The Alberta Journal of Educational Research *Vol. XLV, No. 4, Winter 1999, 344-352*

John O. Anderson
University of Victoria

Does Complex Analysis (IRT) Pay Any Dividends in Achievement Testing?

The study was an exploratory investigation of the consequences of using a complex test-and-item analysis approach in a large-scale testing situation that historically has used a conventional approach of simple number-right scoring. In contemplating modifications to a complex, high-stakes testing program that has a long history of successful operation, any change in operations would have to be carefully evaluated to ensure that there is a high probability of improvement through change. So if a change from number-right-type scoring to item response theory (IRT) scoring is under consideration, the question arises: Does the increase in complexity and difficulty associated with the use of IRT pay significant dividends in better achievement estimates? In terms of consequences, it did not make much difference which domain score estimate was selected for use: any estimate gives approximately the same results in terms of mean, standard deviation, error of estimation, and correlation to other sources of estimation of student achievement.

Cette recherche a étudié, à titre exploratoire, les conséquences d'employer une analyse complexe du test et des items dans un contexte d'évaluation à grande échelle où on avait l'habitude d'employer une approche traditionnelle selon laquelle le score représente le nombre de bonnes réponses fournies. Pendant le processus d'envisager des modifications à un programme d'évaluation complexe à enjeu considérable qui fonctionne bien depuis plusieurs années, il est important d'évaluer soigneusement tout changement pour se donner les meilleures chances possibles que l'adaptation mène à une amélioration. Ainsi, avant de substituer une évaluation qui reflète le nombre de bonnes réponses fournies par une qui repose sur la théorie de la réponse d'item (TRI), il faut se poser la question suivante: Est-ce que l'accroissement de complexité et de difficulté lié à l'emploi de la TRI mène à de meilleures estimations quant à l'évaluation? Pour les conséquences, il importait peu quel test de rendement on étudiait: toute estimation donne à-peu-près les mêmes résultats en ce qui concerne la moyenne, l'écart type, l'estimation de l'erreur et la corrélation avec les autres sources d'estimation du rendement des élèves.

Educational tests are a main source of information about student achievement in schools. In the context of large-scale testing, the analysis of test data is essential in the production of student scores and grades and in monitoring and evaluation of the quality of the test and the information the test generates. There are two main approaches to this analysis: conventional (classical) analysis and item response theory analysis. Conventional analysis is based on classical test theory (Gulliksen, 1950). This approach to test analysis has a long history of use, and results are easy to understand. It is relatively simple in terms of computations, and software programs are standard in that they produce the same results for a given analysis. It uses number-right scoring to produce raw scores, percentage scores, and scaled scores such as standard

---

John Anderson completed his doctorate under the supervision of Tom Maguire and is now the Chair of the Department of Educational Psychology and Leadership Studies. He teaches measurement and evaluation.

scores. Item statistics generated through conventional analysis are the item difficulty (the *p*-value) and an item discrimination index (most often a correlation between the item score and the test score). This analytic approach makes no claims about the generalizability of item and test statistics beyond the test administration that generated the response data.

Item analysis based on item response theory (IRT) models item responses as a function of both item and person characteristics (Lord, 1980). This is logical and reasonable given that in the context of achievement testing the difficulty of an item and the achievement level of a respondent should interact to influence performance on the item. For example, the more difficult an item is, the less likely it is a student will give a *good* (correct) response; whereas the higher the achievement level of the respondent, the more likely he or she will give a *good* response. Given this reasonableness and the inclusion of more information into the estimation procedure, the resulting estimate of achievement level or domain score should be better than a simple raw score on the same set of items. Further, as has been pointed out on more than one occasion (Hambleton & Swaminathan, 1985; Lord, 1980), IRT has significant implications in an item banking situation in that the item characteristic estimates are independent of the sample of respondents, and the ability (achievement) estimates are independent of item sample. This means that in estimating item characteristics (*item calibration* in IRT parlance) the estimates derived from one sample of students are comparable to estimates derived from another sample of students, and these estimates can be scaled onto the same metric. This means that given a calibrated bank of items, an ability estimate for a student can be obtained from any set of items. Further, the score is on the same metric and is therefore comparable to a score obtained from any other set of items. But the use of IRT-based scoring and analysis software is much more complex than conventional number-right scoring, requiring well-trained and experienced personnel to conduct the scoring, maintain the system, and interpret the results. The state of software available for IRT analysis and scoring is not standard in that different approaches (e.g., a one-parameter model versus a two- or three-parameter model, or Bayesian versus maximum-likelihood analytic algorithms) can result in different scores and item characteristics. This means that the choice of IRT model and underlying mathematics influences student results.

