawn.

Weighting procedures can be used to address this apparent imbalance. Sometimes the weights are explicitly determined by policy. For example, a policy may require that the scores obtained from the SR and CR items be equal in value. To achieve the desired weighting, it may be necessary to explicitly weight the SR and CR items differently. Other times the policy is such that the items are weighted implicitly as in pattern scoring under a particular IRT model. For example, the weights that result when the two- or three-parameter IRT model is used for the SR items and Murkaki's (1992) generalized

partial credit model is used for the CR items are optimal in the sense of maximizing test information and minimizing the standard error of measurement, regardless of item format.

Prior to the introduction of IRT models, judgmental procedures were employed to determine the weights to be applied. For example, after correcting for unequal standard deviations, the SR and CR components could be differentially weighted so that each contributed the same or a pre-determined and different number of points to the total score. Alternatively, if an external criterion was available, regression procedures could be used to obtain implicit weights to be applied to the subtest scores derived from the SR items and the subtest scores derived from the CR items that maximized the squared multiple correlation coefficient (Gulliksen, 1950).

While the implicit weights employed in IRT scoring are often acceptable, there may be occasions where explicit weights are required within an IRT framework. Weighting of the CR response component may be required so that the contribution of the constructed responses to the total score matches what is required. One option is to multiply the portion of the Test Characteristic Curve contributed by the item or subset of items to be weighted. For example, if time constraints prevent the administration of two extended response items, then the expected score for the single extended response item can be multiplied by two and then added to the expected item scores for the remaining items to obtain the expected total score. Another alternative, employed by the Education Quality and Accountability Office to score the Ontario Secondary School Literacy Test, involves adding together the scores of the trained raters who score each of the two long writing items. The initial score range was changed from 0 to 8 to 0 to 16, and these

“item” scores were input into the IRT analysis (EQAO, 2005). Thus, there are different options for explicitly weighting items in an IRT framework.

Schaeffer, Henderson-Montero, Julian, and Bene (2002) were among the first to examine the effect on test scores of different weighting systems within an IRT framework. They examined three different ways of combining SR and CR scores yielded by low-stakes Grade 9 Biology and English field tests that contained both item types: *unit weighting procedure* in which a student’s total score was equal to the number of points earned from the SR items plus the number of points earned from the CR items, where the number of points was set by the test developer; a *differential weighting* in which the SR items and CR items contributed the same number of points toward the total score; and *implicit weighting* in which the SR and CR items were implicitly weighted according to the IRT models used. Items within each content area were calibrated using the 3PL model for SR items and the generalized (Muraki, 1992) partial credit model for the CR items so that the items were placed on the same scale simultaneously. The computer program PARDUX, which uses marginal maximum likelihood procedures implemented with the EM algorithm (Bock & Aitken, 1981), was used for this calibration. The item parameters were placed on the score scale using WINFLUX (Burket, 1999). The scaling parameters were a multiplier of 50 and an additive constant of 500. The scale scores were rounded to the nearest integer value and the lowest (LOSS) and highest (HOSS) possible scale scores were set at 300 and 700, respectively.

Whereas the variances of the distributions, correlation patterns, and percentages of students in each of four hypothetical proficiency categories (English only) for the total group and the gender and ethnic subgroups considered were comparable, the scale score

means for Biology were consistently lower than the scale score means for English across the three weighting procedures (495.1, 496.6, and 493.1 vs. 500.9, 501.2, and 503.7 for the total group). Implicit weighting provided the smallest standard errors, particularly at the lower end of the scale score distributions.

Sykes and Hou (2003) also addressed the issue of different weighting procedures using a low-stakes Grade 8 field test for writing that contained both SR and CR items. In particular, they were interested in the influence of CR items and the way they are handled in a writing test. In addition to the three weighting procedures used by Schaeffer et al. (2002), they considered four additional 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 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. Item calibration was completed using PARDUX and WINFLUX as described above for Schaeffer et al. (2002).

Sykes and Hou (2003) found that the test characteristic curves (TCC) for the different weighting procedures were, with the exception of the all multiple-choice form, quite similar. The all multiple-choice form TCC was above the other TCCs for scale scores below 375 and above 525, and was steeper within the middle of the ability scale due to increased item discrimination in the middle of the midrange of scores. Implicit weighting provided the smallest standard errors across the ability range for all of the forms containing CR items, particularly for scale scores less than 400. Differences at the 10th and 90th percentiles between the operational form and each of the other forms varied,

respectively, between -13.0 and -34.0 and 13.5 and 32.0. The all multiple-choice form was found to have the highest test reliability ($\hat{\alpha} = 0.90$), with the two summed CR forms and the CRx2 form having slightly lower and equal reliabilities ($\hat{\alpha} = 0.84$).

Low-Stakes Tests versus High-Stakes Examinations

Large-scale assessments are often described as low-stakes and high-stakes. For example, the Primary, Junior and Grade 9 assessments in Ontario are considered to be low-stakes and the Ontario Secondary School Literacy Test is considered to be high-stakes. Whereas student level data are provided at the student level for all of these assessments, the student level data does not count toward a school mark at the Primary and Junior levels and between 0 and 30% at the Grade 9 level, with the decision as to how much made by individual teachers. In contrast, students must pass the OSSLT or take a special literacy course after failing the OSSLT at least once in order to satisfy graduation requirements for the province. Of concern is that students who perceive the results of an assessment as inconsequential to their personal achievement may, as a result, not work as hard to achieve their best as they would if they perceived the results to be consequential (Blau, Moller, & Jones, 2004; Brown & Walberg, 1993; DeMars, 2000; Ferrer-Caja & Weiss, 2002; Kiplinger & Linn, 1992; Paris, Lawton, & Turner, 1992; Sloane & Kelly, 2003; van Barneveld, 2003; Wise, 1996; Wolf, Smith, & Birnbaum, 1995; Wolf & Smith, 1995). The evidence collected in the studies conducted suggests that up to a quarter of the examinees display a lack of persistence in applying their abilities to answer all the items in a low stakes test and that this lack of effort results in underestimates of their abilities.

