The Alberta Journal of Educational Research *Vol. XLIX, No. 1, Spring 2003, 83-97*

Don A. Klinger
Queen's University

and

W. Todd Rogers
University of Alberta

An Investigation of the Accuracy of Alternative Methods of True Score Estimation in High-Stakes Mixed-Format Examinations

Increasingly, high-stakes large-scale examinations are used to make important decisions about student achievement. Consequently, it is equally important that scores obtained from these examinations are accurate. This study compares the estimation accuracy of procedures based on classical test score theory (CTST) and item response theory (Generalized Partial Credit model, GPCM) for examinations consisting of multiple-choice and extended-response items. Using the British Columbia Scholarship Examination program, the accuracy of the two procedures was compared when the scholarship portions of the examinations were removed. For the subset of examinations investigated, the results indicate that removing these scholarship portions led to an error rate of approximately 10% with approximately seven out of 10 errors resulting in the denial of scholarships. The results were similar for both the CTST and the GPCM, indicating that for mixed-format examinations the two procedures produce randomly equivalent results. Implications for policy and future research are discussed.

Lors de la prise de décisions importantes quant au rendement des élèves, on tient de plus en plus compte d'examens à grande échelle et à enjeu considérable. Il est donc tout aussi important que les résultats qu'obtiennent les élèves à ces examens soient justes et précis. Cette étude compare la justesse de l'estimation de procédures reposant sur la théorie classique des scores (classical test score theory, CTST) d'une part, et la théorie de la réponse d'item (Generalized Partial Credit model, GPCM) d'autre part, pour des examens comprenant des questions à choix multiples et des questions ouvertes. La justesse des deux procédures à été comparée dans le contexte du programme d'évaluation pour les bourses d'études de la Colombie britannique (British Columbia Scholarship Examination program) duquel la section sur les bourses d'études avait été retranchée. Les résultats de l'étude portant sur le sous-ensemble d'examens indiquent que le fait d'enlever la section liée aux bourses d'études donnait un taux d'erreurs d'environ 10% où à-peu-près 7 erreurs sur 10 menait à un refus d'accorder la bourse. Les résultats pour les deux procédures (CTST et GPCM) étaient similaires, ce qui indique que pour les examens à format mixte, les procédures donnent des résultats équivalents au hasard. Une discussion des incidences de l'étude sur les politiques et la recherche termine l'article.

Don Klinger is an assistant professor in Assessment and Evaluation in the Faculty of Education. His research interests include the examination of psychometric and policy issues of large-scale assessments, standard setting, and measures of school effectiveness.
Todd Rogers is a professor and Director of the Centre for Research in Applied Measurement and Evaluation. His research interests are in test translation, ethics in testing, and psychometrics.

Increasingly, large-scale testing is used to make important decisions about student achievement. A review of the provincial Ministries of Education in Canada indicates that British Columbia, Alberta, Manitoba, Quebec, New Brunswick, and Nova Scotia have provincial examination programs that help determine high school students' final grades (see also Cheliminsky & York, 1994; Lafleur & Ireland, 1999). Such examination programs are considered high-stakes because of the implications of the results for the students. The stakes associated with the British Columbia examination program may even be higher because the results are also used to award provincial academic scholarships to high-achieving students who wish to pursue postsecondary education.

Historically, such examinations have been based on classical test score theory (CTST). With this theory, a student's total test score is used as the estimate of that student's level of achievement. The total score may be a simple sum of the item scores or a weighted sum where the weights are determined to reflect the differential importance of content assessed by the test (Wainer & Thissen, 1993). More recently, item response theory (IRT, Lord, 1952) has been used in several testing programs to obtain an estimate of a student's level of achievement or ability. With this theory, characteristics of the items (e.g., item difficulty and discrimination) in the test are combined with the student's response pattern to obtain an estimate of the student's ability, where ability (θ) is described in reference to the domain of subject matter as defined by the examination. Thus a fundamental difference between CTST and IRT is that CTST operates at the examination level, whereas IRT operates at the item level.

IRT uses a series of mathematical models that proponents of IRT (e.g., Hambleton & Swaminathan, 1985) claim have several advantages over the CTST model. Chief among these advantages is the ability to select the IRT model that best fits simultaneously the characteristics of the student responses as well as the characteristics of the examination items. The item characteristics and the response pattern across the items of each student are used to provide an estimate of that student's ability (θ). Thus two students with the same total raw test score but different response patterns could receive different θ estimates. The student who correctly answered the more difficult and discriminating test items would receive a higher θ estimate than the student who correctly answered the same number of easier but less discriminating items. Such discrimination among students is not possible with CTST. Although computationally complex, the increase in computing power has enabled psychometricians to use IRT and its associated mathematical models as an alternative and perhaps more accurate method to estimate student achievement than CTST and its associated model.

