

---

Gender Differential Item Functioning in  
Mathematics and Science: Prevalence and Policy  
Implications

**Mark Gierl**  
**Shameem Nyla Khaliq**  
**Keith Boughton**

Centre for Research in Applied Measurement and Evaluation  
University of Alberta

Paper Presented at the Symposium entitled "Improving Large-Scale  
Assessment in Education" at the Annual Meeting of the Canadian  
Society for the Study of Education

Sherbrooke, Québec, CANADA

June 7-11, 1999

---

This research was supported with funds from the Social Sciences and Humanities Research Council of Canada (SSHRC). Please address correspondence to Mark J. Gierl, Assistant Professor of Educational Psychology, Faculty of Education, University of Alberta, Edmonton, Alberta, Canada, T6G 2G5. Email: [mark.gierl@ualberta.ca](mailto:mark.gierl@ualberta.ca).

### **Gender Differential Item Functioning in Mathematics and Science: Prevalence and Policy Implications**

Large-scale achievement testing is pervasive in Canada today. At the provincial level, 9 of the 10 provinces have large-scale achievement testing programs. At the national level, the Council of Ministers of Education conduct national assessments in language arts, mathematics, and science (e.g., Council of Ministers of Education, Canada, 1996). At the international level, Canada has participated in many international assessments, the most recent being the Third International Mathematics and Science Study (Robitaille, Taylor, & Orpwood, 1996). In these types of assessments, differential item functioning (DIF) is a constant concern. DIF is present when examinees from different groups have a different probability or likelihood of answering an item correctly, after controlling for overall ability (Shepard, Camilli, & Averill, 1981). Once identified, DIF may be attributed to item impact or to item bias. Item impact can be described as any group disparity in item performance that reflects actual knowledge and experience differences on the construct of interest (Clauser & Mazor, 1998). Alternatively, item bias is defined as invalidity or systematic error in how a test item measures a construct for the members of a particular group (Camilli & Shepard, 1994, p.8). Item bias is of particular concern on tests of mathematics achievement where differences between females and males are commonly found (Kimball, 1989; Scheuneman & Grima, 1997). The failure to understand and account for gender differences on any test may lead to interpretations and, consequently, actions that are invalid (Principles for Fair Student Assessment Practices for Education in Canada, 1993).

To ensure that tests are fair for all examinees, most large testing programs have a formal review, which is part of the test development process, where items are screened by content specialists for text that might be inappropriate or unfair to relevant subgroups in the test-taking population including female examinees, minority group examinees such as aboriginal students, or disabled examinees. These reviews are conducted before the tests are administered. Statistical measures of DIF can also help test developers identify items that are biased against examinees since the ultimate criterion of item equivalence must come from an analysis of the examinee responses. Typically, DIF analyses are conducted after the tests are administered using large samples of examinee data.

The purpose of this study is threefold. First, the test review used by one large testing program in Canada—the Student Evaluation Branch at Alberta Education—will be described. This review is conducted prior to test administration. Second, three DIF statistical procedures will be presented and the gender DIF rates for achievement tests administered in Mathematics and Science will be reported. Third, policy implications will be discussed. The results from this study will shed light on the effectiveness of a well-established test review used to screen for gender bias and identify areas where possible refinements and research are needed.

#### Test Development at Alberta Education

Like many large testing programs, the Student Evaluation Branch at Alberta Education has a formal review to scrutinize items during the test development process. The items are reviewed before field testing by content specialists (i.e., test developers and teachers) as well as after field testing when a small sample of student data are available. Practicing teachers are involved throughout the test development process. Participating teachers must be recommended by their principal and superintendent and must have a minimum of two years teaching experience. The involvement of teachers is significant because they are familiar with the students and the curriculum thereby ensuring the test items are appropriate for the students given the provincial expectations and content in a given subject area. These teachers are also provided with basic item writing principles using content-specific manuals which contain explicit guidelines and examples.

The test development process contains three general steps: Item writing, field testing, and creating the final form of the test. Item writing begins when teachers are nominated to serve on the item writing committees. Alberta Education uses a rotating selection procedure to ensure that teachers throughout the province are represented. The item writers meet several times a year to develop new items. If possible, the items are developed using a realistic context that would be familiar or topical for students in the province.

Next, the items are compiled and field tested. The field tests undergo an internal review before they are administered to students. Members of the internal review committee are from the achievement testing unit at Alberta Education. The purpose of the internal review is to examine the field tests for content validity, curricular validity, item appropriateness (e.g., wording, length, interest), bias (e.g., gender, cultural, disability), balance to the test blueprint, and tone. Members of the committee also

critique the tests. Once reviewed, the field tests are administered throughout the province to participating schools. Although participation is voluntary, approximately 400 students write the field tests each year. Teachers and students are encouraged to comment on the field test items. The field tests are then evaluated using classical item analysis. These three sources of information—the internal review committee comments, the teachers' and students' comments from field testing, and the statistical results—are considered by the test developer during the selection of the final test items.