Few studies have reported using empirical response data from large-scale achievement testing applications that evaluate the extent to which the more complex IRT-based analysis has generated superior results. Fan (1998) compared the item and person statistics generated by classical and IRT analyses of the Texas Assessment of Academic Skills, a large-scale assessment administered to grade 11 students. It was found that the achievement estimates for the different analyses were similar. The item difficulty estimates were not only similar across the analyses, but the invariance of estimates (using different samples of students to generate difficulty estimates) was similar for classical and IRT approaches.

Rogers and Ndalichako (in press) evaluated the robustness of conventional scoring and IRT scoring of response data from the administration of a school-leaving examination of reading comprehension to over 1,200 high school students. They found that results of conventional, and one- and two-parameter

IRT scoring were very similar. Three-parameter IRT results showed an adverse sensitivity to testwiseness. They concluded that,

> When coupled with the relative simplicity of conventional number-right scoring and the relative ease of explaining to students and their teachers and parents how the scores are obtained, the findings of this study support the continued use of number-right scoring. (p. 13)

In a contrived setting using spelling tests with university students, Bock, Thissen, and Zimowski (1997) evaluated domain score estimates with empirical results from a bank of 100 spelling items. The domain score (π) was defined as the proportion number-right score on the 100 items. From this bank of 100 items (the domain), they created numerous items sets of various sizes to calculate both raw score and a two-parameter IRT estimates of domain scores. They evaluated the comparability of the results on the basis of the root mean squared errors. They found that the IRT-based results were generally superior to the raw scores in estimating student domain scores. Do these results replicate in a real-world setting with more difficult and complex items? The basic idea of the Bock et al. study and the study reported here is to simulate the situation in which students are completing more than one test for a particular curriculum using existing response data. Each *test* is created by selecting a specific subset of items from the original larger test and the associated response data available for each student. Each *test* thus created would yield an estimate of student achievement. Because the underlying trait (achievement) would remain constant and the items would vary as would the scoring and analysis, this would allow for comparison of the scores from the various created *tests*.

The Study

This study is a small-scale exploratory investigation of the consequences of using a complex test and item analysis approach in a situation that historically has used a conventional approach of simple number-right scoring. In contemplating modifications to a complex, high-stakes testing program that has a long history of successful operation, any change in operations would have to be carefully evaluated to ensure that there is a high probability of improvement through change. Any changes in the program would need to be carefully monitored for any effects on the quality of the system as students, parents, and teachers carefully scrutinize the results. The results are an important component of the final graduating grades awarded to the students in that subject. So if a change from number-right-type scoring to IRT scoring is under consideration, the question arises: Does the increase in complexity and difficulty associated with the use of IRT pay significant dividends in better achievement estimates?

The high school graduation examinations in British Columbia are large-scale, high-stakes tests of achievement. The examination program has been in operation for the past 15 years. Exams are created three to five times annually in over 15 different subject areas. For each administration, each exam is developed to match curriculum specifications, and committees of teachers are actively involved in the creation, marking, and standard-setting of all examinations. The examinations are typically administered over a two and a half hour time period, and consist of 50 or more items of both multiple-choice

and open-ended formats. Exam results are based on number-right scoring and conventional (classical) test analysis. High school graduation examinations can be viewed as means to estimate the achievement levels of students in the subject—to classify students into achievement categories. The examinations are designed on the basis of a table of specifications that is a description of the content and processes representative of the course: a blueprint of the achievement domain. The letter grades and percentage marks reported are directly related to domain scores in that the higher the grade or mark, the higher the proportion of the domain the student has mastered.