Currently, large testing companies use IRT as the foundation for measuring achievement with examinations that have either multiple-choice (MC) and/or extended-response (ER) items. In contrast, despite the apparent advantages of IRT and the availability of computer programs that can quickly complete the analyses, provincial testing officials in Canada continue to rely largely on CTST. Given the high stakes associated with provincial examination programs, it is essential that the results provide an accurate estimate of student achievement. Therefore, it is important to investigate whether the results yielded by the IRT models are superior to the current results yielded by the CTST model.

One of the difficulties of comparing results from the CTST model with the results from the IRT models is the lack of a standard of comparison. Simulated research has been conducted to assess the utility and superiority of different IRT models (Reise & Yu, 1990). However, it is difficult to compare the CTST model and the IRT models using simulated conditions. One approach for comparing the CTST model and the IRT models involves the use of actual examination data obtained from a shortened form of the full examination (Anderson, 1999; Bock, Thissen, & Zimowski, 1997; Folske, Gessaroli, & Swanson, 1999). The shortened form consists of items that when taken together yield scores that can be validly interpreted in terms of the construct measured by the full-length form. The comparison of the examinee score estimates derived from the shortened examination and the full examination provides a measure of the accuracy of the procedure employed to obtain the estimates: the estimates from both forms in the ideal situation should be equal except for sampling error.

A second problem encountered in the comparison of the CTST model and the IRT models is centered on the definition of achievement. Although defined differently and not synonymously with latent trait, the true score (τ) as estimated using the CTST model is theoretically related to the latent trait estimate θ obtained using an IRT model. This relationship is nonlinear (Lord, 1953, 1980). Because the estimates derived from the CTST model and the IRT models are related, the merits of each theory should be based on the quality of the estimates produced by the models used to operationalize the theory. The estimates obtained from the IRT models should be different and somewhat superior to those obtained from the CTST model, because more information is used in the IRT models than in the CTST model. Both item and examinee response vectors in the two- and three-parameter IRT models are used to determine the ability estimates for each examinee in contrast to the use of only the total test score in the CTST model (Thissen, Pommerich, Billeaud, & Williams, 1995).

These hypotheses have not been consistently supported by earlier research. Although Birnbaum (1968) first demonstrated that the scores based on item response patterns differed from those based on the total or summed scores, other research has shown the differences to be small. Fan (1998) illustrated this using the grade 11 Texas Assessment of Academic Skills (TAAS) examination to compare the estimates yielded by the classical model and the one-, two-, and three-parameter IRT models. Fan found the correlations between the estimates of θ provided by the IRT models and the total test score to be at least 0.96 in all cases and concluded that the same or similar conclusions would be drawn regardless of the method used. Rogers and Ndalichako (2000) and Tomkowicz and Rogers (2001) found similar correlations between the estimates yielded by these four models using, respectively, the grade 12 school-leaving examinations in English, social studies, and mathematics in Alberta. Further, in both of these studies the agreement between the values of the pairs of estimates was also found to be high, which suggests that the estimates could then be used interchangeably. Anderson (1999) compared the estimates from the three-parameter IRT model and the total test score for the dichotomously scored items included in the January 1996 grade 12 mathematics provincial examination in British Columbia. His conclusion, although admittedly exploratory, was that

the estimates derived from the two procedures and transformed to the same scale were almost identical in terms of their means, standard deviations, correlations, and the classification decisions made from using the transformed scores. However, a closer analysis of the results reported by Anderson revealed some small but notable differences. For example, although the mean domain scores for each method were the same, the standard deviation for those scores derived from the three-parameter IRT model was less than the standard deviation of the scores derived from the CTST model (Anderson, Table 1, p. 348). The root mean square error value for the three-parameter model was marginally larger than the root mean square error value for the CTST model (Anderson, Table 3, p. 349). Finally, small differences did exist in the assignment of letter grades to students between the two approaches with the three-parameter model providing marginally superior assignment (Anderson, Table 5, p. 350).

