The Alberta Journal of Educational Research *Vol. 51, No. 3, Fall 2005, 200-215*

Joanna Tomkowicz
CTB-McGraw-Hill, Monterey, CA

and

W. Todd Rogers
University of Alberta

The Use of One-, Two-, and Three-Parameter and Nominal Item Response Scoring in Place of Number-Right Scoring in the Presence of Test-Wiseness

Ability estimates yielded by the one- (1PL), two- (2PL), and three-parameter (3PL) models and the nominal response model (NRM) were compared with the number-right (NR) scoring model using items not susceptible to test-wiseness (NTW) and items susceptible to the ID1 test-wiseness strategy. These items were contained in grade 12 diploma examinations for social studies and chemistry. The results were compared for high-, middle-, and low-ability examinees. Differences were found between pairs of ability estimates obtained when 2PL, 3PL, and NRM scores were used in place of NR scores. The differences tended to be greater for chemistry than for social studies, and with the exception of high-ability students in social studies, for the subtest containing items with absurd options than for the subtest containing nonsusceptible test-wise items. It appears at least for the two subject areas considered in the present study, that the scoring models cannot be used interchangeably to obtain estimates of examinees' abilities, particularly when a test contains test-wise susceptible items.

En nous appuyant à la fois sur des items qui ne sont pas susceptibles d'être influencés par des paramètres de discrimination et de pseudo-chance, et à des items qui le sont, nous avons comparé les estimations de niveau d'habileté découlant de modèles à un (1PL), deux (2PL) et trois (3PL) paramètres ainsi que du modèle à réponses nominales (NRM) à celles provenant du modèle basé sur le nombre de bonnes réponses (NR). Ces items étaient présents dans les examens du ministère en études sociales et en chimie pour la 12e année. Nous avons comparé les résultats des élèves de différents niveaux d'habileté (bas, moyen et élevé). Des écarts se sont manifestés entre des paires d'estimations d'habileté quand nous avons analysé des scores reposant sur les modèles 2PL, 3PL et NRM plutôt que les scores NR. De façon générale, les différences étaient plus importantes: (a) en chimie qu'en études sociales et, (b) à l'exception des élèves à capacité élevée en études sociales, dans le sous-test comprenant des réponses absurdes que dans celui avec des items qui ne sont pas susceptibles d'être influencés par des paramètres de discrimination et de pseudo-chance. Du moins pour les deux domaines à l'étude, il semblerait que les modèles de pointage ne sont pas interchangeables dans l'évaluation des capacités des élèves, surtout lorsqu'un examen contient des items susceptibles d'être influencés par des paramètres de discrimination et de pseudo-chance.

Joanna Tomkowicz is a research scientist whose research interests include large-scale assessments, psychometrics, and applied statistics.
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.

Introduction

The multiple-choice test is one of the more popular test forms used to assess students' academic achievement. This is particularly true at the higher grade levels and in large-scale testing programs. Quick and objective scoring is often cited as a factor leading to the preference for multiple-choice tests over other forms of assessment (e.g., performance assessment, Aiken, 1987; Bennet & Ward, 1993; Hambleton & Murphy, 1992).

Various scoring methods have been developed to score multiple-choice items and estimate examinees' abilities. The first of these is number-right scoring associated with the classical test score model. In response to limitations of the classical test score model, Lord (1980) proposed the one-, two-, and three-parameter item response theory (IRT) models. Like number-right scoring, the scoring method associated with these item response models requires that the items be scored dichotomously. Several researchers, however, perceive dichotomous scoring as deficient because it implicitly assumes that examinees act according to the "knowledge-or-random guessing" principle (Lord, 1980). That is, examinees either have the knowledge to answer an item correctly or simply randomly select their answers from among the alternatives provided. However, several researchers (De Ayala, 1989, 1993; Lord, 1980; Rogers & Bateson, 1991a; Tatsuoka, 1983) have found that examinees who do not possess the necessary knowledge to answer an item use their partial knowledge about the item content to select their responses. It is reasonable to presume, however, that some examinees may possess only part of the knowledge necessary to select the correct answer and that they may use this partial knowledge to choose a particular incorrect alternative (De Ayala; Lord). As pointed out by Tatsuoka and others (Brown & Burton, 1978; Jacobs & Vandeventer, 1970; Lane, Stone, & Hsu, 1990) "wrong responses can be more than just one kind, although the binary scoring procedure uniformly assigns a score of zero to all the wrong responses" (Tatsuoka, p. 346).

Bock (1972) proposed the nominal response model to take account of this partial knowledge. With this model, information from both the correct and incorrect responses is used to estimate examinee ability. Bock (1972, 1997), Levine and Drasgow (1983), Thissen (1976), Thissen and Steinberg (1984), and Thissen, Steinberg, and Fitzpatrick (1989) demonstrated that using this additional information yields more precise ability estimates, particularly for examinees whose scores fall below the mean ability level where incorrect responses occur more frequently.