The study is based on empirical response data from the British Columbia Provincial Examination in Mathematics 12. The Mathematics 12 Exam is a graduation exam consisting of both multiple-choice and constructed response items. The response data came from over 6,000 students completing the 50 multiple-choice items from the January 1996 administration of the exam. In addition to the raw scores from the Mathematics 12 examination, the data set also included a School score, a Provincial score, and the reported Mathematics 12 Exam score. The School score is the final Mathematics score that each teacher submits for each student. The Provincial score is the final score the student receives on his or her record of graduation. It is based on the School score submitted by the teacher and the Mathematics 12 Exam score. It should be noted that the reported Mathematics 12 Exam score is not a raw score. Although it is based on exam raw scores, it has been modified by standard setting committees and then scaled onto a 100-point scale on which the cutpoints are: A: x_i>85, B: x_i=73 to 85; C^+: x_i=67 to 72; C: x_i=60-66; C^-: x_i=50 to 65 and F: x_i<50.

For this study the 50 multiple-choice items from the Mathematics 12 Exam served as the *item bank* from which student domain scores would be estimated by both conventional raw score and by IRT-based analyses. In order to compare analyses, two samples of 25 items were selected: odd-numbered items in one sample and even-numbered items in the other. The two *tests* created were not equivalent in that an arbitrary split of items was used. Student responses were scored using the two analytic approaches. Evaluation of item fit for the IRT analysis was conducted for the 1-, 2-, and 3-parameter IRT analyses using BILOG 3 (Mislevy & Bock, 1996). Because the 3-parameter IRT model resulted in the highest level of fit, it was used in all subsequent analyses based on IRT.

The scores generated by conventional analysis were the raw score for each created test and these, expressed as proportion scores, represented one set of domain score estimates. The IRT scoring was conducted using the item parameter estimates from the analysis of the 50-item set of data. The scores generated were the theta scores with a mean of 0.0 and standard deviation of 1.0. Using the item statistics from the item bank (the 50-item set) and the theta score from each 25 item test for each student, the domain score was estimated:

$$\pi_p = \frac{\sum p_j(\theta_p)}{n}$$

where π_p is the domain score for person p, $p_j(\theta_p)$ is the probability of person p with achievement level θ correctly responding to item j, and n is the total number of items comprising the test. The resulting achievement estimates (domain scores) were compared with each other, with the Mathematics 12

Exam score, with estimates generated independently from teachers, with the Provincial score, and with the resulting scores on the total 50-item set.

Because the major outcome of the Mathematics 12 examination is to assign students to appropriate achievement categories—letter grades—each domain score was used to assign the student a letter grade by using the provincial cutpoints described above. This assigned letter grade was then compared with the letter grade awarded the student by the BC Ministry of Education for Mathematics 12, and with other domain score estimates. It should be noted that the Mathematics 12 Exam score is not the Provincial score that incorporates both the Examination score and the School score, but rather the grade based solely on Mathematics 12 Exam results modified by the standard setting procedures.

The Results

The summary statistics (Table 1) indicate the mean domain scores are similar for all item sets regardless of which analysis was used. The small differences that are evident are larger within analysis type than within item set. The mean score on item set 1 is 0.669 for both conventional scoring and IRT scoring. The mean for item set 2 is 0.631 for conventional scoring and 0.632 for IRT scoring. The correlations between scores from the same item set (X_1 : IRT_1, and X_2 : IRT_2) are approximately 0.97, whereas the correlations between scores derived from the different items sets but the same analysis (X_1 : X_2 and IRT_1 : IRT_2) are lower at approximately 0.77. This suggests that the domain scores are more closely related to the item set from which they were derived than to the analysis used to derive them. The correlations of domain scores to related measures (Table 2) show consistency between the estimates. The relationship to the School score is relatively low (approximately .28), higher for the Provincial score (~.79), and higher yet for the reported Mathematics 12 Exam score (~.87).