Purposes

Earlier research has been limited in that the comparisons have been completed only for examinations that have dichotomously scored (2 score points) items, most commonly those using MC formats. However, many high-stakes examinations now contain both dichotomously scored MC items and polytomously scored (more than 2 score points) ER test items. Thus the purpose of the current study was to examine and compare the agreement between, and the accuracy of, the achievement scores generated using the CTST model and the IRT Generalized Partial Credit Model (GPCM, Muraki, 1992) on examinations that contain both dichotomously and polytomously scored items. The accuracy of the models was determined by examining the agreement between the scores and decisions yielded by each model on a scholarship examination that was a shortened form of a full-length scholarship examination and the scores yielded by the full-length examination. Accordingly, the second purpose of this study was to determine which model yielded scores that better led to the same scholarship decisions initially made using the full scholarship examination scores.

Overview of the Models

According to CTST, the observed score for an examinee on an examination consists of two additive components, the true score (τ) and the error score (ε) (Crocker & Algina, 1986). Because neither the true nor error scores are known, the problem is indeterminate. However, the observed score is an unbiased estimate of each examinee's true score (Gulliksen, 1950).

In contrast, IRT is represented by a class of probability models that use the student item response vectors to estimate the item level parameter(s) for each item to best fit the distribution of the students' responses. The item parameters are then used to produce ability (θ) estimates for the examinees using maximum likelihood estimates. The choice of model is based on model data fit and on the number of score points for the items included in the analysis. Although all the IRT models will work with dichotomously scored items, only the models developed to handle polytomously scored items will work with examinations that have both dichotomously and polytomously scored items. The generalized two-parameter partial credit model (GPCM) is one such model[1] (Muraki, 1992).

Since its introduction, the GPCM has been found to produce good approximations of the actual parameter and ability estimates under simulated condi-

tions and is being increasingly used in the measurement of polytomously scored ER items (Carlson, 1996; Donoghue, 1994; Fitzpatrick et al., 1996; Muraki, 1992). Earlier research has also shown that the results produced by the GPCM are comparable to the results yielded by the two-parameter graded response model introduced in 1972 by Samejima (Klinger & Boughton, 2000; Maydeau-Olivares, Drasgow, & Mead, 1994). Further, the two-parameter models have been shown to be superior to the previously developed one-parameter models for polytomously scored items (Fitzpatrick et al., 1996; Sykes & Yen, 2000).

The Study

The Full-Length and Shortened Scholarship Examinations

The scholarship examination program in British Columbia was recently shortened by eliminating a portion of the initial examination used to identify scholarship recipients. Before 1996-1997 (since 1984-1985), students interested in obtaining a provincial scholarship wrote two examinations in at least three grade 12 academic courses. For each academic grade 12 course in which they were enrolled, all students completed the mandatory two-hour provincial examination consisting of a set of MC and ER items that encompassed the major concepts in the curriculum. Students interested in obtaining a provincial scholarship also wrote the optional one-hour scholarship examination, which consisted of more conceptually difficult ER items. The number of items on the provincial and scholarship examinations varied depending on the subject. However, the total raw score of the scholarship examination was half the value of the total raw score for the corresponding provincial examination. The total raw scores were simply a sum of the corresponding item scores.

In 1996-1997 the provincial government shortened the scholarship examinations by eliminating the optional scholarship examinations. Scholarships were then awarded based only on the scores achieved on the mandatory two-hour provincial examinations.

For the purposes of the study, the original procedure, in which both the provincial and optional scholarship examinations were in place, is referred to as the *gold standard procedure.* The shortened scholarship examination scores calculated using only the provincial examination scores were compared with this gold standard to determine if the same students would be awarded or not awarded a scholarship. The current procedure is referred to as the *CTST procedure,* and the procedure in which the item response vectors are used is referred to as the *GPCM procedure.*

Calculation of Scholarship Scores

The calculation of the scholarship scores employed in the present study were computed following the procedures used by the British Columbia Ministry of Education.

Gold standard procedure. In the case of the gold standard, individual scholarship scores were calculated for each subject area by summing both the provincial and the corresponding scholarship examination raw scores for each student who chose to write the scholarship examination. The RANKIT transformation procedure (Chambers, Cleveland, Kleiner, & Tukey, 1983) was then used to transform the scholarship score distribution such that it resembled a normal distribution. The transformed scores were then rescaled so that the

mean scholarship score was 500, the standard deviation was 100, and the minimum and maximum scores were 200 and 800. Scholarships were awarded to students who obtained a scaled scholarship score of at least 475 in three subject areas and had a combined minimum scholarship score of at least 1,700 based on these three scores. For students who wrote more than three scholarship examinations, the three highest scores were used to determine the overall combined scholarship score. Thus the calculation of scholarships was a two-stage procedure, stage 1 being the calculation of examination-level scholarship scores and stage 2 the calculation of total (combined) scholarship scores.