The last step involves creating the final form of a test. A second internal review is conducted after the test developer has made his or her final selection of the test items. The committee for this review includes the achievement test developers and five teachers. The purpose of this review is to finalize the selection of the test items and to ensure the test meets curriculum, assessment, and achievement standards. Again, the review panel scrutinizes each test item looking for content and curricular match, item appropriateness, and possible biases. The final version of the test is then read by two editors who check the final version for grammatical accuracy. These editors also examine the test for any possible biases and attempt to have gender balance by counting the number of references to males and females, either by name, situation, or picture to ensure that the references to males and females are equal in number. The items are also reviewed for the inclusion of males and females in non-traditional roles. The final version of the test is then validated and the key is checked. The validation of the final form is conducted by two teachers who are currently teaching the academic subject but have not participated in the test development process just described. They examine the test for accuracy of information presented in each item, wording, and possible biases. Once this step is completed, the test developer can sign off the test and the extensive development process is finished.

#### Overview of DIF Statistical Procedures

Content review is one method for identifying and eliminating gender bias. The review process just described provides one example. Statistical procedures can also be used to identify biased items. Three procedures are currently popular, although many statistical methods exist (for a review see Millsap & Everson, 1989; Camilli & Shepard, 1994). The three procedures used in this study are the Mantel-Haenszel, Simultaneous Item Bias Test, and logistic regression. The three procedures are briefly described in the next section.

Mantel-Haenszel (MH) is a nonparametric approach for identifying DIF (Mantel & Haenszel, 1959, applied to DIF research by Holland & Thayer, 1988). MH yields a chi-square test with one degree of freedom to test the null hypothesis that there is no relation between group membership and test performance on one item after controlling for ability. MH is computed by matching examinees in each group on total test score and then forming a 2-by-2-by- $K$  contingency table for each item, where  $K$  is the score level on the matching variable of total test score. At each score level  $j$ , a 2-by-2 contingency table is created for each item  $i$ . The MH  $\chi^2$  statistic tests the null hypothesis that there is no relation between group membership and test performance on one item after controlling for overall test performance.

The MH procedure is also used to estimate the constant odds ratio that yields a measure of effect size for evaluating the amount of DIF that is present. Research at the Educational Testing Service has resulted in proposed values for interpreting the constant odds ratio, called  $\Delta - MH$  when transformed onto the delta scale, which serves as the effect size measure for classifying DIF at the item level. DIF is considered negligible when  $\Delta - MH$  is not significantly different from 0 and the magnitude of the  $|\Delta - MH| < 1$ . DIF is considered moderate when  $\Delta - MH$  is significantly different from 0 and has either (a)  $1 \leq |\Delta - MH| < 1.5$  or (b)  $|\Delta - MH|$  is at least 1 but not significantly greater than 1. DIF is considered large when  $\Delta - MH$  is significantly greater than 1 and  $|\Delta - MH| \geq 1.5$ . (Zieky, 1993, p. 342). These ratings are referred to as A-, B-, and C-level DIF to denote negligible, moderate, and large amounts of DIF.

The Simultaneous Item Bias Test (SIBTEST) is an alternative statistical method for detecting DIF proposed by Shealy and Stout (1993). SIBTEST is intended to model multidimensional data, although it can be used for unidimensional data as well. With this procedure, the complete latent space is viewed as multidimensional,  $[\Theta = (\Theta, \eta)]$ , where  $\Theta$  is the unidimensional construct of interest or the target ability and  $\eta$  is the extraneous or nuisance ability. The statistical hypothesis tested by SIBTEST is:

$$H_0: B(T) = P_R(T) - P_F(T) = 0$$

vs.

$$H_1: B(T) = P_R(T) - P_F(T) \neq 0,$$

where  $B(T)$  is the difference in probability of a correct response on the studied item for examinees in the Reference and Focal groups matched on true score;  $P_R(T)$  is the probability of a correct response on the studied item for examinees in the Reference group with true score  $T$ ; and  $P_F(T)$  is the probability of a correct response on the studied item for examinees in the Focal group with true score  $T$ . In other words  $B(T)$ , the parameter representing the amount of unidimensional DIF when a single test item is evaluated, is 0 when there is no DIF and nonzero when DIF is present. With the SIBTEST approach, items on the test are divided into two subsets, the suspect subtest and the matching subtest. The suspect subtest contains the biased item and the matching subtest contains the rest of the items. For each matching subtest score,  $k$ , the corresponding subtest true score for the Reference and Focal groups is estimated using linear regression. The estimated true scores are then adjusted using a regression correction technique to ensure the estimated true score is comparable for the examinees in the Reference and Focal groups on the matching subtest. In the final step,  $B(T)$  is estimated using  $\hat{B}$  which is the weighted sum of the differences between the proportion-correct true scores on the studied item for examinees in the two groups across all score levels.