For example, Wolf et al. (1995) explored the consequences of performance on a low-stakes pilot test and a high-stakes operational 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 same test and the results were used as a major determinant of 11th grade placement into remedial programs. Although 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 were not analyzed.

Wolf et al. reported that the overall performance on the subtest of SR items was not significantly different between the two grade levels. 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 made this finding suspect. Anticipating this finding, Wolf et al. included an attitudinal scale which the students completed after the test. One question asked the students to indicate, 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 ($p < 0.001$).

In a second study, 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 pilot administration of the forms ($n = 3,596$) or the 1997 high-stakes operational administration, which was used for state diploma endorsement ($n = 8,334$). 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 scale scores were estimated for each student, one for the subset containing the SR items and one for the subtest containing the CR items. The SR and CR item difficulties were simultaneously estimated with the 1-PL model and the one parameter partial-credit model, and then the items were divided into the two subscales. Item parameter estimates were obtained using PARSCALE 3 (Muraki & Bock, 1997). After the item difficulties were estimated, Bayesian Expected a Posterior (EAP) scores were estimated for each person on the two subscales, with a normal prior. A hierarchical linear model HLM4 (Bryk, Raudenbush, & Congdon, 1996) that included both a student level and school level was used. In both math and science, students scored significantly higher on the high-stakes operational forms than on the low-stakes pilot forms ($p < 0.001$).

Both Schaeffer et al. (2002) and Sykes and Hou (2003) worked with field test forms, the marks from which did not count toward a school grade. Hence, the authors considered these tests to be low stakes. However, it was not made clear how motivated the students who responded to these field tests were. It may be possible that different weighting procedures may produce different scores when applied to low-stakes examinations with multiple formats than when applied to high-stakes examinations with multiple formats. Consequently, one of the purposes of the present study was to examine the differences, at the group and student levels, between the scores yielded with unit weighting, a weighted procedure where the SR items and CR items were equally

weighted, and implicit weighting obtained in pattern scoring when applied to low- and high-stakes examinations.

Content Area

Schaeffer et al. (2002) found differences among scale means between, but not within, Biology and English, with the latter approximately five points lower for unit weighting and implicit weighting, and approximately 10 points when the SR and CR items were weighted so that they contributed equally to the total score. Part of the difference is likely attributable to the differences in the difficulty of the two tests (students in Biology answered about 47% of the SR items correctly and achieved about 29% of the CR points; the corresponding values for English were, respectively, 60% and 55%) (Schaeffer et al., 2002, p. 323). And, in the case of the explicitly weighted methods, greater weight was assigned to the CR items which, given the lower performance on the CR items for Biology, would serve to increase the difference between the total Biology and English scores. Therefore, the second purpose of the present study was to examine the scores yielded with unit weighting, a weighted procedure where the SR items and CR items were weighted equally, and implicit weighting when applied to low- and high-stakes examinations in the Language Arts and the Mathematics content areas.

Method

Examinations

The low-stakes and high-stakes examinations considered in the present study were the Grade 9 2003 Provincial Achievement Tests (PATs) for Language Arts and Mathematics and the Grade 12 2003 Diploma Examinations (DEs) for Language Arts and Mathematics administered in Alberta (Alberta Education, n.d.) The PATs, which are

administered at Grades 3, 6, and 9 at the end of the year, are considered low-stakes tests as their main purposes are to help the province, school districts, and schools evaluate how well the curriculum is being taught and improve student achievement. Although student results are reported to parents in the following Fall, student results are not included as part of a final end of year grade for the course (Note 1). In contrast, the DEs are considered to be high-stakes examinations. Administered at the end of each Grade 12 examinable course, the DEs are school exit examinations used to certify individual student competence. The scores from these examinations are combined with the school awarded mark for the courses each student takes and the blended marks (50% and 50%, respectively) are used to determine whether each student has passed or not passed each course and to determine scholarship winners.

The examinations in the areas of Language Arts and Mathematics for Grade 9 and Grade 12 were selected to allow comparisons between low-stakes and high-stakes examinations within and across two content areas that differed in orientation and structure of the examination. Language Arts is more process oriented and Mathematics is more content oriented. Further, the weight of the CR items was greater for Language Arts than for Mathematics. Thus, both purposes of this study could be addressed.

Language Arts

The Grade 9 Language Arts test contained 55 SR and 2 CR items. The SR items assessed the students' ability to recognize explicit and implicit ideas and details and make inferences; interpret text organization; use contextual cues to determine the connotative meaning of words, phrases, and figurative language; and generalize information from an entire reading passage in order to identify the purpose, theme, main idea, or mood of the

passage. Informational, narrative, and poetic texts were used as prompts. One CR item, a narrative essay, required the students to respond to a prompt that consisted of a topic and materials related to the topic (e.g., graphics, quotes, short literary excerpts or ideas from previous experience and/or readings). They were to establish a purpose, select ideas and supporting details, produce a unified and coherent essay, use good sentence structure and vocabulary appropriate for their audience, and use conventions accurately and effectively. The second CR item, a functional essay (business letter and envelope to be addressed) required the students to develop, organize, and evaluate ideas for a specified purpose and audience and to communicate clearly and effectively. The reading and writing components contributed equally to the total score (50% each).

The Grade 12 English Arts examination contained 70 SR and 2 CR items. The SR items required the students to respond to a variety of literary texts, including poetry and prose, and then answer a series of multiple-choice items that assessed the students' ability to construct knowledge from content and context; relate textual forms, elements, and techniques to content, purpose, and effect; and to connect self, culture, and milieu to text and text writer. One CR item, a personal response to text, required the students to reflect on and explore ideas and impressions prompted by the topic and to select an appropriate and effective prose form to convey impressions and create a unifying effect and effective voice. The second CR item, a critical/analytical response to a literary assignment, required the students to demonstrate an understanding of the ideas developed by the text creator(s) by analyzing and explaining the personality traits, roles, relationships, motivations, attitudes, and values of the characters developed in the text. They were also expected to present relevant evidence from the text to support their ideas; develop a

coherent, unified composition by choosing an appropriate method (e.g., implicit or explicit controlling idea); demonstrate a repertoire of stylistic choices and vocabulary in a deliberate and controlled manner; and write in a clear and correct manner. As at Grade 9, the writing and reading components contributed equally to the total examination score.