Another factor that has been known to affect ability estimates is test-wiseness. Test-wiseness has been defined as an examinee's cognitive capacity to utilize the characteristics and formats of the test and/or the test-taking situation to improve a test score. Examinees who do not know the correct answer to a test question but possess both test-wiseness and relevant partial knowledge have a greater probability of selecting a correct response than examinees who possess partial knowledge but not test-wiseness or who have knowledge of test-wiseness principles but low partial knowledge of item content (Diamond & Evans, 1972; Millman, 1966; Rogers & Bateson, 1991a; Rogers & Yang, 1996; Towns & Robinson, 1993). The most common types of test-wiseness clues found in multiple-choice items include absurd options, stem-option associa-

tion, similar and opposite options, and options containing specific determiners (Crehan, Koehler, & Slakter, 1974; Diamond & Evans; Hughes, Salvia, & Bott, 1991; Millman; Rogers & Bateson; Rogers & Wilson, 1993; Sarnacki, 1979). Absurd options are options known by most examinees to be incorrect. Students knowledgeable of this strategy and possessing relevant partial knowledge avoid absurd options and choose from among the remaining ones. Stem-option association allows examinees to recognize and make use of a resemblance between an option and an aspect of the stem. Similar options tend to be considered by an examinee simultaneously, and given that there is only one correct response, both options are rejected. In contrast, opposite options will guide a skilled test-taker toward choosing neither or one (but not both) of two options, one of which, if correct, would imply the incorrectness of the other. Recognizing and making use of a specific determiner included in an option has also been found helpful in distinguishing the correct answer from incorrect alternatives (see Rogers & Yang, 1996, for a more complete review of test-wiseness).

Given these findings, a question arises about the relevance of partial knowledge and test-taking skills when determining ability estimates. If the partial knowledge is considered not to be relevant to the individual ability estimate, then dichotomous scoring of students' responses to multiple-choice items may be warranted. However, if partial knowledge is considered relevant, then a total test score that takes into account the additional information coming from incorrect responses may be a more valid indicator of performance (Messick, 1989). For example, Levine and Drasgow (1983) demonstrated that at least for some items, examinees at varying levels of ability tended to have varying patterns of wrong responses (i.e., very able examinees differed from low-ability examinees in their patterns of wrong responses). Further, if a test contains test-wise-susceptible items and if examinees employ both test-wiseness and partial knowledge, then the distribution of incorrect responses to these items may differ from the distribution of incorrect responses to the items that are not test-wise-susceptible (Rogers & Bateson, 1991b). Subsequently, when information from wrong responses is taken into account, this may affect their ability scores (Nedelsky, 1954; Rogers & Ndalichako, 2000; Rogers & Yang, 1996).

Consequently, the question addressed in the present study was: If we were to use any one of the item response scoring models—one-, two-, and three-parameter IRT, and nominal response—to score a test containing items susceptible to test-wiseness and a test containing items not susceptible to test-wiseness, would we obtain the same estimate of examinee ability yielded by the commonly used number-right scoring model? The number-right scores were taken as the base for comparisons as this was the score reported to students and others with a legitimate right to know for the examinations considered in the present study. Because several researchers (Bock, 1972; Levine & Drasgow, 1983; Rogers & Bateson, 1991b; Thissen, 1976; Thissen & Steinberg, 1984) have found an examinee's ability level may probably be a factor associated with selection of a particular item option, this question was answered separately for high-, middle-, and low-ability examinees.

Method

Data

Two data sets were used. Each set consisted of responses of high school students to the multiple-choice items contained in the grade 12 school-leaving examinations for social studies and chemistry administered at the end of June 1999. These high-stakes tests count for 50% of a student's final grade and are intended for students who are planning or wish to leave open the opportunity to pursue some form of tertiary education (Alberta Learning, 1999a, 1999b).

Identification of test-wise susceptible items. Two panels of experts, one for each subject area, were formed to analyze the items in each test for the presence of test-wise cues. The panel for social studies included two members who were currently teaching high school social studies, and the remaining two were graduate students who before beginning graduate school had taught high school social studies. The panel for chemistry included one current high school chemistry teacher, one just-retired high school chemistry teacher, and one graduate student who had previously taught chemistry at the high school level. Following the procedures used by Rogers and Bateson (1991b) and Rogers and Wilson (1993), the test-wise strategies identified above were explained and illustrated using examples from the study conducted by Rogers and Wilson. Following this training, the panel members working alone first identified items that they believed contained one or more test-wise-susceptible options and items that did not. Then each member was shown the results of the classical item analyses of the items and asked to use the results to verify their initial identifications and to see if there were any other items that they felt contained test-wise susceptible options. For social studies, three of the four panel members agreed on 44 of the 70 items considered; for chemistry, two of the three panel members agreed on 42 of the 44 items considered. A possible reason for the difference between the two rates of agreement is ascribable to the nature of the curriculum: social studies involves the use of personal and societal values in contrast to chemistry, which involves fewer of these values. The panel members then met to discuss their individual findings. Following discussion of each item by each panel, complete agreement was achieved among the respective panel members for all 70 items in the social studies test and all 44 items in the chemistry test. The proportions of items identified with the test-wiseness elements considered in the present study closely matched those found by Rogers and Wilson. Further, as in Rogers and Bateson and Rogers and Wilson, the most common test-wiseness element found by both panels was the absurd option. The numbers of items sensitive to other test-wiseness elements (e.g., similar options, opposite options, stem-option connection, and specific determiners) in both examinations were not sufficient to obtain stable ability estimates.