Like the MH procedure, SIBTEST yields an overall statistical test as well as a measure of the effect size for each item ( $\hat{B}$  is an estimate of the amount of DIF). Roussos and Stout (1996, p. 220) suggested a range of values for interpreting  $\hat{B}$ . When the null hypothesis is rejected and when  $|\hat{B}| < 0.059$ , DIF is considered negligible; when the null hypothesis is rejected and when  $0.059 \leq |\hat{B}| < 0.088$ , DIF is considered moderate; and when the null hypothesis is rejected and when  $|\hat{B}| \geq 0.088$ , DIF is considered large. These guidelines are used to classify items in category A, B, or C (i.e., negligible, moderate, and large amounts of DIF).

Logistic regression is a third approach commonly used to identify DIF (LR; Swaminathan & Rogers, 1990). Whereas MH and SIBTEST are designed to measure uniform DIF, these

approaches are poor at detecting non-uniform DIF (Rogers & Swaminathan, 1993). Uniform DIF exists when there is no interaction between ability level and group membership. That is, the probability of answering an item correctly is greater for one group uniformly over all ability levels. In item response theory terminology, uniform DIF is indicated by parallel item characteristic curves. Non-uniform DIF occurs when there is an interaction between ability level and group membership. In this case, the difference in the probabilities of a correct response for the two groups is not the same at all levels of ability. Again, using item response theory terminology, non-uniform DIF is indicated by nonparallel item characteristic curves. LR can detect uniform and non-uniform DIF, which may provide a noteworthy advantage over both MH and SIBTEST.

The presence of DIF in the LR approach is determined by testing the improvement in model fit that occurs when a term for group membership and a term for the interaction between test score and group membership are successively added to the regression model. A chi-square test is then used to evaluate the presence of uniform and non-uniform DIF on the item of interest by testing each term included in the model. The general model for logistic regression takes the form:

$$P(u = 1) = \frac{e^z}{1 + e^z},$$

where  $u$  is the score on the studied item. Performance on the studied item is first conditioned on total test score. In this step,  $z = \mathbf{b}_0 + \mathbf{b}_1X$ , where  $X$  is the test score (Model 1). This serves as the baseline model. The presence of uniform DIF is then tested by examining the improvement in chi-square model fit associated with adding a term for group membership ( $G$ ) against the baseline model. That is, Model 2 (i.e.,  $z = \mathbf{b}_0 + \mathbf{b}_1X + \mathbf{b}_2G$ ) subtracted from Model 1. The presence of non-uniform DIF is tested by examining the improvement in chi-square model fit associated with adding a term for group membership ( $G$ ) and a term for the interaction between test score and group membership ( $XG$ ) against model 2. In other words, Model 3 (i.e.,  $z = \mathbf{b}_0 + \mathbf{b}_1X + \mathbf{b}_2G + \mathbf{b}_3XG$ ) subtracted from Model 2.

Zumbo and Thomas (1996) developed an index to quantify the magnitude of DIF for the LR procedure based on partitioning a weighted least-squares estimate of  $R^2$  that yields an effect size measure (also see Zumbo, 1999). This index is obtained, first, by computing the  $R^2$  measure of fit

for each term in the LR model (i.e., test score, group membership, test score-by-group membership interaction) and then by partitioning the  $R^2$  for each of the terms. A DIF effect size for the group membership term is produced by subtracting the  $R^2$  for the group membership term (Model 2) from the  $R^2$  for the total test score term (Model 1). The result is an effect size measure associated with group membership that quantifies the magnitude of uniform DIF (herein called  $R^2\Delta - U$ ). A second DIF effect size is produced for the total score-by-group membership term by subtracting the  $R^2$  for the group

group membership interaction that quantifies the magnitude of non-uniform DIF (herein called  $R^2\Delta - N$ ). As with the MH and SIBTEST effect size measures,  $R^2\Delta$  can be used with the LR significance test to identify items with DIF. Jodoin (1999) empirically-established guidelines for interpreting  $R^2\Delta$ . An item has negligible or A-level DIF when the chi-square test for model fit is not statistically significant or when  $R^2\Delta < 0.035$ . An item has moderate or B-level DIF when the chi-square test is statistically significant and when  $0.035 \leq R^2\Delta < 0.070$ . An item has large or C-level DIF when the chi-square test is statistically significant and when  $R^2\Delta \geq 0.070$ . These guidelines are applicable to both uniform and non-uniform DIF, and were used to classify DIF items in this study.

## Method

### Student Sample and Achievement Tests

Students' response data from the 1994, 1995, 1996, 1997, and 1998 administrations of the Alberta Education Mathematics and Science achievement tests were analyzed in this study at Grades 3 (Mathematics only), 6, and 9. 12000 students (6000 males and 6000 females) were randomly selected from the Alberta Education database in each content area for each grade during each year, and their data were analyzed using three DIF statistical procedures to identify items that function differently for males and females.