Table 1
Summary Statistics for Domain Scores

	Conventional		*IRT*	
	Test 1 X_1	*Test 2* X_2	*Test 1* IRT_1	*Test 2* IRT_2
Mean	0.669	0.631	0.669	0.632
Standard Deviation	0.169	0.177	0.141	0.151
Minimum	0.120	0.040	0.368	0.317
Maximum	1.000	1.000	0.928	0.952
Correlations				
X_1				
X_2	0.783			
IRT_1	0.971	0.775		
IRT_2	0.766	0.977	0.765	
n = 6,147				

Table 2
Correlations to Related Measures

Estimate	*School Score*	*Reported Math 12*	*Provincial Score*
X_1	0.289	0.873	0.794
X_2	0.295	0.884	0.804
IRT_1	0.283	0.858	0.783
IRT_2	0.281	0.860	0.783

Table 3
Root Mean Square Errors

Estimate	*RSME*
X_1	0.060
X_2	0.060
IRT_1	0.067
IRT_2	0.065

The root mean square errors (RMSE) were calculated for each set of scores using the percent correct score of the whole 50-item set as the domain score and the scores from the 25-item sets as the estimates:

$$RMSE = \sqrt{\frac{\sum (estimate - domain)^2}{n}}.$$

The results (Table 3) show little difference between the four estimated domain scores. All root mean square errors are approximately the same (average = 0.063).

Each domain score estimate was converted into a letter grade (6 = A to 1 = Fail) and compared with the letter grade actually produced by the Mathematics 12 Exam procedures of the BC Ministry of Education. The summary statistics (Table 4) once again suggest that the results from both 25-items sets and from both analyses are similar. As previously noted the Mathematics 12 Exam results are derived from the exam raw scores (from both multiple-choice and open-ended items) that have been modified through the incorporation of committee-set standard cut points and then rescaled onto a 100-point reporting

Table 4
Summary Statistics for Letter grades

	Mean	*Standard Deviation*	*Correlation to Mathematics 12*
Mathematics 12 Exam	3.50	1.72	
X_1	3.35	1.55	0.86
X_2	3.02	1.57	0.87
IRT_1	3.48	1.63	0.85
IRT_2	3.08	1.67	0.87

Table 5
Distributions of Differences in Student Classification
(Mathematics 12 Letter Grades)

Estimates	Percentage Distribution of Differences						
	<–2	–2	–1	0	+1	+2	>+2
X_1	0.2	4.5	27.6	48.9	15.9	2.6	0.1
X_2	0.5	9.2	37.3	44.8	7.4	0.4	0.3
IRT_1	0.4	3.7	22.0	50.6	19.0	3.7	0.7
IRT_2	1.5	9.1	31.7	47.4	8.8	1.1	0.4

metric. This score was then used as the domain score because it is the best available target domain score.

To further explore the classification of students, the differences between the 25-item estimates and the Mathematics 12 letter grades were taken as an index of consistency of classification. The distribution of differences to the Mathematics 12 letter grade (Table 5 and Figure 1) again show that estimates are similar in categorizing students into one of the six letter grade categories. The 25-item multiple-choice tests (item set 1 or item set 2) result in the same pattern of letter grade assignment whether based on conventional or IRT analysis. On average about 48% of student classifications with the Mathematics 12 letter grades are matched whether one analysis is compared with another or one item set with another item set. Looking at classifications that are within one letter grade, 91% of classifications are consistent within (1 letter grade of the Mathematics 12 results). These difference values provide a view of error of estimate that is similar to the RMSE. However, they are different in that they report both the magnitude and the direction of differences, and they are directly related to the application under consideration—the assignment of students to achievement categories.