Consequently, two subtests were identified in each diploma examination for further analyses. For the social studies examination, one subtest consisted of 35 items not susceptible to any test-wiseness element (SS-NTW), and the second contained 23 items susceptible to the absurd test-wiseness strategy: "eliminate option(s) that are known to be incorrect and choose from among the remaining alternatives" (SS-ABS). For the Chemistry 30 Diploma Examination, one subtest contained 21 items that were not susceptible to test-wiseness (CH-

NTW), and the other subtest contained 14 items that were susceptible to the absurd option strategy (CH-ABS). For both subject areas, the distribution of the items in the subtest not susceptible to test-wiseness and the items in the subtest susceptible to the application of the absurd test-wiseness strategy were similarly distributed across the topic-by-level of thinking cells contained in the table of specifications for the subject area.

Student Samples

The total numbers of students that completed the social studies and chemistry examinations were respectively 10,905 and 8,594. Two samples of 4,000 students each were randomly drawn without replacement from the total number of examinees for each test. The initial data analyses were performed using one sample for each test. The initial analyses were then replicated using the second sample for each test to assess the stability of the results.

Analyses

Number-right scores were obtained for each of the four subtests using LERTAP (Nelson, 1983). The estimated abilities for the remaining four models were obtained in two stages. First, marginal maximum likelihood estimates of item parameters for the 1PL, 2PL, 3PL, and NRM models were obtained for each subtest. The four IRT models differ in terms of a number of item parameters they estimate. The 1PL model assumes that item difficulty is the only factor affecting student performance. The 2PL model allows items to vary in terms of their difficulty and discrimination. The 3PL model assumes that examinee performance is influenced by item difficulty, item discrimination, as well as examinee guessing behavior. The NRM is similar to the 2PL model in that variation in item difficulty and discrimination is allowed. However, whereas the 2PL model is applied to dichotomous data only, the NRM considers all the item options under the assumption that item alternatives are measured at a nominal level of measurement (Bock, 1972). As such, the NRM allows for the description of the relationship between each item alternative and the cognitive ability measured by the test.

Following item parameter estimation, maximum a posteriori (MAP) estimates of individual ability were computed using a Gaussian prior distribution in each model. Both stages were completed using MULTILOG (Thissen, 1991). An advantage of the MAP estimation is that it is defined for all response patterns, including patterns in which all items are answered correctly or all items are answered incorrectly (Thissen & Steinberg, 1997).

Before computing the item-response model estimates, the assumptions underlying the use of these models 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 test and the two subtests in each test were essentially unidimensional (Nandakumar, 1994). Fewer than 0.4% of the students did not complete the items in each test, indicating that speed was not a factor. The ranges of the uncorrected point-biserial correlation coefficients were too large to meet the assumption of equal discrimination needed for the one-parameter model (Hambleton & Murray, 1983). Further, use of the 1PL model revealed

relatively poor data-model fit for some items as indicated by the Chi-Square test (Wright & Stone, 1979). Despite not meeting the assumption of equal item discrimination and in some cases relatively poor item fit, the 1PL model was retained for data analysis for comparison purposes. In addition, the assumption of guessing behavior was tested. It was found that the lowest-ability students chose correct answers to difficult multiple-choice items at a rate that would be expected if they were guessing.

The analyses described above were completed for each subtest with the full sample of 4,000 students for each. The ability estimates obtained from each item response scoring method were compared with the number-right estimates in terms of the extent to which (a) the item response and number-right ability estimates provided similar rankings of students, (b) the values of the item response ability estimates agreed with the corresponding number-right estimate, and (c) the proportions of students who received an item response ability estimate higher and lower than the corresponding number-right estimate. The ability estimates were first transformed to *T*-scores (μ=50; σ=10) to express the estimates in the same metric (Ndalichako & Rogers, 1997; Rogers & Ndalichako, 2000). The ability estimates were then compared separately for high-, middle-, and low-ability students. Correlations among the scores in each ability group were computed for the rank comparisons. The root mean squares between pairs of scores and the proportions of students whose *T*-scores were greater than or less than one standard error of measurement for the number-right scores were computed to examine the closeness of scores in each group.

Formation of ability groups. Because an external measure of examinees' ability was not available, the examinees' total raw scores on the social studies and chemistry tests were used to create three ability groups. The means, standard deviations, skewness, and kurtosis for each subject area distribution for the total sample are reported in Table 1. These data reveal that the distributions for both subject areas were skewed negatively and were somewhat platykurtic. Therefore, although the guidelines provided by Kelley (1927) could not be strictly followed, they nevertheless were used as a starting point to form the three ability groups in each subject area.

Results

Because the results of item and ability estimation obtained in the initial and replication study were essentially the same for both examinations, the results presented and discussed in this article are based on the analysis for the initial sample.[1] The results for the items not susceptible to test-wiseness are presented in regular type; the results for the items susceptible to the absurd option test-wiseness strategy are reported in boldface type.

Item Analysis

Although the emphasis of the study was on the agreement among scores, item analyses were first conducted to assess possible differential effects on the item characteristics. Briefly, the results of the classical test model item analyses conducted for the subtest of items not susceptible to test-wiseness and of the subtest of items with absurd options revealed that the two subtests differed less in terms of their difficulty in the case of the social studies (0.64 vs. **0.68**) test than in the case of the chemistry test (0.65 vs. **0.75**). For both examinations, the two

Table 1
Properties of the Distribution of the Observed Scores for the Total Sample and Ability Groups