The achievement tests are used in the province of Alberta to monitor student achievement. All students are expected to write these tests, and participation rates typically exceed 95%. The test score

does not necessarily contribute to a student's final course grade although teachers are encouraged to mark the tests and use the results for student grading. The test items are based on concepts, topics, and facts from the province-wide Program of Studies. Sample test items can be downloaded from the Alberta Education internet site (<http://ednet.edc.gov.ab.ca/studenteval/>).

### Statistical Analysis

DIF statistical analyses were conducted for each item using MH, SIBTEST, and LR. The item under consideration was included in forming the score groups for MH and LR and no iterative purification was used in any of the DIF analyses. All test statistics were interpreted at an alpha-level of 0.05. In all comparisons described below, items with a B- or C-level rating are considered DIF items whereas those with an A-level rating are not. This decision seems justified since B- and C-level DIF items are scrutinized for potential bias in tests reviews (Zieky, 1993).

## Results

### Psychometric Characteristics of the Test Forms and Items

A summary of the exam characteristics on the Mathematics and Science tests is presented in Tables 1 and 2, respectively. Typically, the differences reported in Tables 1 and 2 are tested between groups. However, the large samples used in this study resulted in many differences that were statistically but not practically significant. Hence, statistical outcomes are not reported. Instead, some general trends are highlighted. First, males consistently scored higher than females in both Mathematics and Science. The magnitude of this mean difference favoring males increased in both content areas as students moved from grades 3 to 9. Second, the score variability and distribution characteristics were similar between males and females in both Mathematics and Science across all test administrations.

### Comparison of DIF Procedures

Relations Among DIF Effect Size Measures. Since there is little agreement on which DIF statistical procedure is most accurate, three different methods, all currently popular, were used in this study. Table 3 contains the mean correlation coefficients and standard deviations for each DIF effect size measure using the Mathematics and Science achievement tests across grade and test administration. These results provide some information about the DIF classification consistency across the three procedures. The effect size measures were highly correlated across DIF

procedures except the measure for non-uniform DIF. For the Mathematics tests, the MH effect size measure,  $\Delta - MH$ , was highly correlated with the SIBTEST effect size measure,  $\hat{B}$ , at  $-.95$  and with the LR effect size measures for uniform DIF,  $R^2\Delta - U$ , at  $.84$ .  $\Delta - MH$  and the non-uniform DIF measure,  $R^2\Delta - N$ , were correlated at  $-.04$ .  $\hat{B}$  and  $R^2\Delta - U$  were highly correlated at  $.82$  whereas  $\hat{B}$  and  $R^2\Delta - N$  were correlated at  $.02$ . The correlation between the  $R^2\Delta$  measures was  $.05$ . The standard deviations for the uniform measures were consistently small while the standard deviation for the non-uniform measures were much larger. This outcome is likely due to the increased variability in the  $R^2\Delta - N$  measure as the amount of non-uniform DIF varied by grade and test administration (i.e., some data sets had more non-uniform DIF than others).

Similar results were found using the Science data.  $\Delta - MH$  was highly correlated with  $\hat{B}$  and  $R^2\Delta - U$  ( $-.95$  and  $.83$ , respectively) but not with  $R^2\Delta - N$  ( $r = .04$ ).  $\hat{B}$  and  $R^2\Delta - U$  were highly correlated at  $.87$  whereas  $\hat{B}$  and  $R^2\Delta - N$  were weakly correlated at  $.06$ . The  $R^2\Delta$  measures also had a weak correlation at  $.15$ . These results strongly suggest that the three DIF procedures in this study, while based on different assumptions about the data, produced very similar item rankings with their respective effect size measures. It also demonstrates that  $R^2\Delta - N$  was not strongly related to the uniform effect size measures, as would be expected, because the procedures flag different types of DIF items which provides some criterion-related evidence to support the use of this new non-uniform DIF effect size measure.

DIF Classification Consistency. DIF classification consistency was also assessed by comparing the items classified as B- or C-level DIF across procedures. The results are presented in off-diagonals in Table 4 for Mathematics and Table 5 for Science. In Mathematics, classification consistency is relatively strong for Grade 3 DIF items but not for Grades 6 or 9 DIF items. Each statistical procedure tended to flag a subset of the same items flagged by the other two procedures but rarely is the match perfect. Rather, each procedure tends to flag about half the same items as the other two procedures. In addition, MH tended to be the most conservative procedure flagging

fewer items than either SIBTEST or LR (this was not the case in Grade 3 where MH flagged the largest number of items but it was in Grades 6 and 9).

