

Identifying Content and Cognitive Skills that Produce
Gender Differences in Mathematics:
A Demonstration of the DIF Analysis Framework*

Mark J. Gierl
Jeffrey Bisanz
Gay L. Bisanz
Keith A. Boughton
University of Alberta

Paper Presented at the Annual Meeting of the
National Council on Measurement in Education (NCME)
at the Symposium entitled, "New Approaches for Identifying
and Interpreting Differential Bundle Functioning"

New Orleans, Louisiana
April 2-4, 2002

* This paper can also be downloaded from the CRAME website at <http://www.education.ualberta.ca/educ/psych/crame/>

Identifying Content and Cognitive Skills that Produce Gender Differences in Mathematics: A Demonstration of a DIF Analysis Framework

Gender differential item functioning (DIF) is a constant concern on large-scale standardized achievement tests in mathematics because differences between females and males are often found (e.g., Bielinski & Davison, 2001; Boughton, Gierl, & Khaliq, 2000; DeMars, 1998; Garner & Engelhard, 1999; Scheuneman & Grima, 1997; Willingham & Cole, 1997). DIF is present when examinees from different groups have a different probability or likelihood of answering an item correctly, after conditioning on ability. A number of "preferred" DIF statistical methods are available with theoretical strengths and empirical support that are well documented (Clauser & Mazor, 1998).

Statistical DIF analyses are often followed by judgmental reviews to identify *why* items are functioning differentially between groups (see, for example, Camilli and Shepard, 1994, p. xiii). Unfortunately, little progress has been made in understanding why DIF occurs using this approach (e.g., Camilli & Shepard, 1994, Englehard, Hansche, & Rutledge, 1990; Gierl, Bizanz, Bisanz, Boughton, & Khaliq, 2001; Gierl, Rogers, & Klinger, 1999, O'Neill & McPeck, 1993; Plake, 1980; Sudweeks & Tolman, 1993). For example, Roussos and Stout (1996) reviewed the DIF literature and concluded, "attempts at understanding the underlying causes of DIF using substantive analyses of statistically identified DIF items have, with few exceptions, met with overwhelming failure" (p. 360). More recently, authors of the 1999 *Standards for Educational and Psychological Testing* stated:

Although DIF procedures may hold some promise for improving test quality, there has been little progress in identifying the causes or substantive themes that characterize items exhibiting DIF. That is, once items on a test have been statistically identified as functioning differently from one examinee group to another, it has been difficult to specify the reasons for the differential performance or to identify a common deficiency among the identified items. (p. 78)

This impasse represents a fundamental problem in the study of group differences using DIF statistical methods.

To overcome this impasse, Roussos and Stout (1996) proposed the *multidimensionality-based DIF analysis paradigm*. It is a two-stage *confirmatory* approach intended to bridge the gap

between substantive and statistical analyses. The first stage is a substantive analysis in which DIF hypotheses are generated. To understand gender differences in mathematical achievement, for example, some researchers contend it is critical to identify specific content areas and cognitive skills that could produce these differences (Gallagher, 1998; Halpern, 1997; see also Gierl et al., 2001). The second stage is a statistical analysis where the DIF hypotheses are tested. To evaluate whether specific cognitive skills produce gender differences, for example, items from a large-scale mathematics achievement test that measure these skills could be identified in a judgmental review and then tested for DIF. A confirmatory DIF analysis provides better Type I error control than an exploratory DIF analysis (i.e., a DIF analysis where each item is tested individually) because only a small number of DIF hypotheses are tested. The DIF analysis framework also allows researchers to systematically identify and study the sources of DIF and begin to create a body of confirmed DIF hypotheses. (Stout & Roussos, 1995). These confirmed hypotheses, accumulated over studies, may lead to a better understanding of why DIF occurs.

In the present study we illustrate and evaluate an application of the Roussos-Stout (1996) DIF analysis framework to the study of gender differences in mathematics. Four characteristics distinguish this study from previous research. First, the substantive analysis was guided by past research on the cognitive and content-related sources of gender differences in mathematics achievement, as presented in the taxonomy by Gallagher, De Lisi, Holst, McGillcuddy-De Lisi, Morely, and Cahalan (2000; see also Gallagher, 1998; Gallagher & De Lisi, 1994). This approach enables us to use the Roussos-Stout framework to test the adequacy of the Gallagher et al. taxonomy in accounting for gender differences in mathematics achievement. Second, the substantive analysis was conducted by reviewers who were not only expert in the content domain but who, by virtue of their experience in tutoring, were highly knowledgeable about the diverse kinds of cognitive strategies students use to solve particular problems. If gender differences arise, in part, because males and females tend to use different solution strategies in response to certain problems characteristics (Gallagher & De Lisi, 1994), then reviewers must be sensitive to likely cognitive and content differences among problems. Third, three statistical methods were used to test DIF hypotheses. Each method is noteworthy because it provides a different kind of