CTST procedure. In the case of the CTST procedure, students with a provincial examination score of at least 70% on a subject area examination received a scholarship score for that examination. In this subsample of students, the examination scores were normalized and scaled following the same procedure used for the gold standard to obtain a distribution of scaled scores with mean 500, standard deviation 100, and lowest and highest scores of 200 and 800. Scholarships were awarded to those students who had three subject area scholarship scores of at least 475 and a combined scholarship score of at least 1,700 based on these three scores. As above, if students had a minimum score of 475 in more than three subjects, the three highest scores were used to determine the combined score.

GPCM procedure. For each examination, estimates of θ_i were calculated for all the students who wrote that examination. These estimates were determined using the *expected a posteriori* (EAP) estimator (Bock & Mislevy, 1982). The sample sizes ranged from 1,318 students to 12,449 across the examinations considered in this study. The computer program PARSCALE 3.1 (Muraki & Bock, 1997) was used to complete the necessary computations.

After the θ estimates were computed for all the students, those with a total examination score less than 70% were dropped. The θ estimates were then transformed into examination scores following the same procedure used to calculate the CTST scholarship scores.

Data and Data Analysis

The student responses for the last two years, 1994-1995 and 1995-1996, in which the full scholarship examinations were administered formed the data set for the present study. The use of two years allowed for a replication of the analyses. Due to sample size considerations, student responses from the January and June sittings of the biology, chemistry, geography, geology, mathematics, and physics examinations were used. The total test score value and distribution of marks allocated to MC and ER items varied across the examinations although they were consistent in each examination. Total test scores varied from a low of 70 marks (mathematics) to a high of 120 marks (physics). The proportion of marks allocated to the MC items was 52% for biology, 60% for chemistry and geology, 40% for geography, 71% for mathematics, and 50% for physics. The examinations contained between 5 and 14 ER items having score values commonly between 2 and 6 marks.

The gold standard scores were those actually calculated in 1994-1995 and 1995-1996 based on both the provincial and optional scholarship examinations. The CTST and GPCM scholarship scores were calculated using only the students' responses on the provincial examination for these two years. In effect,

then, the CTST and GPCM scores were determined from a shortened form of the full-length examination used to determine the gold standard scores, thus replicating the methods used in earlier studies to compare estimation accuracy (Anderson, 1999; Bock et al., 1997; Folske et al., 1999).

Before computing the GPCM scholarship scores, the assumptions underlying the use of the GPCM model were tested. Taken together, the dominance of the first component and the difference between the first and second factors (Hambleton, Swaminathan, & Rogers, 1991), the shape of the Scree plot (Cattell, 1952), and Stout's T statistic (Nandakumar & Stout, 1993; Stout, 1987) indicated that each of the examinations considered in this study was essentially unidimensional (Nandakumar, 1994). Based on the responses to the last three MC items, fewer than 1.0% of the students did not complete the MC items included in each examination. The omission rates for the last two (ER) items were no greater than 10%. Although high, these rates are consistent with the omission rates across a number of years, indicating that the degree of difficulty of the ER items and not speed was the factor that accounted for the omission rates. The lack of guessing could not be assumed. However, Lord (1980) and Hambleton and Swaminathan (1985) suggested that the presence of guessing may be a factor only for those students in the lower ability ranges and would not generally affect the estimate of high-ability students. A preliminary analysis revealed that as suggested, the estimates of the students with examination scores of at least 70% determined using a two-parameter model and a three-parameter model were essentially identical (Klinger, 2000). Consequently, the presence of guessing had no influence on the comparisons made.

Assessment of Agreement

At the examination level the agreements between the gold standard and CTST scholarship scores and between the gold standard and the GPCM scholarship scores were assessed using (a) correlations, (b) root mean square errors (RMSE), and (c) classification errors. For each examination only those students who wrote the optional scholarship examination were included in the study. Further, correlations and RMSE values were based on a subsample of students who also obtained a minimum of 70% on the provincial examinations. In the case of scholarship classification, students who did not obtain the 70% minimum were included in the group of students not meeting the minimum 475 score. The correlations provided a measure of agreement between procedures in terms of ranking; however, they were expected to be high because the provincial examination scores were an integral component in the gold standard, CTST, and GPCM procedures. The RMSE values were used to determine if the students received the same scholarship scores across procedures. Finally, classification errors were used to examine if the same students would obtain the minimum scholarship score of 475 across procedures. The results for the combined scholarship scores based on the three best scholarship scores include those students who wrote at least three scholarship examinations in the subjects considered. As above, students with less than 70% on one of the provincial examinations would be classified as not meeting the minimum score (1,700). The accuracy of agreement was measured in terms of classification errors in the identification of scholarship recipients and nonrecipients.