In Science, MH consistency identified fewer DIF items compared to SIBTEST and LR indicating that it was a more conservative procedure. Most items identified by MH were also identified by SIBTEST and LR. SIBTEST and LR flagged a larger number of DIF items, and they tended to flag the same items on the Science tests but agreement between the two procedures was not perfect.

#### Prevalence and Nature of DIF

Item classification produced by the three DIF procedures—Mantel-Haenszel (MH), Simultaneous Item Bias Test (SIBTEST), and Logistic Regression (LR)—was compared across the Mathematics and Science tests. Classification results across the procedures are summarized along the main diagonal in Table 4 for Mathematics and Table 5 for Science. In all comparisons that follow, items with a B- or C-level rating are considered DIF items whereas those with an A-level rating are not. Since the number of items per test and year of each test administration differed by grade and content area, the percentage of DIF items is also reported.

For the Grade 3 Mathematics achievement tests, MH identified eight (4.2%) DIF items, SIBTEST four (2.1%) DIF item, and LR six (3.2%) uniform and no non-uniform DIF items for the 1994 through 1997 administrations. For the Grade 6 Mathematics achievement tests, MH identified 13 (8.7%) DIF items, SIBTEST 15 (10.0%) DIF items, and LR 25 (8.0%) uniform and one (0.7%) non-uniform DIF item for the 1995 through 1997 administrations (the Mathematics achievement tests were not administered in 1994 at either Grades 6 or 9). For the Grade 9 Mathematics achievement tests, MH identified 13 (10.0%) DIF items, SIBTEST 21 (16.2%) DIF items, and LR 18 (13.9%) uniform and no non-uniform DIF items for the 1995 through 1997 administrations. Across all grades and administrations, 469 Mathematics items were analyzed. Of these items, MH identified 34 (7.3%) DIF items, SIBTEST found 40 (8.5%) items, and LR flagged 49 (10.5%) items with uniform DIF and one (0.2%) item with non-uniform DIF.

More DIF items were found in Science. Analyses were only conducted at Grades 6 and 9 because this test is not administered at Grade 3. For the Grade 6 Science achievement tests, MH identified 15 (7.4%) DIF item, SIBTEST 26 (13.8%) DIF items, and LR 35 (18.5%) uniform and no non-uniform DIF items for the 1995 through 1998 administrations. For the Grade 9 Science

achievement tests, MH identified 21 (14.0%) DIF items, SIBTEST 37 (24.7%) DIF items, and LR 39 (26.0%) uniform and two (1.3%) non-uniform DIF items for the 1995 through 1998 administrations. Across all grades and administrations, 428 Science items were analyzed. Of these items, MH identified 36 (8.4%) DIF items, SIBTEST found 63 (14.7%) items, and LR flagged 74 (17.3%) items with uniform DIF and two (0.5%) items with non-uniform DIF. Taken across both content areas, these results indicate that the majority of items do not display B- or C-level DIF, although Science tended to have more DIF than Mathematics. A comparison of the DIF items across the three statistical procedures in Mathematics and Science is shown in Figure 1.

LR has one noteworthy conceptual advantage over MH and SIBTEST: It is a model-based approach that can be used to identify uniform and non-uniform DIF. However, this advantage will only be realized if non-uniform DIF is present in test data. In addition, there is some debate about the prevalence of non-uniform DIF in student response data from large-scale assessments with some researchers claiming that non-uniform DIF is rare (Camilli & Shepard, 1994, p. 66). In the present study, non-uniform DIF was not prevalent over a large number of items in different content areas, grades, and years of administration suggesting that non-uniform DIF is rare in this type of student achievement test data.

### Conclusions and Policy Implications

The purposes of this study were to describe the test review process at Alberta Education, to identify items that function differently for males and females in Mathematics and Science, and to discuss some policy implications of this research. Conclusions and policy implications are presented in three related areas: Consistency among DIF statistical procedures, substantive outcomes, and test development.

#### Consistency Among DIF Statistical Procedures

The number of items flagged with DIF varied by statistical procedure across content area. For Mathematics, the results were relatively consistent. Correlations among the effect size measures were strong while classification consistency varied across grade and statistical procedure. MH, SIBTEST, and LR for uniform DIF flagged a comparable number of items across administrations (see Table 9) although MH tended to be more conservative than either SIBTEST or LR. Across all

grades and administrations, the proportion of DIF items in Mathematics flagged by each procedure was similar ranging from 7.3% with MH to 10.5% for LR-U.

For Science, the results were less consistent across the three procedures. Correlations among the effect size measures were strong and classification consistency was relatively high but not perfect. MH consistently flagged the smallest number of items, LR the largest number of items, and SIBTEST was in between (see Table 10). Across all grades and administrations, the proportion of DIF items flagged by each procedure was variable ranging from 8.4% with MH to 17.3% for LR-U.