Sample	*Sample Size*	*Minimum Score*	*Maximum Score*	*Mean*	*Standard Deviation*	*Skewness*	*Kurtosis*
Social Studies							
Total Sample	4,000	11	70	46.96 67.1%	11.79	–0.36	–0.57
Low Ability	1,086	11	39	31.60 45.1%	5.96	–0.80	0.04
Middle Ability	1,141	44	53	48.54 69.3%	2.82	0.00	–1.19
High Ability	988	57	70	61.25 87.5%	3.22	0.48	–0.67
Chemistry							
Total	4,000	10	44	30.70 69.7%	6.76	–0.34	–0.47
Low Ability	1,117	10	26	22.07 50.2%	3.46	–1.05	0.60
Middle Ability	1,255	29	34	31.67 72.0%	1.65	–0.14	–1.16
High Ability	1,066	36	44	38.75 88.0%	2.12	0.47	–0.70

Note. For the initial sample. The properties for the replication sample were essentially the same.

subtests were comparable in terms of their item discriminations, and the distribution of wrong responses was more uniform across item foils of the items not susceptible to test-wiseness than of the items containing absurd options.

The results for the one-, two-, and three-parameter analyses were consistent with the results of the classical test model item analyses. The subtest of items not susceptible to test-wiseness and the subtest of items containing absurd options were found to be less different in terms of difficulty for the social studies examination than for the chemistry examination. For example, the b_i mean estimates obtained from the two-parameter model were –0.80 for the SS-NTW and –1.06 for the SS-ABS, and –0.82 for the CH-NTW and –1.76 for the CH-ABS. The distribution of item discrimination estimates and the pseudo-guessing estimates were similar across both subtests in both subject areas.

Examination of the item option trace lines indicated that the distracters appeared to be fairly attractive for examinees at a given proficiency level for both the SS-NTW and CH-NTW subtests. In contrast, the trace lines for the absurd options indicated a very low probability of selection or attractiveness to low-ability examinees only for both the SS-ABS and CH-NTW subtests.

Comparison of Ability Estimates

The results of the comparison of the ability estimates are reported in Table 2 for social studies and Table 3 for chemistry. The first panel in each of these tables contains the results for the high-ability group, the second panel the results for

Table 2
Comparison Between Ability Estimates: SS-NTW and SS-ABS

	θ_{NR}	θ_1	θ_2	θ_3	θ_{NM}
High-Ability Group (n=988)					
Mean (SD)	61.88 (3.56) **61.10 (3.96)**	62.54 (4.94) **61.67 (5.08)**	62.70 (5.01) **61.77 (5.11)**	62.39 (4.97) **61.68 (5.15)**	62.72 (4.95) **61.72 (5.04)**
$r_{\theta_{NR,j}}$	-	0.99 **0.99**	0.97 **0.96**	0.96 **0.95**	0.94 **0.94**
$RMS_{\theta_{NR}-\theta_j}$	-	1.60 **1.33**	2.00 **1.85**	1.92 **1.94**	2.18 **2.04**
$P_{greater}$	-	1.92 **0.00**	5.47 **0.00**	4.45 **2.33**	6.68 **3.44**
P_{less}	-	0.00 **0.00**	0.00 **0.40**	0.00 **0.71**	0.71 **1.21**
Middle-Ability Group (n=1,141)					
Mean (SD)	51.10 (3.77) **51.04 (4.92)**	50.31 (3.63) **50.44 (4.89)**	50.12 (3.70) **50.42 (4.82)**	50.43 (3.59) **50.49 (4.75)**	50.06 (2.77) **50.48 (4.80)**
$r_{\theta_{NR,j}}$	-	0.99 **0.99**	 **0.95**	0.95 **0.94**	0.94 0.91 **0.92**
$RMS_{\theta_{NR}-\theta_j}$	-	0.81 **0.69**	1.51 **1.67**	1.45 **1.76**	1.93 **2.04**
$P_{greater}$	-	0.00 **0.00**	0.00 **0.00**	0.00 **0.00**	0.00 **0.52**
P_{less}	-	0.00 **0.00**	0.18 **0.61**	0.35 **1.40**	3.51 **3.07**
Low-Ability Group (n=1,086)					
Mean (SD)	37.63 (5.49) **38.20 (6.62)**	38.23 (4.85) **38.62 (5.88)**	38.36 (4.69) **38.54 (5.61)**	38.18 (5.24) **38.50 (5.81)**	38.38 (4.58) **38.46 (5.59)**
$r_{\theta_{NR,j}}$	-	1.00 **1.00**	0.98 **0.97**	0.97 **0.96**	0.94 **0.94**
$RMS_{\theta_{NR}-\theta_j}$	-	0.89 **0.86**	1.51 **1.83**	1.53 **1.98**	2.06 **2.43**
$P_{greater}$	-	0.00 **0.00**	0.09 **1.29**	0.64 **2.39**	4.51 **4.70**
P_{less}	-	0.00 **0.00**	0.00 **0.09**	0.00 **0.92**	0.28 **3.22**

Note. Values in regular type are for the SS-NTW subtest and in boldface type for the SS-ABS subtest.

the middle group, and the third panel the results for the low-ability group. The means and standard deviations of the scores yielded by each scoring method are presented at the top of each panel followed by four rows containing respectively the correlations (row 1) and root mean squares (row 2) between each of the item response scores and the number-right scores, and the percentages of scores yielded by each item response item model at least one standard error of measurement greater (row 3) and less (row 4) than the corresponding number-right scores. The internal consistency (Cronbach coefficient alpha) of each test, adjusted so that each had the same length, were 0.83 for the SS-NTW and **0.82** and for the SS-ABS. The corresponding values for the two chemistry subtests were respectively 0.80 and **0.78**

Social Studies Subtests

As expected, the mean level of performance decreased with decreasing ability for both the SS-NTW and SS-ABS subtests. The means of the five sets of transformed *T*-scores at each ability level were approximately equal (i.e., differed by one score point or less) both within and between each subtest. In contrast, the standard deviations of the five sets of scores were more variable. The standard deviations of the number-right scores for both subtests were lower than the standard deviations of the IRT scores by slightly more than one score point in the high-ability group and essentially equal in the middle- and low-ability groups. Further, although the standard deviations for the SS-ABS were essentially equal to the corresponding standard deviations for the SS-NTW in the high- and low-ability groups, with the exception of the nominal response scores, they were greater than the corresponding standard deviations in the SS-NTW by slightly more than one score point in the middle-ability group.