The agreement between the CTST and GPCM procedures was assessed using correlations and by examining the degree of agreement in the identification of scholarship winners using the combined scholarship scores.

Results

Agreement Between Gold Standard and CTST Procedures

The correlations, RMSEs, and classification errors between the gold standard and CTST procedures are presented in Table 1 for the subject area examinations included in the study for both years. Comparison of the results across the two years reveals that the correlational and RMSE results are essentially stable. The highest and lowest correlations in 1994-1995 tend to be the highest and lowest in 1995-1996; the ranges are also similar (0.74-0.92, 1994-1995 vs. 0.71-0.92, 1995-1996). The lowest and highest root mean square errors in 1994-1995 tend to be the lowest and highest in 1995-1996 and the ranges are again similar (42.6-71.69 vs. 38.68-74.03). It is clear that the values of the RMSEs are not small, which suggests that for some students the two sets of scores produce different decisions. This is shown in the three right columns of the panel for each year. False positive decisions are those in which students would obtain the minimum standard (475 at the examination level) using the CTST procedure but not using the gold standard. In contrast, false negative decisions are those in which students would not obtain the minimum standard using the CTST procedure, but did according to the gold standard. Although the percentages of false positives in each year are less than 3.7, the percentages of false negatives are somewhat higher, varying from 6.8% to 18.9% in 1994-1995 and from 5.7% to 19.5% in 1995-1996. On average, the occurrence of false negative classifications was approximately five times more likely than false positive classifications. Taken together, the results at the subject level suggest that the fit between the gold standard and the CTST scholarship scores is not good; approximately one to two out of every 10 students would have had a different outcome at stage 1, namely, a scholarship score of at least 475 in a given subject area.

Reported in the last row of Table 1 are the classification error percentages for the combined scholarship score used at the second stage of the scholarship decision process. At this stage the three highest scholarship scores above 475 are added together and compared with the 1,700 cut score. Again, the percentage of false negatives is greater than the percentage of false positives, but not to the same degree as the differences noted for the individual subject area examinations. The ratio of false negatives to false positives is approximately two to one after the second stage. Overall, in the case of the second stage scholarship scores, approximately one out of 10 students (9.9% in 1994-1995 and 10.3% in 1995-1996) would have a different decision depending on which of the two selection procedures was used. Most of the classification errors would result in students not receiving a scholarship using the CTST procedure although they received a scholarship using the gold standard procedure.

Agreement Between Gold Standard and GPCM Procedures

The correlations, RMSEs and classification errors between the gold standard and GPCM procedures are presented in Table 2. Like the findings for the comparison of the gold standard and CTST procedures (compare Table 1), the correlational and RMSE results across the two years are essentially stable. The highest and lowest correlations between the gold standard scholarship scores

Table 1
Comparison of Examination Scholarship Results Between the Gold Standard and the CTST Procedures

Subject	*1994-1995*						*1995/1996*					
	N	*r*	*RMSE*	*False Positive (%)*	*False Negative (%)*	*Error Rate (%)*	*N*	*r*	*RMSE*	*False Positive (%)*	*False Negative (%)*	*Error Rate (%)*
Biology (Jan.)	1,587	0.91	48.26	0.5	10.5	11.0	1,867	0.92	40.80	1.4	7.1	8.5
Biology (June)	3,227	0.91	42.60	2.0	7.0	8.9	3,604	0.92	38.68	2.4	5.7	8.1
Chemistry (Jan.)	1,534	0.91	58.42	0.2	15.6	15.8	1,833	0.90	47.50	2.3	7.1	9.4
Chemistry (June)	3,913	0.92	44.01	1.6	7.7	9.3	4,170	0.91	45.67	2.0	8.3	10.3
Geography (Jan.)	854	0.74	71.69	2.1	18.9	20.6	1,052	0.71	74.03	3.3	19.4	22.7
Geography (June)	1,717	0.76	64.40	3.4	15.0	18.5	1,904	0.77	62.94	3.7	10.9	14.5
Geology (June)	306	0.81	57.04	3.6	9.8	13.4	308	0.85	64.40	1.3	19.5	20.8
Mathematics (Jan.)	2,259	0.90	51.43	0.7	10.8	11.5	2,765	0.87	55.33	1.4	13.9	15.3
Mathematics (June)	5,283	0.90	51.36	0.9	8.9	9.8	5,252	0.88	52.67	1.6	10.3	11.9
Physics (Jan.)	772	0.88	64.34	1.0	16.5	17.5	945	0.88	55.84	0.8	12.8	13.7
Physics (June)	2,601	0.88	48.40	3.5	6.8	10.4	2,860	0.89	47.49	2.2	11.6	13.8
Combined Score	2,524	-	-	3.4	6.5	9.9	2,769	-	-	3.2	7.0	10.3

Note. Differences in the overall error rate are due to rounding.