Policy Implication: Results from the effect size correlations, classification consistency rates, and total item classifications indicate that the three DIF methods used in this study produce results that are relatively consistent but not identical. Moreover, these outcomes suggest that the three procedures produce relatively consistent effect size rankings and item classification, but that two procedures should be used to screen items for DIF. The results from this study indicate that MH is more conservative (i.e., flags fewer items) than either SIBTEST or LR. Policy makers and test developers should consider these results when choosing a DIF statistical approach since MH will likely identify fewer DIF items (i.e., MH will make fewer Type I errors) but possibly at the expense of power (i.e., MH will make more Type II errors). Both SIBTEST and LR consistently identify more DIF items. Researchers who are interested in studying DIF will likely accept a more powerful statistic even at the risk of identifying non-DIF items. Test developers and practitioners who are often pressed for time and resources may not accept such a trade-off and opt for a more conservative approach.

#### Substantive Outcomes

Two trends were apparent in this study. First, the number of DIF items tended to increase in Mathematics and Science as students progressed through school. Second, there is more DIF in Science than in Mathematics. In other words, when the total number of DIF items was assessed across all grades and administrations, the proportion of DIF items flagged by each procedure was higher in Science than in Mathematics.

Policy Implication: These trends have policy implications for the test review. For example, test developers should anticipate more DIF in Mathematics and Science at higher grades. Perhaps the review panel should be forewarned of this trend and encouraged to be more attentive to possible

gender DIF—and gender bias—in the higher grades. Test developers should also expect more DIF, overall, in Science compared to Mathematics.

Interpreting DIF becomes the next important step. Recall, DIF is not synonymous with bias. If the performance differences can be attributed to construct irrelevant test difficulty which unfairly affects the test performance for members of one group, then the item is biased. If, on the other hand, the performance difference can be attributed to actual knowledge and experience differences the test is designed to measure, then the outcome can be interpreted as item impact. The distinction between DIF, item bias, and item impact is important since DIF is a statistical concept and bias and impact are substantive concepts. Typically, explanations for DIF are sought from panels of content specialists who study the items and try to identify why some items are more difficult for one group of examinees compared to another (Berk, 1982; Ramsey, 1993). However, experience and research has shown that it is unusually difficult to account for DIF using judgmental analyses (e.g., Angoff, 1993; Camilli & Shepard, 1994; Englehard, Hansche, & Rutledge, 1990; Gierl & McEwen, 1998; O'Neill & McPeck, 1993). Thus, more research is needed to substantively interpret DIF statistical outcomes when males and females are compared. In the current study, it is not clear whether DIF can be attributed to item bias or item impact until content reviews with test developers and teachers are conducted in addition to cognitive analyses with male and female students solving DIF and non-DIF items. This research is currently being conducted by the authors.

#### Test Development

As previously reported, the number of DIF items tended to increase in Mathematics and Science from Grade 3 to 9 and more DIF was found in Science compared to Mathematics. Both trends were found despite the fact that Alberta Education uses the same test review in all grades and content areas.

Policy Implication: Statistical and substantive methods for detecting differential item functioning should be an essential part of test development and test evaluation efforts. Moreover, quantitative and qualitative analyses that can inform the test development process should be conducted after the administration of a test. These types of studies focusing on item, test, and DIF analyses are often not routine since they are guided by specific questions applicable to a particular content area or test administration. Yet, these studies are essential since they help establish a

feedback loop between developers and psychometrician so that information gained from each test administration can be used to improve the existing test development process. This point seems particularly applicable to DIF analyses since the factors that produce gender differences on tests are elusive and poorly understood. Currently, test reviews at Alberta Education are conducted before the tests are administered and statistical measures of DIF are not produced at all making it almost impossible to evaluate the effectiveness of the review or to improve the review process. Perhaps the test review should be conducted during test development using a set of reviewers, as is currently the policy, but also after the test is administered using the results from DIF statistical analyses to flag the DIF items and the same reviewers to interpret the DIF. This two-stage process may help identify sources of DIF that could be used to improve the test review (e.g., provide concrete, interpretable examples of item bias and impact), to sensitize developers and item writers to the sources of DIF, and to reduce the number of DIF items on a test.

## References

- Angoff, W. (1993). Perspective on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 3-24). Hillsdale, NJ:Lawrence Erlbaum.
- Berk, R. A. (Ed.). (1982). Handbook of methods for detecting test bias. Baltimore: Johns Hopkins Press.
- Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. Newbury Park, CA: Sage.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. Educational Measurement: Issues and Practice, 17, 31-44.
- Council of Ministers of Education. (1996). SAIP 1996 report on science. Toronto, ON: Council of Ministers of Education.
- Englehard, G., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. Applied Measurement in Education, 3, 347-360.
- Gierl, M., & McEwen, N. (1998, May). Differential item functioning on the Alberta Education Social Studies 30 Diploma Exams. Paper presented at the annual meeting of the Canadian Society for Studies in Education, Ottawa, Ontario, Canada.
- Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 129-145). Hillsdale, NJ: Erlbaum.
- Jodoin, M. (1999). Reducing type I error rates using an effect size measure with the logistic regression DIF procedure. Unpublished Masters thesis, University of Alberta, Edmonton, AB, Canada.
- Kimball, M. M. (1989). A new perspective on women's math achievement. Psychological Bulletin, 105, 198-214.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.
- Millsap, R. E., & Everson, H. T. (1993). Statistical approaches for assessing measurement bias. Applied Psychological Measurement, 17, 297-334.