The two CR items were administered approximately one week prior to the multiple-choice items at both grade levels. The administration times at Grade 9 were 2 hours for the CR items and 75 minutes for the SR items, while 2 ½ hours was allowed for the CR items and 2 ½ hours for the SR items at Grade 12.

Mathematics

The Grade 9 Mathematics examination contained 44 SR and 6 numerical response (NR) items that assessed the students' knowledge and problem solving ability in four content areas: Number, Patterns and Relations, Space and Shape, and Statistics and Probability. The Grade 12 Mathematics examination contained 33 SR items, six NR items, and three 3 CR items that emphasized knowledge of mathematical theory, procedural knowledge, and problem solving ability in six content areas: Transformations of Functions; Exponents, Logarithms, and Geometric Series; Trigonometry; Conic Sections; Permutations and Combinations; and Statistics. The NR and CR items contained content from at least two of the six content areas.

Students at both grade levels were required to work through each NR item and then indicate their solution on the machine-scored answer sheet. However, as with the multiple-choice items, one mark was awarded for each correct solution; in contrast to the CR items, part marks were not awarded for the process use. The contributions of the numeric and multiple-choice items were proportional to the number of each item type at

Grade 9 (88% vs. 12%). At Grade 12, the constructed response items contributed 45% to the total score and the multiple-choice items contributed 55%.

Both assessments were administered in one sitting. The administration time at Grade 9 was 1 ½ hours, while one hour was allowed for the CR items and 1/1/2 hours for the SR items at Grade 12. At both grade levels, an additional 30 minutes was permitted to allow students to complete the assessment; this decision was made by the teacher administering the test.

Summary

The descriptions above reflect differences in topics and the level of complexity of the topics covered between the two grade levels within each content area. The differences are attributable to the sequential developmental nature of the curriculum, which is common throughout the province. The descriptions also reveal differences between language arts and mathematics, with reading and writing processes emphasized in Language Arts and content emphasized in Mathematics.

Scoring

The SR and NR items were machine scored and the CR items were scored by paid, trained teachers. The teacher markers were selected from teachers nominated by their school superintendents. To be eligible to mark a provincial achievement test, a teacher must have taught the course with a PAT within the past three years, have a valid Alberta Permanent Professional Certificate, and be employed by a school authority at the time of marking. To be eligible to mark a diploma examination, a teacher must have taught the diploma examination course for two or more years (one or more for Language Arts), be teaching the course in the current year, and have a valid Alberta Permanent

Professional Certificate. Selection criteria included previous experience as a marker subject to a turnover of about 20% for new markers, regional representation, and proportional representation by student population. Approximately 70 teachers were selected for each of the language arts examinations and about 70 teachers were selected for the Grade 12 mathematics examination.

Five, 5-point scoring scales were used to score the CR Language Arts Narrative/Essay (Grade 9) and Critical/Analytical (Grade 12) items and two 5-point scales were used to score the Function Essay (Grade 9) and Personal Response to Test (Grade 12) items. Prepared solution guides were used to score the Grade 12 CR mathematics items. Initial training, which took place on the first day of marking, required approximately one day for language arts assessments and a half day for the Grade 12 mathematics assessment. Continuous training was conducted at the beginning of the morning and again after lunch on each marking day. Each Language Arts response was separately marked by two markers. When the two marks for a response differed by more than two points, the response was marked by a third marker. The Mathematics CR responses were marked by one marker. Inter-marker reliability was checked on a regular basis, with all markers in the marking room marking a common paper. Three weeks were needed to mark language arts and two weeks were needed to mark the Grade 12 mathematics examination.

Student Samples

The total number of students who wrote the Grade 9 Language Arts and Mathematics tests were 39,493 and 39,604, respectively; the corresponding numbers for the Grade 12 examinations were 26,566 and 21,114. Two random samples of 2,000

students were selected without replacement for each examination to allow estimation of the stability of the results across samples selected from the same population.

Calibration

The assumption of unidimensionality was assessed using principal component analysis and was found to be tenable for all four assessments in both samples. For example, principal component analysis of the 55 English 9 SR items yielded 18 components with eigenvalues greater than 1.0 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. 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 of the analytic scale scores for the 2 English 9 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. The successive differences between remaining components were small (0.37, 0.13, 0.04, 0.02, and 0.02). Thus, the assumptions of uni-dimensionality and local independence were met for the SR and CR subtests included in each examination for both samples. Further, less than 1% of the students did not complete each question in each of the four examinations and since extra time could be provided where needed, speededness was not a factor (Hambleton, Swaminathan, & Rogers, 1991).

The calibration procedures used by Schaeffer et al. (2002) and Sykes and Hou (2003) were adopted in the present study so as to facilitate comparisons of their results with the results obtained in the present study. The item parameters for each test were

simultaneously calibrated on the same scale using three-parameter logistic model (3-PL; Lord, 1980) for the SR items and Muraki's (1992) "generalized" partial credit model for the CR items. Given the NR items in Mathematics required the students to develop their answers and not select their answers from among options provided, they were included with the CR items. The item parameters were estimated using PARDUX (Burket, 1998). WINFLUX (Burket, 1999) was used to place the item parameter estimates onto a scale score scale. The scaling parameters were a multiplier of 50 and an additive constant of 500; the lowest obtainable score (LOSS) and highest obtained scale score (HOSS) were set at 300 and 700, respectively, to allow for a range of scale scores (four standard deviations) sufficiently wide to accommodate different weightings of the SR and CR items (Schaeffer et al., 2002; Sykes & Hou, 2003).