It may be noted in Table 5 that the item sets tend to result in lower letter grade assignments than does the reported Mathematics 12 Exam score. This is probably due to the fact that the Exam score is dependent not only on the

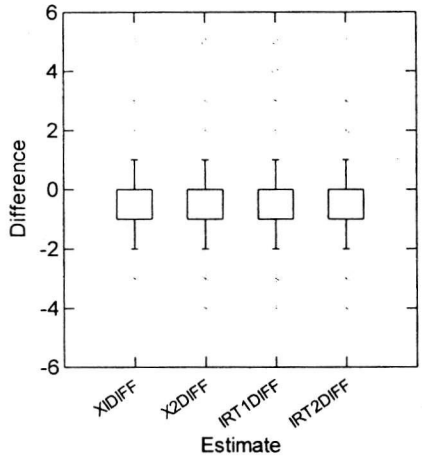

Figure 1. Estimates—Mathematics 12 Exam letter grades.

Table 6
Letter Grade Consistency Between Item Sets and Between Analyses

Comparison	*Mean*	*Standard Deviation*
Between Analyses		
X_1 - IRT_1	–0.129	0.536
X_2 - IRT_2	–0.059	0.503
Between Items		
X_1 - X_2	0.332	1.091
IRT_1 - IRT_2	0.402	1.222

multiple-choice items on the exam but also the open-ended items and this is further modified by the standard setting procedures.

To continue the exploration of the consistency of classification of students, the letter grades derived from different analyses using the same set of items were compared as difference scores (X_1 - IRT_1 and X_2 - IRT_2), and letter grades derived from different sets of items using the same analysis were likewise compared (X_1 - X_2 and IRT_1 - IRT_2). The results (Table 6 and Figure 2) clearly show that there is greater consistency (less difference in letter grade assignment) between the same items using different analyses than between different items using the same analysis.

Conclusion

This study explored the use of item response theory-based analysis with graduation exam response data, comparing results with those produced with conventional number-right scoring and analysis. Because the item pool consisted of only a small number of items (n=50) and only two samples of items were selected for investigation, the results should be interpreted cautiously. However, the items and the response data are characteristic of an operational large-scale, high-stakes testing program, and therefore the findings should have relevance to applied achievement testing.

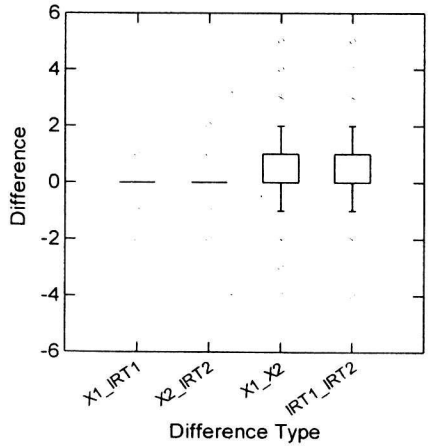

Figure 2. Differences between letter grade estimates.

The main finding is the striking consistency in the results. The results from either analysis were almost indistinguishable, particularly in regard to the assignment of letter grades to students. Another evident pattern was that there were more differences between item sets than between analyses. Although differences were expected between item sets because the division of items (an odd-even split) was rather arbitrary and unlikely to result in parallel tests, there was some expectation that IRT should yield more accurate domain score estimates than those from raw score procedures. In terms of consequences it did not make much difference which domain score estimate was selected for use: any estimate gives approximately the same results in terms of mean, standard deviation, error of estimation, and correlation to other sources of estimation of student achievement.

References

Bock, R.D., Thissen, D., & Zimowski, M.F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement*, 34(3), 197-211.

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

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

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

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

Mislevy, R.J., & Bock, R.D. (1996). *BILOG 3.09.* [Software program]. Chicago, IL: Scientific Software International.

Rogers, W.T., & Ndalichako, J. (in press). Number right, item response and finite state scoring: Robustness with respect to lack of equally classifiable options and item option independence. *Educational measurement: Issues and Practice.*