Examination of the correlations reported in the first row of Table 2 reveals that the correlations between the number-right scores and each of the IRT scores from the SS-NTW subtests and the SS-ABS subtest were within 0.01 of each other (e.g., the correlations between the number-right and nominal model in the middle group for the SS-NTW subtest and SS-ABS subtest were respectively 0.91 and **0.92**). However, the correlations between the scores obtained from the number-right and the nominal response models were lower than the remaining correlations for each ability group (e.g., for the number-right and nominal response models versus the number-right and two- and three-parameter models: 0.94 vs. 0.97 and 0.96 in the high-ability group, 0.91 versus 0.95 and 0.94 in the middle-ability group, and 0.94 versus 0.98 and 0.97 in the low-ability group). As expected, the correlations between the number-right and one-parameter scoring models were essentially one. These findings indicate that with the exception of the using nominal response scores in place of the number-right scores, the students' ranked positions are essentially maintained across the five scoring models for both subtests.

The closest agreements between transformed scores were found for the pairs of scores yielded by the number-right and one-parameter item response scoring models for both the SS-NTW and SS-ABS subtests in the middle- and low-ability groups. For these pairs of scores, the root mean squares varied from 0.69 to 0.89 (second row, Table 2). The root mean squares for the remaining pairs of scores varied from 1.45 to 1.93 for the SS-NTW and **1.67** to **2.04** for the

SS-ABS in the middle-ability group, and from 1.51 to 2.06 for the SS-NTW and **1.83** to **2.43** for the SS-ABS in the low-ability group. These results suggest that there are differences between score values for some of the pairs of scores being compared, with slightly larger differences for the SS-ABS, particularly for the nominal response model in the low-ability group (**2.43** vs. 2.06). This suggestion is reflected in the percentages of examinees who received IRT scores that were at least one standard error of measurement greater than or less than their number-right scores (see rows 3 and 4, Table 2, middle- and low-ability students). Although not large, ranging from 0.00% to **4.70**%, these percentages tend to be greater for the SS-ABS than for the SS-NTW, and for the number-right scores vs. the nominal scores. Although the percentage of low-ability students for whom the nominal scores were higher than their number-right scores is greater than the percentage for whom the nominal scores were less than their number-right scores for the SS-NTW (4.51 vs. 0.28), these percentages are more equal for the SS-ABS (**4.70** vs. **3.22**). For the high-ability group, the root mean squares varied from 1.60 to 2.18 for the SS-NTW and from **1.33** to **2.04** for the SS-ABS (row 2, Table 2). The percentages of examinees who received IRT scores that differed from their number-right by at least one standard error of measurement were greater for the SS-NTW than for the SS-ABS (1.92 to 6.68 vs. **0.00** to **3.44**; rows 3 and 4, Table 2). Further, for both subtests, the percentages of students whose nominal scores exceeded their number-right scores by at least one standard error of measurement was greater than the percentages of students whose nominal scores were less than their number-right scores by at least one standard error of measurement (6.68 vs. 0.71 for SS-NTW and **3.44** vs. **1.21** for SS-ABS). Thus although the rankings—with the one exception noted above—are essentially the same across the five scoring models, there will be differences between some score values, with the differences being somewhat more noticeable when two- and three-parameter and particularly nominal scores are used in place of number-right scores.

Chemistry Subtests

Again, as expected the mean level of performance decreased with decreasing ability. However, in contrast to social studies, although the means of the five sets of transformed scores were approximately equal across the five scoring models in each subtest, the means for the CH-NTW were between 1.64 and 1.95 *T*-score points higher than the means for the CH-ABS in the high-ability group, whereas for both the middle- and low-ability groups the differences between corresponding subtest means were less (0.63 to 1.11 and 0.29 to 1.24 respectively). Except for the difference between the standard deviations of the number-right and three-parameter IRT scores for the low-ability group, the differences between the number-right standard deviation and each of the IRT standard deviations were less than one score point for both subtests in each ability group. However, although the standard deviations of the five sets of scores for the CH-NTW and for the CH-ABS are essentially equal (differences range from 0.08 to 0.38) in the high-ability group, the standard deviations for the five sets of scores for the CH-NTW are lower than the corresponding standard deviations for the CH-ABS in both the middle-ability (by 1.15 to 1.51) and low-ability (by 1.28 to 2.31) groups.