Weighting

The scores on the SR items and CR items on each examination were combined according to the following three weighting procedures: unit weighting of both the SR and CR items (UNW); differential weighting of the SR items and CR items so that each is weighted equally (WN/M); and implicit weights (IW) obtained using IRT pattern scoring. The *explicit* unit and weighted total scores were computed from:

$$\xi(X_a / \hat{\theta}_a) = w_m \left\{ \sum_{i=1}^{n_{sr}} w_i P_i(\hat{\theta}_a) + \sum_{j=1}^{n_{cr}} w_j \sum_{k=1}^{m_j} (k-1) P_{jk}(\hat{\theta}_a) \right\},$$

where $\xi(X_a / \hat{\theta}_a)$, the expected total score for randomly chosen student a with estimated ability $\hat{\theta}_a$,

$P_i(\theta_a)$ is the probability that student a with ability θ_j answers dichotomously scored item i correctly,

$P_{ik}(\hat{\theta}_a)$ is the probability that student a is awarded the k^{th} score from m_j possible scores of CR item j ,

w_i and w_j are the weights to be applied to the n_{sr} selected response items and the n_{cr} constructed items, respectively,

and w_m , which multiplies each item probability, is used to determine the total number of points in the total score. Set to 1, the number of test points is allowed to increase as w_i and/or w_j exceed 1. Alternatively, w_m can be set to a fractional value so as to preserve the original number of points in the total score (Sykes & Hou, 2003, pp. 263-264).

The UNW scores were obtained by setting w_i and w_j to one. To obtain the UN/M scores, the students' scaled scores were weighted so that the expected scores were equal to the number of SR point plus n/m times the number of points earned from the CR items, where n = total number of possible SR points and m = total number of possible CR points and, when included, NR points (Schaeffer et al., 2002). For Grade 9 and 12 Language Arts, the n/m weights were, respectively, 1.22 (55/45) and 2.33 (70/30); the corresponding weights for Mathematics were, respectively, 7.33 (44/6) and 1.57 (33/21). The IW weights were produced by the 3PL/PC model, which used *implicit* item weights that maximize test information and, correspondingly, minimize the standard error of measurement. In contrast to the explicit weights, which are constant for the items in the SR subtest and the CR subset, the implicit weights vary by item. Lastly, the common weight, w_m , was set to a fractional value that the number of points for the total score

using each of the explicit weighting procedures equaled the number of points for the total score using implicit weighting.

Information

The test information at ability θ for the explicitly weighted items was computed from:

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

where w_l is the weight for item l , $l = 1, 2, \dots, n$, n the total number of items,

$P'_{lk}(\theta)$ is the first derivative of the l^{th} item probability, $P_{lk}(\hat{\theta})$

and $\sigma^2(w_m w_l X_l | \theta)$ is the variance of the weighted expected score (Sykes & Hou, 2003, p. 264).

Given the implicit weights were obtained in pattern scoring, the values of w_l and w_m were one. Therefore, the test score information for the implicit weight obtained from pattern scoring is (Sykes & Hou, 2003, p. 264):

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

The corresponding standard error for each θ value is the square root of the inverse of these two functions.

Results

Given space limitations and the marked similarity between the results for the two samples for each examination, only the results for one sample are provided (Note 2).

Group level results are presented first followed by student level results.

*Agreement among Unit Weighted, Differentially Weighted, and Implicit Weighted Scores
at the Group Level*

Characteristics of the Examinations

The summary classical test score statistics for the four examinations are reported in Table 1. With the exception of the CR items for Grade 9 Language Arts, the mean performance on the SR and CR components ranged between 65% and 71%. Further, the distributions of scores were negatively skewed, with the skewness values for the total scores ranging from -0.43 (Grade 9 Mathematics) to -0.25 (Grade 12 Language Arts). These results suggest that the examinations were relatively easy. It is expected that province wide, 85% of the students will achieve the standard of acceptable performance. The values of Cronbach's alpha for the SR items were at least 0.85 across the four examinations (Note 3). The correlations between the subsets of SR and CR items were moderate, 0.59 and 0.64 for Language Arts 9 and 12 and 0.77 and 0.70 for Mathematics 9 and 12, suggesting that the SR and CR items were measuring, in part, something different.

TABLE 1

Descriptive Statistics: Observed Scores

| Item | # of | # of | | | | | | |
|------------------------|-------|--------|-------|------|-------|-------|-------|------------------|
| Type | Items | Points | M | M % | SD | Skew | Kurt | Rel ^a |
| Grade 9 Language Arts | | | | | | | | |
| SR | 55 | 55 | 36.63 | 66.6 | 8.23 | -0.51 | -0.24 | 0.86 |
| CR | 2 | 45 | 23.75 | 52.8 | 4.76 | 0.14 | -0.10 | |
| Total | 57 | 100 | 60.38 | 60.4 | 11.67 | -0.28 | -0.31 | |
| Grade 12 Language Arts | | | | | | | | |

| | | | | | | | | |
|----------------------|----|-----|-------|------|-------|-------|-------|------|
| SR | 70 | 70 | 47.36 | 67.6 | 10.16 | -0.37 | -0.46 | 0.89 |
| CR | 2 | 30 | 21.14 | 70.4 | 4.53 | -0.01 | -0.35 | |
| Total | 72 | 100 | 68.68 | 68.7 | 13.50 | -0.25 | -0.48 | |
| Grade 9 Mathematics | | | | | | | | |
| SR | 44 | 44 | 29.79 | 67.7 | 8.45 | -0.39 | -0.70 | 0.92 |
| NR | 6 | 6 | 3.88 | 64.7 | 1.75 | -0.58 | -0.63 | |
| Total | 50 | 50 | 33.66 | 67.3 | 9.84 | -0.43 | -0.68 | |
| Grade 12 Mathematics | | | | | | | | |
| SR | 33 | 33 | 22.90 | 69.4 | 5.64 | -0.38 | -0.40 | 0.85 |
| NR | 6 | 6 | 4.25 | 70.8 | 1.44 | -0.64 | -0.23 | |
| CR | 3 | 15 | 9.84 | 65.6 | 3.03 | -0.24 | -0.61 | |
| Total | 42 | 54 | 36.99 | 68.5 | 9.23 | -0.33 | -0.46 | |

^a Cronbach's alpha (Note 3).