Table 2

Comparison of Examination Scholarship Results Between the Gold Standard and the GPCM Procedures

	1994-1995						1995/1996					
Subject	*N*	*r*	*RMSE*	*False Positive (%)*	*False Negative (%)*	*Error Rate (%)*	*N*	*r*	*RMSE*	*False Positive (%)*	*False Negative (%)*	*Error Rate (%)*
Biology (Jan.)	1,587	0.91	46.25	0.4	9.8	10.1	1,867	0.91	40.98	1.4	6.4	7.8
Biology (June)	3,227	0.91	42.67	1.7	7.7	9.4	3,604	0.91	41.15	2.5	5.6	8.1
Chemistry (Jan.)	1,534	0.91	59.30	0.1	14.1	14.2	1,833	0.90	48.36	2.1	7.1	9.2
Chemistry (June)	3,913	0.92	43.83	1.7	7.9	9.6	4,170	0.91	45.19	1.6	9.3	10.9
Geography (Jan.)	854	0.73	74.24	2.6	19.4	22.0	1,052	0.67	77.59	3.6	20.5	24.1
Geography (June)	1,717	0.74	66.59	3.6	14.7	18.3	1,904	0.74	67.93	3.6	11.9	15.5
Geology (June)	306	0.80	57.89	3.9	11.4	15.4	308	0.85	64.23	2.3	19.2	21.4
Mathematics (Jan.)	2,259	0.89	52.68	0.6	11.9	12.5	2,765	0.86	56.03	1.0	14.5	15.6
Mathematics (June)	5,283	0.90	51.40	0.8	9.5	10.3	5,252	0.88	52.50	1.7	10.7	12.3
Physics (Jan.)	772	0.87	65.61	0.8	16.7	17.5	945	0.88	56.41	1.0	13.5	14.5
Physics (June)	2,601	0.88	49.51	3.9	7.4	11.3	2,860	0.88	50.51	2.6	11.1	13.6
Combined Score	2,524	-	-	3.2	6.6	9.9	2,769	-	-	3.8	7.3	11.0

Note. Differences in the overall error rate are due to rounding.

and the GPCM scholarship scores in 1994-1995 tend to be the highest and lowest in 1995-1996; the ranges are also similar (0.73-0.92, 1994-1995 vs. 0.67-0.91, 1995-1996). The lowest and highest root mean square errors in 1994-1995 tend to be the lowest and highest in 1995-1996, and the ranges are again similar (42.67-74.24 vs. 41.15-77.59). The RMSEs are not small, which suggests that for some students the two sets of scores produce different decisions. And again, as shown in the three right columns for each year, although the percentages of false positives in each year were less than 3.9, the percentages of false negatives were larger, varying from 7.9% to 19.4% in 1994-1995 and from 5.6% to 20.5% in 1995-1996. On average, the occurrence of false negative classifications was approximately five times more likely than false positive classifications. The results at the subject level suggest that the fit between the gold standard and the GPCM scholarship scores is not good; approximately one to two of every 10 students would have had a different outcome with respect to obtaining the minimum examination scholarship score of 475.

The classification error percentages for the combined scholarship scores used at the second stage of the scholarship decision process are reported in the last row of Table 2. Like the case for the gold standard and CTST procedures, the percentage of false negatives is approximately twice the percentage of false positives. And again, approximately one of 10 students (9.9% in 1994-1995 and 11.0% in 1995-1996) would have a different decision depending on which of the two selection procedures was used. Most of the errors would result in students not receiving a scholarship using the GPCM procedure although they received a scholarship using the gold standard procedure.

Comparison Between the Current and GPCM Procedures

The results of the agreement analyses presented in Tables 1 and 2 suggest that the CTST and GPCM procedures produced similar results. The highest and lowest correlations and the lowest and highest RMSE values obtained when the CTST scholarship scores were compared with the gold standard scholarship scores are similar to the highest and lowest correlations and the lowest and highest RMSE values obtained when the GPCM scholarship scores were compared with the gold standard scholarship scores. For example, the 1995-1996 January geography examination had the poorest results (low correlation and high RMSE) and the 1995-1996 January and June biology examinations had the best results (high correlation and low RMSE) for both procedures. Similarly, classification error rates appear to be similar at both the first and second stages of the scholarship selection process. The use of either the CTST or GPCM procedure in place of the gold standard procedure results in substantial error rates at the examination level and an approximately 10% error rate in the awarding of scholarships.