Table 3
Comparison Between Ability Estimates:
CH-NTW and CH-ABS

	θ_{NR}	θ_1	θ_2	θ_3	θ_{NM}
High-Ability Group (n=1,066)					
Mean (SD)	61.40 (4.24) **59.54 (4.55)**	61.68 (5.20) **59.98 (5.38)**	61.85 (5.13) **59.90 (4.90)**	61.66 (5.20) **60.02 (5.28)**	61.78 (5.13) **59.85 (4.75)**
$r_{\theta_{NR,j}}$	-	0.99 **0.99**	0.97 **0.89**	0.96 **0.86**	0.94 **0.87**
$RMS_{\theta_{NR}-\theta_j}$	-	1.09 **0.98**	1.48 **2.26**	1.65 **2.75**	1.91 **2.44**
$P_{greater}$	-	0.00 **0.00**	0.00 **3.19**	0.00 **4.32**	0.28 **4.78**
P_{less}	-	0.00 **0.00**	0.09 **1.31**	0.56 **2.72**	1.31 **2.63**
Middle-Ability Group (n=1,255)					
Mean (SD)	50.99 (4.49) **51.64 (5.64)**	50.491 (4.42) **51.23 (4.89)**	50.41 (4.55) **51.52 (5.91)**	50.52 (4.53) **51.34 (6.04)**	50.49 (4.66) **51.55 (5.85)**
$r_{\theta_{NR,j}}$	-	0.99 **0.99**	0.95 **0.88**	0.93 **0.85**	0.90 **0.85**
$RMS_{\theta_{NR}-\theta_j}$	-	0.55 **0.59**	1.54 **2.86**	1.81 **3.20**	2.11 **3.13**
$P_{greater}$	-	0.00 **0.00**	0.00 **5.66**	0.08 **7.17**	1.31 **7.49**
P_{less}	-	0.00 **0.00**	0.48 **4.78**	1.51 **8.29**	2.23 **6.77**
Low-Ability Group (n=1,117)					
Mean (SD)	38.32 (6.06) **39.15 (8.37)**	38.69 (5.52) **39.39 (7.47)**	38.37 (5.40) **39.04 (7.25)**	38.73 (5.69) **39.19 (6.97)**	38.73 (5.34) **39.97 (7.36)**
$r_{\theta_{NR,j}}$	-	1.00 **0.99**	0.97 **0.93**	0.92 **0.91**	0.93 **0.91**
$RMS_{\theta_{NR}-\theta_j}$	-	0.66 **0.99**	1.63 **3.17**	2.39 **3.54**	2.36 **3.45**
$P_{greater}$	-	0.00 **0.00**	0.18 **7.52**	3.67 **10.30**	4.12 **7.43**
P_{less}	-	0.00 **0.00**	0.09 **6.54**	2.95 **7.97**	2.15 **9.13**

Note. Values in regular type are for the CH-NTW subtest and in boldface type for the CH-ABS subtest.

Examination of the correlations presented in the first row of Table 3 reveals that the presence of absurd options influenced the two-parameter, three-parameter, and nominal response scoring models, particularly in the high- and middle-ability groups. For these two ability groups, the values of the correlations between pairs of scores obtained from the number-right and scores and the two-parameter, three-parameter, and nominal scores for the CH-ABS score were from 0.05 to 0.10 smaller than the corresponding values for the CH-NTW. For the low-ability group, the differences between the values of the corresponding pairs were smaller, varying between 0.01 and 0.04. However, the correlations between the three-parameter and to a greater degree the nominal response scores and the number-right scores for both the CH-NTW and CH-ABS subtests tended to be lower than the correlations between the number-right scores and the one- and two-parameter scores for each ability group (e.g., for the number-right and nominal response scores versus the number-right and two-parameter scores: 0.94 vs. 0.97 and **0.87** versus **0.89** in the high-ability group, 0.90 versus 0.95 and 0.85 vs. 0.88 in the middle-ability group, and 0.93 versus 0.97 and **0.91** versus **0.93** in the low-ability group). In contrast, the correlations between the number-right and one-parameter scores were essentially perfect for both subtests. Taken together, these findings indicate that students' ranked positions will probably vary somewhat when the three-parameter or nominal scores are used in place of the number-right scores.

The patterns of root mean square deviations and proportions of students who received two-parameter, three-parameter, and nominal scores greater than their number-right scores by at least one standard error of measurement followed the same pattern as the correlations for the three ability groups. Congruent with the correlational findings, the root mean square deviations between ability estimates yielded by the number-right model and the two-parameter, three-parameter, and nominal response models were consistently greater for the CH-ABS than for the CH-NTW (e.g., **2.44** vs. 1.91, **3.13** vs. 2.11, and **3.45** vs. 2.36 for the number-right and nominal scores; row 2, Table 3). Compatible with this finding, the percentages of students who received two-parameter, three-parameter, and nominal response scores that differed from their number-right scores by at least one standard error of measurement were consistently greater for the CH-ABS than for the CH-NTW (e.g., **4.78**% vs. 0.28% for the nominal scores greater than number-right scores, high-ability students; 3rd row, Table 3). The percentages of two-parameter, three-parameter, and nominal response scores greater than or less than their corresponding number-right scores tended to vary across scoring model and group. For example, for the CH-ABS, the percentage of students with higher three-parameter scores than number-right scores exceeded the percentage of students with lower three-parameter scores in the high-ability group (**4.32** vs. **2.72**), was less in the middle-ability group (**7.17** vs. **8.29**), and greater in the low-ability group (**10.30** vs. **7.97**). In contrast, although the percentage of nominal scores greater than the corresponding number-right scores exceeded the percentage of lower nominal scores in the high-ability group (**4.78** vs. **2.63**), these percentage were closer in the middle-ability group (**7.49** vs. **6.77**), and in reverse order in the low-ability group (**7.43** vs. **9.13**). Of note, though, is that at least one in 10 middle- and low-ability students received two-parameter, three-parameter, or

nominal response scores that differed from their number-right scores for the CH-ABS. In contrast and in agreement with the correlations, no students had a one-parameter score that exceeded their number-right scores for both subtests. These findings suggest that for several students, if the test contains items susceptible to the ID1 test-wiseness strategy, the two- and three-parameter and nominal models will yield ability estimates that differ from the number-right or one-parameter model for the CH-NTW and more so for the CH-ABS.