Scale Scores

The means and standard deviations of the three scale score distributions and the correlations among the three sets of scale scores are provided in Table 2 for the four examinations. Inspection of the means reveals that the three scale means for each assessment were all less than 500. As mentioned above, the examinations were relatively easy. Consequently, the mean ability estimate on the theta scale will be to the left of zero (Lord, 1980). Given the scaling process, the transformed scale score means would likewise be to the left of 500.

Although there were significant differences among the three scale score means for each examination ($p < 0.05$) and between the means across examination, the effect sizes, Δ , were less than 0.18. With the exception of Grade 9 Mathematics, the standard deviations of the three score distributions are comparable for each examination. Lastly, the correlations among the three sets of scale scores for each examination are all above

0.96, suggesting that the students were ranked in essentially the same order for each weighting procedure.

TABLE 2

Descriptive Statistics: Scale Scores

| | M | SD | Correlations | | |
|-------------------------------|--------|-------|--------------|------|------|
| | | | UNW | WN/M | IW |
| Grade 9 Language Arts | | | | | |
| UNW | 494.31 | 57.00 | | | |
| WN/M | 493.97 | 56.89 | 1.00 | 1.00 | |
| IW | 494.71 | 54.68 | 0.97 | 0.96 | 0.96 |
| Grade 12 Language Arts | | | | | |
| UNW | 483.02 | 56.96 | | | |
| WN/M | 480.40 | 55.21 | 1.00 | 1.00 | |
| IW | 483.40 | 55.59 | 0.98 | 0.98 | 0.98 |
| Grade 9 Mathematics | | | | | |
| UNW | 492.23 | 63.87 | | | |
| WN/M | 496.33 | 54.25 | 0.99 | 0.99 | |
| IW | 492.29 | 64.21 | 0.99 | 0.99 | 0.98 |
| Grade 12 Mathematics | | | | | |
| UNW | 495.37 | 59.66 | | | |
| WN/M | 496.34 | 58.97 | 1.00 | 1.00 | |
| IW | 495.92 | 59.45 | 1.00 | 1.00 | 1.00 |

Note:- UNW = Unit Weighting; WN/M – Weighting SR/CR; IW = Implicit Weighting

*Agreement among Unit Weighted, Differentially Weighted, and Implicit Weighted Scores
at the Student Level*

The three scaled scores at the student level were compared in three ways. First, differences among the three distributions of conditional standard errors of measurement were examined. Second, the differences among the three scale scores were compared at

the 10th, 50th, and 90th percentile points. Third, the differences between pairs of scores were summarized using the root mean square deviation to provide a measure of the average difference between the scale scores yielded by the two weighting procedures being compared:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{2,000} (X_{i1} - X_{i2})^2}{n-1}},$$

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

X_{i2} is the scale score for a second weighting procedure for the same person i , and

2,000 is the number of students.

Standard Error of Measurement

The distributions of the conditional standard errors of measurement for each weighting procedure are shown in Figure 1 for Grade 9 Language Arts and in Figure 2 for Grade 9 Mathematics. The plots for the corresponding Grade 12 examinations were very similar and, therefore, are not provided (Note 2).

Comparison of Figures 1 and 2 reveals that there are differences between the shapes of the distributions for the two subject area examinations. The smallest values of standard error for the three weighting procedures occurred around the means (between 475 and 575) for the language arts examinations and increased much more rapidly for scale scores above 575 than for scale scores below 475. In contrast, the distributions of the standard errors for each of the weighting procedures for the mathematics examinations were parabolic in shape. The lowest standard errors occurred between 450 and 550, with a sharp increase for scale scores below 450 and above 550. Further, the

distribution of the standard errors of the UNW procedure begins to deviate from the distributions for the other two weighting procedures at about a score of 400 and the increase is sharper in the right tail than in the left tail. Consequently, while the values of the vast majority of standard errors were between approximately 10 and 20 points for scores up to 600 for Language Arts for each weighting procedure, the values of the majority of standard errors were between 10 and 20 score points for a more restricted range of scores, approximately 450 to 550, for Mathematics. As found by others (Schaeffer et al., 2002; Sykes and Hou, 2003), the standard errors for implicit weighting were smaller than the standard errors for the two explicit weighting procedures and, again, particularly for low scores.