O'Neill, K. A., & McPeck, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 255-276). Hillsdale, NJ:Lawrence Erlbaum.

Principles for Fair Student Assessment Practices for Education in Canada (1993). University of Alberta, Edmonton, AB.

Ramsey, P. A. (1993). Sensitivity reviews: The ETS experience as a case study. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 367-388). Hillsdale, NJ:Lawrence Erlbaum.

Robitaille, D. F., Taylor, A., & Orpwood, G. (1996). TIMSS-Canada Report, Volume 1: Grade 8. Vancouver, BC: University of British Columbia.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Applied Psychological Measurement, *17*, 105-116.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. Journal of Educational Measurement, *33*, 215-230.

Scheuneman, J. D., & Grima, A. (1997). Characteristics of quantitative word items associated with differential item functioning for female and black examinees. Applied Measurement in Education, *4*, 299-320.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. Psychometrika, *58*, 159-194.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of six procedures for detecting test item bias using both internal and external ability criteria. Journal of Educational Statistics, *6*, 317-375.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, *27*, 361-370.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.) Differential item functioning (pp. 337-347). Hillsdale, NJ: Erlbaum.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning: Logistic regression modeling as a unitary framework for binary and Likert-type item scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D., & Thomas, D. R. (1996, October). A measure of DIF effect size using logistic regression procedures. Paper presented at the National Board of Medical Examiners, Philadelphia, PA.

Table 1

Descriptive Statistics By Gender for the Mathematics Achievement Tests as a Function of Year of Administration

GRADE 3 MATH								
Characteristic	<u>1994</u>		<u>1995</u>		<u>1996</u>		<u>1997</u>	
	Males	Females	Males	Females	Males	Females	Males	Females
No. of examinees	6000	6000	6000	6000	6000	6000	6000	6000
No. of items	50	50	49	49	45	45	45	45
Mean	37.97	37.07	37.60	37.07	34.64	33.98	34.50	33.57
Standard Deviation	8.51	8.55	8.16	8.01	7.07	7.02	7.60	7.73
Skewness	-0.87	-0.74	-1.03	-0.86	-0.91	-0.80	-0.89	-0.68
Kurtosis	0.20	-0.05	0.80	0.34	0.50	0.37	0.37	-0.22

  

GRADE 6 MATH						
Characteristic	<u>1995</u>		<u>1996</u>		<u>1997</u>	
	Males	Females	Males	Females	Males	Females
No. of examinees	6000	6000	6000	6000	6000	6000
No. of items	50	50	50	50	50	50
Mean	35.86	34.74	37.56	36.59	35.93	34.26
Standard Deviation	8.76	8.80	8.31	8.20	8.37	8.46
Skewness	-0.67	-0.53	-0.81	-0.71	-0.52	-0.36
Kurtosis	-0.27	-0.43	0.04	-0.03	-0.41	-0.61

  

GRADE 9 MATH						
Characteristic	<u>1995</u>		<u>1996</u>		<u>1997</u>	
	Males	Females	Males	Females	Males	Females
No. of examinees	6000	6000	6000	6000	6000	6000
No. of items	40	40	45	45	45	45
Mean	30.51	28.75	34.70	33.14	32.71	31.64
Standard Deviation	9.90	9.88	10.90	11.02	11.30	11.17
Skewness	-0.24	-0.31	-0.34	-0.18	-0.16	-0.08
Kurtosis	-0.72	-0.88	-0.73	-0.89	-0.87	-0.89

Table 2

Descriptive Statistics By Gender for the Science Achievement Tests as a Function of Year of Administration

GRADE 6 SCIENCE								
Characteristic	<u>1995</u>		<u>1996</u>		<u>1997</u>		<u>1998</u>	
	Males	Females	Males	Females	Males	Females	Males	Females
No. of examinees	6000	6000	6000	6000	6000	6000	6000	6000
No. of items	49	49	50	50	50	50	50	50
Mean	41.53	40.38	32.53	31.38	32.23	30.54	34.94	33.66
Standard Deviation	9.30	9.16	7.50	7.24	7.64	7.40	8.15	8.21
Skewness	-0.63	-0.59	-0.59	-0.47	-0.42	-0.27	-0.66	-0.47
Kurtosis	-0.14	-0.09	0.12	-0.18	-0.22	-0.37	0.03	-0.33