An examination of the correlations between the two procedures across examinations indicates the two rank students closely. The lowest correlation was 0.92 (geography, January 1994-1995) with the vast majority of correlations equalling 0.97. Given this similarity, a final comparison was made between the CTST and GPCM procedures using a confusion matrix. This matrix, which is presented in Table 3, summarizes the agreement between the scholarship decisions made using the CTST scores and the scholarship decisions made using the GPCM scores on the shortened scholarship examinations. In Table 3

Table 3
Confusion Matrix for the Scholarship Decisions Comparing the CTST and the GPCM Procedures

		GPCM Procedure		
		Not awarded scholarship	Awarded scholarship	Row totals
CTST Procedure	Not awarded scholarship	1,457 (97.1%) 1,643 (96.8%)	44 (2.9%) 55 (3.2%)	1,501 1,698
	Awarded scholarship	50 (4.9%) 46 (4.3%)	973 (95.1%) 1025 (95.7%)	1,023 1,071
	Column totals	1,507 1,689	1,017 1,080	2,524 2,769

Note. The top row of each column is for 1994-1995. The second row is for 1995-1996.

the number and percentages reported in the top row of each column are for 1994-1995; the numbers and percentages in the second row are for 1995-1996. First, as expected, the degree of agreement across the years is consistent. Second, as expected, the degree of agreement is quite high; the overall agreement in scholarship decisions was 96.3% for 1994-1995 and 96.4% for 1995-1996.

Discussion

The first purpose of the present study was to examine and compare the accuracy of, and agreement between, achievement scores generated using the CTST model and the IRT Generalized Partial Credit Model (GPCM, Muraki, 1992) where the examinations to be analyzed contained both dichotomously and polytomously scored items. Comparison of examinee score estimates derived from the shortened examination and from the full examination provides a measure of the accuracy of the procedure employed to obtain the estimates: the estimates from both forms should be equal except for sampling error. Based on the results, the CTST and GPCM estimation procedures yielded scholarship scores that closely agreed and were equally inaccurate. The correlations between the two procedures were similar to those found in other studies although with more variability (Anderson, 1999; Fan, 1998). Further, the amount of inaccuracy varied across examinations. The highest error rates occurred in geography, which contained the highest proportion of ER items. However, examinations that contained lower ER proportions also exhibited high error rates. Thus it appears that other factors are affecting the rate of classification errors, and research is needed to identify these factors. Although use of either the CTST model or the IRT GPCM model yielded scores on the shortened examination that resulted in essentially the same scholarship decision (96% agreement), for approximately one out of 10 students the scholarship decisions based on the shortened examination and using either of the two models did not agree with the initial decision made using scores derived from the full-length examination and using a classical test score theory approach.

Although Samejima (1996) reported that the use of response patterns to determine student ability produced results superior to those produced using the total score for simulated examinations consisting of polytomously scored items, the results of the present study indicate that the use of the GPCM with mixed-item format examinations will not provide superior estimates of student ability than did the sum of the item scores (total scores). Seemingly, differences in response vectors for students with the same observed score were related more to random differences among the students and less to differences in ability. Replacement of the CTST model with the more complex GPCM model cannot be justified in the present situation. However, because the students included in the agreement analyses included only higher-achieving students interested in obtaining a scholarship, further study is required to determine if such similarities also exist for students of lower ability.

The common lack of accuracy is probably attributable to the exclusion of the conceptually and cognitively more difficult scholarship examination that comprises only ER items included as part of the gold standard scholarship examination. No account was made of this exclusion in the shortened scholarship examinations. The number and type of items included in the provincial mandatory examination have not changed. Consequently, as expected, some students who received scholarships using the gold standard procedure did not receive scholarships using the CTST and GPCM procedures, whereas other students who did not receive a scholarship using the gold standard procedure did with the CTST and GPCM procedure. On the other hand, the increased false negative error rate is more likely to be associated with the current practice of using only provincial examination scores of at least 70% to calculate the scholarship scores. This procedure has in effect removed previously lower-achieving students who wrote the optional scholarship scores and thus increased the number and range of scores at the lower end of the scholarship score distribution.

Clearly the definition of exemplary academic performance changed with the introduction of the current CTST procedure. Consequently, for the two years and 11 examinations considered, approximately 350 students who were awarded a scholarship based on performance on the gold standard full examination would have been denied a scholarship based on performance on the current shortened examination. In contrast, approximately 175 students who were denied a scholarship based on performance on the gold standard would have been awarded a scholarship based on performance on the shortened examination. Further study is needed to clarify what constitutes academic scholarship and what changes are needed so that the scores obtained can be validly interpreted as indicative of academic excellence commensurate with the performance expected of a scholarship winner.