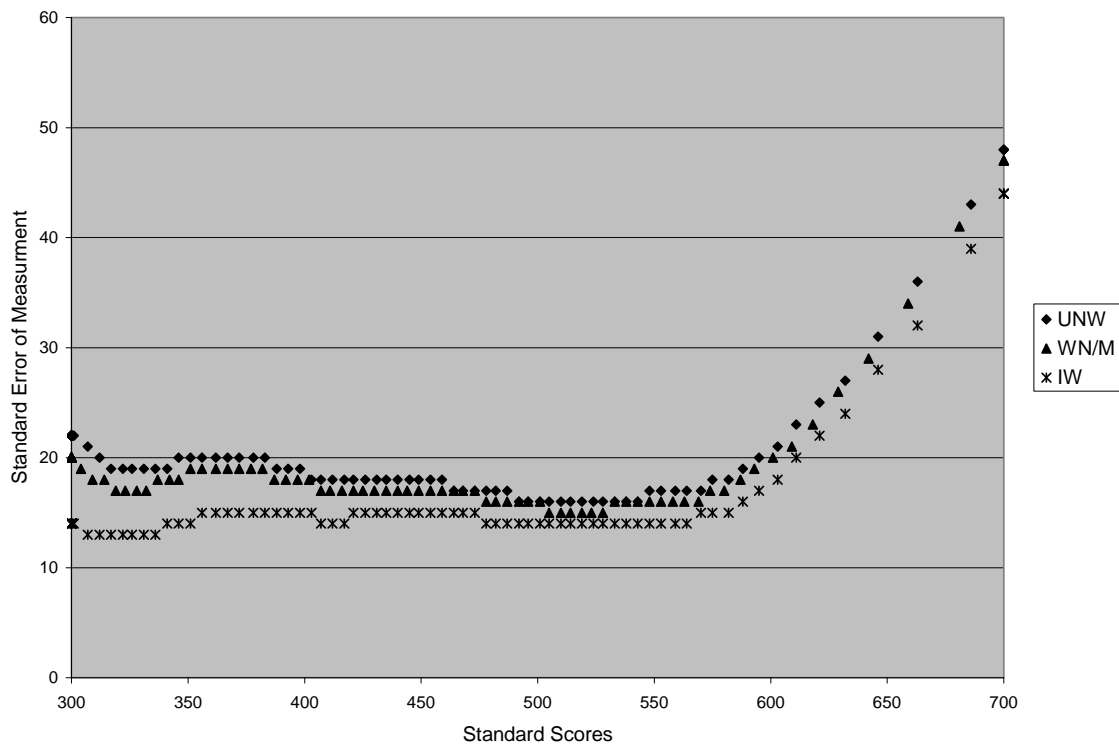


FIGURE 1. Standard Error of Measurement for Grade 9 Language Arts

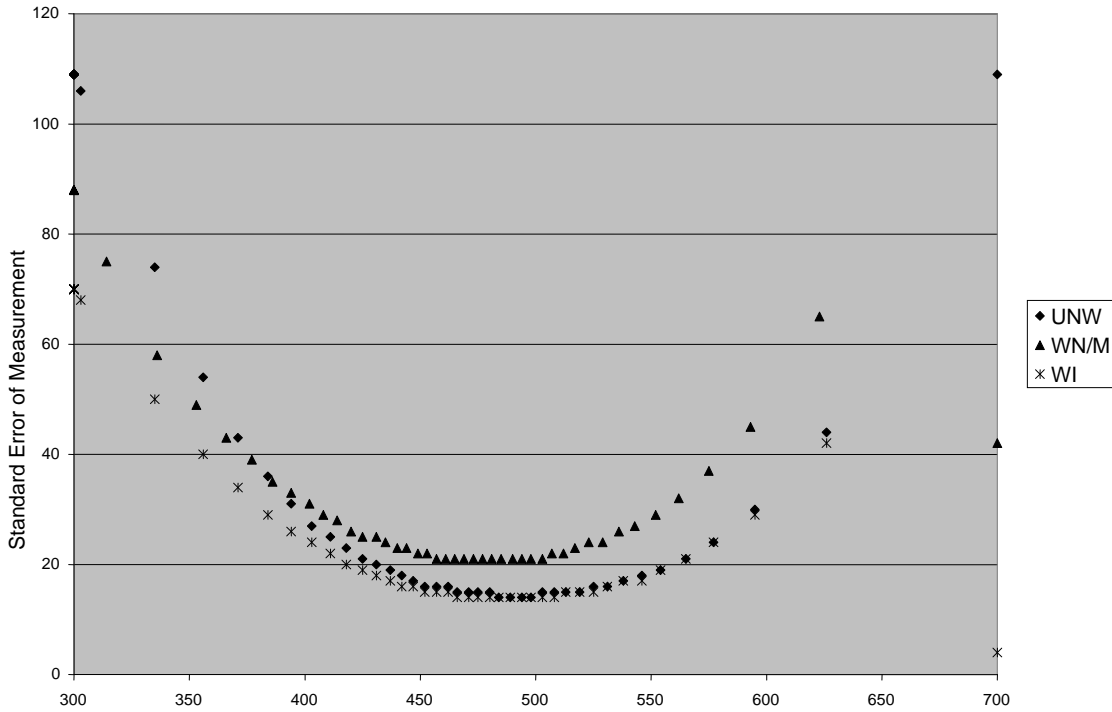


FIGURE 2. Standard Error of Measurement for Grade 9 Mathematics

Scale Score Differences at the 10th, 50th and 90th Percentile Points

The differences between pairs of scale scores yielded by the three weighting procedures at the 10th, 50th, and 90th percentiles are reported in Table 3. A negative difference indicates that the percentile value for the first named weighting procedure was less than the percentile value for the second named procedure in a comparison. For example, the value of - 19 for the 10th percentile for Grade 9 Language Arts indicates that the scale score for the UNW weighting procedure was 19 scale points lower than the scale score for the IW weighting procedure.

TABLE 3

Scale Score Differences at the 10th, 50th, and 90th Percentiles and Root Mean Square Values

| Comparison/ | Percentile | Root Mean |
|-------------|------------|-----------|
|-------------|------------|-----------|

| Examination | 10 th | 50 th | 90 th | Square |
|---------------------------|------------------|------------------|------------------|--------|
| Grade 9 Lang Arts | | | | |
| UNW vs. WN/M | 0 | 0 | 1 | 0.74 |
| UNW vs. IW | -19 | 0 | 18 | 14.95 |
| WN/M vs. IW | -20 | 0 | 18 | 14.97 |
| Grade 12 Lang Arts | | | | |
| UNW vs. WN/M | 1 | 2 | 5 | 3.69 |
| UNW vs. IW | -15 | 0 | 14 | 11.89 |
| WN/M vs. IW | -18 | -3 | 12 | 12.25 |
| Grade 9 Math | | | | |
| UNW vs. WN/M | -15 | 0 | 3 | 13.56 |
| UNW vs. IW | -7 | 0 | 7 | 9.78 |
| WN/M vs. IW | -6 | 0 | 16 | 15.83 |
| Grade 12 Math | | | | |
| UNW vs. WN/M | -2 | -1 | 0 | 1.76 |
| UNW vs. IW | -7 | 0 | 6 | 6.20 |
| WN/M vs. IW | -6 | 0 | 7 | 5.97 |

There was closer agreement among the values of pairs of weighted scores at the 50th percentile than at the 10th and 90th percentiles. With one exception of Grade 9 Mathematics at the 10th percentile, the largest score differences are between each of the explicit weighting procedures and the implicit weighting procedure, with the explicit weighted scores being less than the implicit weighted scores at the 10th percentile and but

greater at the 90th percentile. For example, the UNM scale score was 19 score points below the IW scale score at the 10th percentile and 18 score points above the IW scale score at the 90th percentile for Grade 9 Language Arts.