  

GRADE 9 SCIENCE								
Characteristic	<u>1995</u>		<u>1996</u>		<u>1997</u>		<u>1998</u>	
	Males	Females	Males	Females	Males	Females	Males	Females
No. of examinees	6000	6000	6000	6000	6000	6000	6000	6000
No. of items	65	65	54	54	55	55	55	55
Mean	43.40	41.01	37.17	35.12	38.24	36.11	37.89	35.49
Standard Deviation	10.50	10.49	9.71	9.48	9.06	9.13	8.70	8.89
Skewness	-0.70	-0.40	-0.61	-0.35	-0.64	-0.40	-0.63	-0.40
Kurtosis	0.19	-0.43	-0.17	-0.48	-0.02	-0.43	0.08	-0.35

Table 3

Mean Correlation Coefficient and Standard Deviation Across Effect Size Measures for Three DIF Procedures as a Function of Content Area

	<u>Mathematics</u>				<u>Science</u>			
	<u>1994-1997</u>				<u>1995-1998</u>			
	$\Delta - MH$	$\hat{B}$	$R^2\Delta - U$	$R^2\Delta - N$	$\Delta - MH$	$\hat{B}$	$R^2\Delta - U$	$R^2\Delta - N$
$\Delta - MH$	—				—			
$\hat{B}$	-.95 (.03)	—			-.95 (.02)	—		
$R^2\Delta - U$	.84 (.09)	.82 (.09)	—		.83 (.06)	.87 (.07)	—	
$R^2\Delta - N$	-.04 (.15)	.02 (.17)	.05 (.38)	—	.04 (.12)	.06 (.15)	.15 (.21)	—

Note.  $\Delta - MH$  is Delta-Mantel-Haenszel;  $\hat{B}$  is the effect size measure in the Simultaneous Item Bias Test;  $R^2\Delta - U$  is  $R^2$  change for the Group variable in Logistic Regression associated with uniform DIF; and  $R^2\Delta - N$  is  $R^2$  change for the Total Score-by-Group Membership interaction term in Logistic Regression associated with non-uniform DIF. Both  $\Delta - MH$  and  $\hat{B}$  are directional tests. A positive  $\Delta - MH$  indicates DIF in favor of the Female examinees whereas the opposite is true for  $\hat{B}$ . Since  $R^2\Delta$  does not provide a directional test of DIF, the absolute value of  $\Delta - MH$  and  $\hat{B}$  are used when these effect size measures are correlated with the  $R^2\Delta$ . The standard deviation is shown in parentheses beside the mean correlation coefficient.

Table 4

DIF Counts for Mathematics as a Function of Grade

	<u>Grade 3 Total 1994 – 1997 (n=189)</u>				<u>Grade 6 Total 1995 – 1997 (n=150)</u>			
	MH	SIBTEST	LR-U	LR-N	MH	SIBTEST	LR-U	LR-N
MH	8 (4.2)				13 (8.7)			
SIBTEST	4	4 (2.1)			7	15 (10.0)		
LR-U	5	3	6 (3.2)		13	12	25 (8.0)	
LR-N	0	0	0	0 (0.0)	0	1	1	1 (0.7)

---

	<u>Grade 9 Total 1995 – 1997 (n=130)</u>			
	MH	SIBTEST	LR-U	LR-N
MH	13 (10.0)			
SIBTEST	10	21 (16.2)		
LR-U	11	13	18 (13.9)	
LR-N	0	0	0	0 (0.0)

Note. MH is Mantel-Haenszel; SIBTEST is the Simultaneous Item Bias Test; LR-U is logistic regression with uniform DIF, and LR-N is logistic regression with non-uniform DIF. The diagonal of each matrix indicates the total number of items flagged using each procedure and the off-diagonal indicate the number of matches across procedures. The number of items is on the top of each column and the percentage of DIF items is in parentheses along the main diagonal.

Table 5

DIF Counts for Science as a Function of Grade

	<u>Grade 6 Total 1995 – 1998 (n=189)</u>				<u>Grade 9 Total 1995 – 1998 (n=150)</u>			
	MH	SIBTEST	LR-U	LR-N	MH	SIBTEST	LR-U	LR-N
MH	15 (7.4)				21 (14.0)			
SIBTEST	9	26 (13.8)			18	37 (24.7)		
LR-U	11	22	35 (18.5)		19	31	39 (26.0)	
LR-N	0	0	0	0 (0.0)	0	0	0	2 (1.3)

Note. MH is Mantel-Haenszel; SIBTEST is the Simultaneous Item Bias Test; LR-U is logistic regression with uniform DIF, and LR-N is logistic regression with non-uniform DIF. The diagonal of each matrix indicates the total number of items flagged using each procedure and the off-diagonal indicate the number of matches across procedures. The number of items is on the top of each column and the percentage of DIF items is in parentheses along the main diagonal.