Discussion

Differences were found between pairs of ability estimates obtained when two-parameter, three-parameter, and nominal response model scores were used in place of number-right scores. The differences tended to be greater for chemistry than for social studies, and with the exception of high-ability students in social studies, for the subtest containing items with absurd options than for the subtest containing nonsusceptible test-wise items. Consequently, it appears—at least for the two subject areas considered in the present study—that the scoring models cannot be used interchangeably to obtain estimates of examinees' abilities, particularly when a test contains test-wise susceptible items. Further, in agreement with Bock (1972), Levine and Drasgow (1983), Thissen (1976), and Thissen and Steinberg (1984), the differences tended to be greatest in the low-ability group. These findings indicate that for some students the number-right scores will be either greater than or less than the scores yielded by the two- and three-parameter models and the nominal response model. Consequently, for these students a different, perhaps incorrect, decision such as pass-fail or placement in an alternate program of studies could be made if a two-, three-, or nominal response score were used in place of the corresponding number-right score.

These findings suggest that the five scoring models can be grouped into two sets. The first set contains the number-right and one-parameter models and the second set includes the two-parameter, three-parameter, and nominal response models. The difference between the models in the two sets seems to be associated with the presence of the discrimination parameter found in the models in the second set. Model-data fit analyses (likelihood-ratio χ^2 test for the SS-NTW, SS-ABS, and CH-NTW, and because the number of items was less than 20, root mean square residuals for the CH-ABS, Mislevy & Bock, 1990) revealed that lack of item fit was not an explanation for the lower correlations, higher root mean squares, and greater proportions of students with different score values between pairs of number-right and two- and three-parameter scores and the nominal response scores. The test of nonlinearity (Glass & Hopkins, 1996) revealed that for all pairs of scores the relationships were linear. Furthermore, the shapes, although wider, were cigar-like, suggesting that the lower correlations, higher root mean squares, and greater proportions were not attributable to specific score levels (compare Fan, 1998, Table 4). Thus it appears that the discrimination index present in each of the two- and three-parameter response models and the nominal response model is sensitive to the information contained in all four options.

The finding that the difference between the CH-NTW and CH-ABS subtests was greater than the difference between the SS-NTW and SS-ABS subtests may be attributable to a substantive difference between the two subject areas and

the nature of the students who take these take these two courses and write the year-end examinations. As indicated above, the subject matter of social studies involves the use of personal and societal values in contrast to the subject matter of chemistry, which is more objective and scientific. As indicated above, examination of the item-option trace lines revealed that although many students were able to identify absurd options for both social studies and chemistry, many experienced more difficulty identifying the correct answer from among the remaining options for social studies than for chemistry. The non-absurd options, including the correct option, appear to be more equally attractive to the students, particularly the middle- and low-ability students for social studies. In contrast, when the items measured more objective scientific knowledge and skill (i.e., chemistry), the students were better able to identify the correct answer from among the non-absurd options. It should be noted that the similar behavior was observed on the items not susceptible to test-wiseness. The trace lines for the distractors in the case of social studies were more equal to one another and equally attractive to the correct option, particularly again for the middle- and low-ability students; in the case of chemistry, the trace lines for the distractors were less attractive than the correct option, with the degree of unattractiveness increasing with decreasing ability. These findings are reflected in the observation that with the exception of the high-ability students in each area, the mean scores for social studies are lower than the mean scores for chemistry (compare Table 1). Individual think-aloud interviews of students while they are responding to the test items and protocol analyses (Ericsson & Simon, 1993) are needed to clarify this issue. Finally, a greater number of students take and write social studies than take and write chemistry. The students who take chemistry tend as a group to be academically stronger than the students who take social studies. Thus although top-ability students take both courses (which are needed to enter science programs in universities, colleges, and institutes of technology), a greater proportion of lower-ability students take social studies and instead of chemistry, biology as their required science. Consequently, given how the groups were formed in the present study, it is likely that the middle- and lower-ability students in chemistry are better able to take advantage of test-wise cues than the corresponding groups in social studies.

However, there is no apparent reason for the finding that for the high-ability students in social studies, performance on the SS-NTW exceeded the performance on the SS-ABS. Again, as suggested above, individual think-aloud interviews of students as they respond to the items followed by protocol analyses (Ericsson & Simon, 1993) of the responses are needed to clarify this issue.

Note

1. The replication study results are available on request from the first author: JTomkowicz@ctb.com.

References

Alberta Learning. (1999a). *Chemistry 30 grade 12 diploma examination, June 1999*. Edmonton, AB: Author.

Alberta Learning. (1999b). *Social studies 30 grade 12 diploma examination, June 1999*. Edmonton, AB: Author.