The differences between the UNW and IW and WN/M and IW scale scores were greater than the differences between the UNW and WN/M scale scores for both Language Arts examinations. Further, the Grade 9 differences were somewhat greater than the Grade 12 differences. In contrast, with two exceptions (Grade 9 Mathematics, UNW vs. WN/M at 10th percentile, $d = -15$; Grade 9 Mathematics, WN/M vs. IW at 90th percentile, $d = 16$), there was closer agreement between the scale scores for Mathematics across both grade levels.

Root Mean Square Deviation

As shown in Table 3, the values of the root mean squares generally mirrored the percentile results for Language Arts; the root mean square deviations between the UNW and IW scale scores and the WN/M and IW scale scores were greater than the root mean square deviations between the UNW and WN/M scale scores. Further, the Grade 9 root mean deviations, with one exception (UNW and WN/M), were somewhat greater than the Grade 12 root mean deviations. The root mean square deviations for Grade 9 Mathematics were greater than the root mean square deviations for Grade 12 Mathematics, which, overall, were the smallest of all the root mean deviations.

In all cases, the largest differences were between the explicit weighting procedures and the explicit weighting procedure. This is likely due to the use subtest (multiple-choice subtest, constructed response subtest) weights in the case of the UNW and WN/M procedures instead of item weights which are used in the IW procedure.

Discussion

The results obtained in the present study revealed 1) that there was greater comparability among the three sets of weighted scores at the group level than at the individual student level, particularly in the extremes of the distributions, and 2) that whereas stakes of a test was a not factor, subject area was.

The three scale means obtained using UNW (unweighted), WN/M (selected response and constructed response equally weighted), and IW (item response) procedures were comparable within each of the four examinations. However, the means for Grade 12 Language Arts were approximately 10 scale points lower than the means for Grade 9 Language Arts and Mathematics and Grade 12 Mathematics, which were comparable to the means obtained by Schaeffer et al. (2002) and Sykes and Hou (2003). The standard deviations and correlations were consistent with those found by Schaeffer et al. (2002) and Sykes and Hou (2003). Thus, it appears that the three weighting procedures yield comparable group level results and rank the students in the essentially the same order, thus justifying the use of any one without introducing bias. The observation that the Language Arts 12 scale means were less than the means for the other assessments may be attributable to lower reliability. While reliability values for the CR items was not available (Note 3), the number of third reads for the Language Arts 12 CR items was greater than the number of third reads for the Language Arts 9 CR items. As reflected in the descriptions of these two tests, there is opportunity of a greater range and complexity of responses at Grade 12 than at Grade 9. This is to be expected given the progression of the curriculum and actual responses received (Tim Coates, Personal Communication, December 19, 2008). Wainer and Thissen (1993) pointed out the need for equal reliability

of the components to be weighted. It may be that difference in reliabilities of the SR and CR components of the Language Arts 12 examination was greater than the difference in reliabilities for the remaining three examinations (Note 2). However, the three scale means for all four examinations were similar, suggesting that any disparity in the reliabilities influenced the two explicit weighting procedures and the implicit weighting procedure in the same way, resulting in the comparability of the three scale means within each examination.

In contrast to the group findings, the three weighting procedures did not yield scale score distributions that were sufficiently similar to warrant using the procedures interchangeably at the student level. In agreement with the findings of Schaeffer et al. (2002) and Sykes and Hou (2003), the standard errors for the IW procedure were generally smaller than the standard errors for the UNW and WN/M procedures, particularly for low scores, for all four examinations. Interestingly, whereas the three distributions of conditional standard errors of measurement for Mathematics were parabolic in shape, the three standard error distributions for Language Arts were shaped similar to a reversed capital “L.” With the exception of one unexplained result (Grade 9 Mathematics), there was good agreement between the scale scores yielded by the UNW and WN/M procedures for both Language Arts examinations and the Grade 12 Mathematics examination at the 10th, 50th, and 90th percentiles and on average. However, there was less agreement among between each these explicit weighting procedures and the implicit IW procedure at the 10th and 90th percentiles and, in agreement with Sykes and Hou (2003), particularly for Language Arts. Consequently, whereas the explicit UWN and WN/M procedures and the implicit IW procedures appear to be

interchangeable when only group level results are required, they are not interchangeable at the student level, particularly at the lower end and higher ends of the score distribution.

The results also revealed that the stakes of an examination – high or low – was not a factor in the present study. While the Grade 9 students may have initially perceived the consequences of the Grade 9 tests as low does not necessarily mean their teachers and principals do, particularly in light of the public reporting of school level results (Rogers & Klinger, 2007). Consequently, teachers and principals of Grade 9 students may have provided an atmosphere that led their students to ignore the low consequences and to do their best.