Note

1. The Generalized Partial Credit model is given by the probability that examinee i with a given θ will achieve category k on item j, ($P_{jk}(\theta_i)$:

$$P_{jk}(\theta_i) = \frac{\exp\left[\sum_{v=1}^{k} a_j(\theta_i - b_{jv})\right]}{\sum_{c=1}^{m_j} \exp\left[\sum_{v=1}^{c} a_j(\theta_i - b_{jv})\right]},$$

where m_j is the number of possible score categories,
c denotes the score categories,
$c = 1, 2, \ldots, m_j$,
b_{jv} is the item category or threshold parameter, the ability at which a category score of k or k-1 is equally likely, thus representing the "difficulty" in obtaining a subsequent score category, and
a_j is the discrimination (slope) parameter, which when combined with the set of threshold parameters determines the discrimination of the item if there are more than two score categories (Muraki, 1992).

Acknowledgments

The authors would like to acknowledge the feedback of the anonymous reviewers in helping to provide clarity in those parts of the article that were unclear or unnecessarily complex. We believe their comments have helped to make the article more readable for the general academic audience.

References

Anderson, J.O. (1999). Does complex analysis (IRT) pay any dividends in achievement testing? *Alberta Journal of Educational Research, 45*, 344-352.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Bock, R D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Bock, R.D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement, 34*, 197-211.

Carlson, J.E. (1996, April). *Information provided by polytomous and dichotomous items on certain NAEP instruments*. Paper presented at the annual meeting of the American Educational Research association, New York.

Cattell, R.B. (1952). *Factor analysis*. New York: Harper & Brothers.

Chambers, J.M., Cleveland, W.S., Kleiner, B., & Tukey, P.A. (1983). *Graphical methods for data analysis*. Belmont CA: Wadsworth International Group; Boston, MA: Duxbury Press.

Cheliminsky, E., & York, R.L. (1994). Educational testing: The Canadian experience with standards, examinations, and assessments. *General Accounting Office Report PEMD-93-11*. Gaithersburg, MD: General Accounting Office.

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

Donoghue, J.R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement, 31*, 295-311.

Fan, X. (1998). Item response theory and classical test score theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*, 357-381.

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.

Folske, J.C., Gessaroli, M.E., & Swanson, D.B. (1999, April). *Assessing the utility of an IRT-based method for using collateral information to estimate subscores*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley & Sons.

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

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

Klinger, D.A. (2000). *Recovery of decision consistency in shortened and modified scholarship examinations*. Unpublished doctoral dissertation, University of Alberta.

Klinger, D.A., & Boughton, K.A. (2000, May). *The accuracy of the generalized partial credit model and the graded response model in performance based assessments*. Paper presented at the annual meeting of the Canadian Society for Studies in Education, Edmonton.

Lafleur, C., & Ireland, D. (1999). *Canadian and provincial approaches to learning assessments and educational performance indicators*. Technical report submitted to Commonwealth Caribbean Program, Americas Branch: Canadian International Development Agency.

Lord, F.M. (1952). A theory of test scores. *Psychometric Monographs, No. 7.*

Lord, F.M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517-548.

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

Maydeau-Olivares, A., Drasgow, F., & Mead, A.D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. *Applied Psychological Measurement, 18*, 245-256.

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

Muraki, E., & Bock, R.D. (1997). *PARSCALE 3.0: IRT item analysis and test scoring for rating-scale data* [Computer program]. Chicago, IL: Scientific Software International.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses: Comparison of different approaches. *Journal of Educational Measurement, 31*, 17-35.

Nandakumar, R., & Stout, W.F. (1993). Refinements of Stout's procedure for assessing latent trait dimensionality. *Journal of Educational Statistics, 18*, 41-68.

Reise, S.P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133-144.

Rogers, W.T., & Ndalichako, J. (2000). Number right, item response, and finite state scoring models: Robustness with respect to the lack of equally classifiable options and item option independence. *Educational and Psychological Measurement, 60*(1), 5-19.

Samejima, F. (1996, April). *Polychotomous responses and the test score*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Stout, W.F. (1987). A non-parametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

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

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

Tomkowicz, J., & Rogers, W.T. (2001, April). *Effect of a subject area and a scoring model on ability estimates obtained from testwise susceptible and non-susceptible items*. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle.

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