Aiken, L.R. (1987). Testing with multiple-choice items. *Journal of Research and Development in Education, 20*, 44-58.
Bennet, R.E., & Ward, W.C. (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*. Hillsdale, NJ: Erlbaum.
Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.
Bock, R.D. (1997). The nominal categories model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 33-49). New York: Springer-Verlag.
Brown, J.S., & Burton, R.R. (1978). Diagnostic model for procedural bugs in basic mathematical skills. *Cognitive Science, 4*, 379-426.
Cattell, R.B. (1952). *Factor analysis*. New York: Harper & Bros.
Crehan, K.D., Koehler, R.A., & Slakter, M.J. (1974). Longitudinal studies of test-wiseness. *Journal of Educational Measurement, 11*, 209-212
De Ayala, R.J. (1989). A comparison of the nominal response model and the three-parameter logistic model in computerized adaptive testing. *Educational and Psychological Measurement, 49*, 789-805.
De Ayala, R.J. (1993). Methods, plainly speaking: An introduction to polytomous item response theory models. *Measurement and evaluation in Counseling and Development, 25*, 172-189.
Diamond, J., & Evans, W.J. (1972). An investigation of the cognitive correlates of test-wiseness. *Journal of Educational Measurement, 11*, 209-212.
Ericsson, K.A., & Simon, H.A. (1993). *Protocol analysis*. Cambridge, MA: MIT Press.
Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58,* 375-381.
Glass, G.V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology*. Needham Heights, MA: Allyn and Bacon.
Hambleton, R.K., & Murphy, E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education, 5*, 1-16.
Hambleton, R.K., & Murray, L. (1983). Some goodness of fit investigations for item response models. In R.K. Hambleton (Ed.), *Applications of item response models.* Vancouver, BC: Educational Research Institute of British Columbia.
Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
Hughes, C.A., Salvia, J., & Bott, D. (1991). The nature and extent of test-wiseness cues in seventh- and tenth- grade classroom tests. *Diagnostique, 16*, 153-163.
Jacobs, P., & Vandeventer, M. (1970). Information in wrong responses. *Psychological Reports, 36*, 311-315
Kelley, T.L. (1927). *Interpretation of educational measurements*. New York: World Book Company.
Lane, S., Stone, C.A., & Hsu, H. (1990). *Diagnosing students' errors in solving algebra word problems*. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.
Levine, M.V., & Drasgow, F. (1983). The relation between incorrect response option choice and estimated ability. *Educational and Psychological Measurement, 43*, 675-685.
Lord, F.M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington, DC: American Council on Education and Macmillian.
Millman, J. (1966). *Test-wiseness in taking objective achievement and aptitude examination: Its nature and importance. Final report*. New York: College Entrance Examination Board.
Mislevy, R.J., & Bock, R.D. (1990). *BILOG3: Item analysis and test scoring with binary logistic model* [computer program]. Moresville, IN: Scientific Software.
Ndalichako, J.L., & Rogers, W.T. (1997). Comparison of finite score theory, classical test theory, and item response theory in scoring multiple-choice items. *Educational and Psychological Measurement, 57*, 580-589.
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.
Nedelsky, L. (1954). Ability to avoid gross ever as a measure of achievement. *Educational and Psychological Measurement, 14*, 459-472.
Nelson, L. (1983). *Lertap3: A test, survey and general data analysis system for small computers.* Dunedin, NZ: University of Otago, Department of Education.

Rogers, W.T., & Bateson, D.J. (1991a). Verification of the model of test-taking behavior of high school seniors. *Journal of Experimental Education, 59*, 331-350.
Rogers, W.T., & Bateson, D.J. (1991b). The influence of test-wiseness upon performance of high school seniors on school leaving examination. *Applied Measurement in Education, 4*, 159-183.
Rogers, W.T., & Ndalichako, J. (2000). Number-right, item response, and finite scoring: Robustness with respect to lack of equally classifiable options and item option independence. *Educational and Psychological Measurement, 60*, 5-19.
Rogers, W.T., & Wilson, C. (1993). *The influence of test-wiseness upon performance on high school students on Alberta Education's diploma examinations*. (Alberta Education, Student Evaluation Branch, Contract No. 91-0143). Edmonton, AB: University of Alberta.
Rogers, W.T., & Yang, P. (1996). Test-wiseness: Its nature and application. *European Journal of Psychological Assessment, 12*, 247-259.
Sarnacki, R.E. (1979). An examination of test-wiseness in the cognitive test domain. *Review of Educational Research, 49*, 252-279.
Stout, W.F. (1987). A non-parametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.
Tatsuoka, K.K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 13*, 201-214.
Thissen, D. (1976). Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Measurement, 13*, 201-214.
Thissen, D. (1991). *MULTILOG user's guide version 6*. Chicago, IL: Scientific Software.
Thissen, D., & Steinberg, L. (1984). A response model for multiple-choice items. *Psychometrika, 49*, 501-519.
Thissen, D., & Steinberg, L. (1997). A response model for multiple-choice items. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51-65). New York: Springer-Verlag.
Thissen, D., Steinberg, L., & Fitzpatrick, A.R. (1989). Multiple-choice models: The distracters are also part of the item. *Journal of Educational Measurement, 26*, 161-176.
Towns, M.H., & Robinson, W.R. (1993). Student use of test-wiseness strategies in solving multiple-choice chemistry examinations. *Journal of Research in Science Teaching, 7*, 709-722.
Wright, B.D., & Stone, M.H. (1979). *Best test design*. Chicago, IL: MESA