In contrast, subject area may be a factor. The shape of the distributions of conditional standard errors for both Language Arts examinations was essentially a reversed “L”; the shape of standard error distributions for both Mathematics examinations was essentially parabolic. Differences among the scale scores corresponding to the 10th and 90th percentiles tended to be more pronounced for the Language Arts examinations than the Mathematics examinations. This latter finding needs to be tempered by the differences between the weights for the four examinations and differences between the natures of the two content areas. While the weights applied to the selected response and constructed response items for the UNW (1.00 and 1.00) procedure was constant across the four examinations, the weights for the WN/M procedure varied (1.00 and 1.22 for Grade 9 Language Arts, 1.00 and 2.33 for Grade 12 Language Arts, 1.00 and 7.33 for Grade 9 Mathematics, and 1.00 and 1.57 for Grade 12 Mathematics). These weights were applied to the subset of SR items and the subset of CR items, respectively. Consequently, each item in each subset was explicitly weighted the same. In contrast, the

weights in the IW procedure are implicitly set for and applied to each individual item and, therefore vary by item. The differences between the explicit and implicit weights may have influenced the agreement between the scales scores obtained using each of the UNW and WN/M procedures and the scales scores obtained using the IW procedure at the 10th and 90th percentile points where there were fewer students and items with the corresponding difficulties. Also, it is necessary to consider the difference in the content assessed by the Language Arts and Mathematics examinations. Language Arts is more process oriented while mathematics is more content oriented. This can best be seen in the nature of the CR items and their accompanying scoring guides. Much more latitude is provided to students composing a personal response than to students responding to a mathematics question with a single correct answer. Additional work is needed to ascertain if there is or are not any systematic differences among the weighting procedures across other subject area examinations. Lastly, existing data sets were used for the present study. The Language Arts teachers at both grade levels were encouraged to tell their students that performance on the written part of the Language Arts test would count the same as their performance on the selected response Reading part. For mathematics, the Grade 9 teachers were encouraged to tell their students that each item multiple-choice and constructed response item counted the same while the Grade 12 teachers were encouraged to tell their students that the constructed response and constructed numeric response items would account for 45% of the total score and the multiple-choice items would count 55%. What is not known is to what degree this information, when provided, influenced the way the students responded to the items and tasks presented and how this might have affected the results.

Conclusion

Since each weighting procedure has its advantages, such as the smallest standard error when number of CR items is small (IW), likely easiest to explain to students and parents (UNW), and explicitly takes into account time and effort spent by students on each item type (WN/M), no firm recommendation about what weighting procedure to use can be made from the results of this study. While the results of the present study confirmed and extend the results reported by Schaefer et al. (2002) and Sykes and Hou (2003), additional research is needed to clarify the role that the stakes of the examination on the weighting procedure and the apparent interaction between weighting procedure and content area. Is it the case that the results obtained using explicit vs. implicit weighting for the biological, chemical, mathematical, and physical sciences will systematically differ from the results for the humanities and the social sciences at the student level but not at the group level? More research is needed to clarify the role of student performance and the nature of the shape of score distributions when using different weighting procedures.

Researchers and government officials with provincial/state or national assessment programs need to carefully consider the implications of which weighting procedure is chosen in those situations where student level decisions are to be made. For example, curriculum specialists may call for the equal weighting of SR and CR items in provincial/state examinations (WN/M). On the other hand, special education specialists concerned with the placement of low performing students may call for low standard errors (IW). Other policy makers, teachers, and principals most familiar with unweighted scores and concerned about the difficulty of explaining WN/M and IW weighting and

fear that the use of such weights is to make things look “good” may call for the familiar unweighted scores (UNW). It follows then, that a detailed justification and procedure for the weighting procedure to be used should be developed and included in a technical report for the assessment and provided, when required, to the stake holders, including members of the education community, students and parents.

Notes

¹ Beginning in 1996, teachers were encouraged to mark the tests before sending them to the central marking centre and to include them as part of the final grade for each student in their classes. Anecdotal evidence suggests the contribution of the PAT toward final course grades varies from zero to 10%.

² All results are available from the first author.

³ The inter-rater reliability for the constructed response items was not available from the agency that provided the data for this study (Ping Yang, Personal Communication, October 18, 2007). However the number of third markers is greater for Language Arts 30 than for Language Arts 9 (Tim Coates, Personal Communication, December 19, 2008).

References

- Alberta Education. (n. d.). *About provincial testing*. Retrieved December 18, 2008, from <http://www.education.alberta.ca/admin/testing.aspx>
- Blau, J. R., Moller, S., & Jones, L. V. (2004). Why test? Talent loss and enrolment loss. *Social Science Research, 33*, 409-434.
- Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of EM algorithm. *Psychometrika, 37*, 29-51.
- Brown, S. m., & Walberg, H. J. (1993). Motivational effects on test scores of elementary students. *Journal of Educational Research, 86*, 133-136.
- Burket, G. R. (1998). PARDUX for windows (Version 1.17). Monterey, CA: CTB/McGraw-Hill.
- Burket, G. R. (1998). WINFLUX (Version 1.01). Monterey, CA: CTB/McGraw-Hill.
- Demars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*, 55-77.
- Ferrer-Caja, E., & Weiss, M. R. (2002). Cross-validation of a model of intrinsic motivation with students enrolled in high school elective courses. *Journal of Experimental Education, 71*, 41-65.
- Fitzpatrick, A. R., Link, V., 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.

Gulliksen, 1950

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

Haertel, 2006; Kolen, 2006

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

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

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

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Muraki, E. (1992). A generalized partial credit model: Application of the EM algorithm. *Applied Measurement in Education*, 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). Hillsdale, NJ: Lawrence Erlbaum Associates.

- Rodriquez, 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.
- Rogers, W. T., & Klinger, D. A. (2007).
- 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.
- Tanaka, J. S. (1993). Multifacet conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10-39). Newbury Park, CA: Sage.
- van Barneveld, C. (2003). The effects of examinee motivation on multiple-choice item parameter estimates. *Alberta Journal of Educational Research, XLIX*, 277-289.
- 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.

Wise, L. L. (1996). *Indicators of student effort on the National Assessment of Educational Progress* (Tech. Rep. for Grant No. R99B60002). Alexandria, VA: Human Resources Research Organization.

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.