evidence about these hypotheses. In particular, SIBTEST was used to test DIF hypotheses; DIMTEST was used to test the dimensionality of the data; and multiple linear regression was used to determine whether cognitive and content differences among items, as assessed using the Gallagher et al. taxonomy, predict gender-related DIF. Fourth, the study was conducted using data from a curriculum-based achievement test developed, in consultation with teachers, with the explicit goal of minimizing obvious, content-related gender differences. Our study is thus a particularly stringent test of the extent to which the Roussos-Stout approach, in combination with a substantive analysis based on the Gallagher et al. taxonomy, is adequate for detecting subtle gender differences that may reflect cognitive differences between females and males.

DIF Analysis Framework

The Roussos-Stout (1996) DIF analysis framework is based on the multidimensional model for DIF (MMD) proposed by Shealy and Stout (1993). MMD is a theoretical account for how DIF occurs. It is based on the premise that DIF is produced by multidimensionality. A dimension is a substantive characteristic of an item that can affect the probability of a correct response. The main construct the test is intended to measure is the primary dimension. DIF items are believed to elicit at least one dimension in addition to the primary dimension (e.g., Ackerman, 1992; Camilli & Shepard, 1994; Gierl & Khaliq, 2001; Kok, 1988; Lord, 1980; McDonald, 1999; 2000; Roussos & Stout, 1996; Shealy & Stout, 1993; Walker & Beretvas, 2001). Dimensions that produce DIF are referred to as secondary dimensions. Data are deemed to be multidimensional when primary and secondary dimensions characterize item responses.

Substantive Analysis

The first stage of the Roussos-Stout (1996) DIF analysis framework requires generating DIF hypotheses. A DIF hypothesis specifies whether a single item or bundle of items (i.e., two or more items sharing some important characteristic; see Douglas, Roussos, & Stout, 1996; Gierl et al., 2000; Nandakumar, 1993) designed to measure the primary dimension also measures a secondary dimension, thereby producing group differences. To decide whether the data contain distinct dimensions, *organizing principles* are used to identify items or bundles of items that share certain characteristics. Gierl et al. (2001) described four general organizing principles that can

guide the substantive analysis. These principles include content-related properties (e.g., items may be bundled according to content categories identified by experienced reviewers), psychological characteristics (e.g., items that appear to elicit particular problem-solving strategies may be bundled), test specifications (e.g., items may be bundled according to categories outlined during test development), or empirical outcomes (e.g., items may be bundled according to outcomes from statistical analyses of the data). Although relatively unexplored, the use of organizing principles to create bundles greatly enhances the potential of DIF methods for identifying *and* interpreting group differences on standardized achievement tests.

Recently, Gallagher et al. (2000) presented a taxonomy of content and cognitive characteristics to account for gender differences in mathematics (also see Gallagher, 1998). The taxonomy was based on outcomes reported in the educational and psychological literature. According to the Gallagher et al. taxonomy, females should perform better than males on items with contextual characteristics likely to be more familiar or interesting to females, on items that require a high level of verbal skill, and on items that require mastery of mathematical content. Males should perform better than females on items that have contextual characteristics likely to be more familiar or interesting to males, on items that are likely to place heavy demands on spatial skills, and on items that have multiple solution paths.

To evaluate the taxonomy, Gallagher et al. (2000) asked two mathematicians with test development experience to code the content and cognitive attributes for items from the Graduate Record Exam-Quantitative section (GRE-Q). Standardized mean differences between males and females were used to compute a gender effect-size measure with data from four administrations of the GRE-Q to students majoring in three different areas (social sciences, arts and humanities, or technical sciences). Gallagher et al. found that males outperformed females on all items, but generally the effect-size difference was greater for items expected to favor males than for those expected to favor females. Gallagher et al. interpreted these results as support for their taxonomy because items were classified according to categories that produced gender differences in performance on the GRE-Q. This conclusion is not entirely compelling, however, for four reasons. First, items with characteristics that favored both males and females were

classified arbitrarily as favoring males. This method of collapsing categories may have obscured possible gender differences consistent with or contrary to the authors' hypothesis. Second, the expected pattern of results was statistically significant in only seven of 12 independent tests of the hypothesis (four types of students x three different administrations). The sources of this inconsistency are not clear. Third, although Gallagher et al.'s taxonomy contained six categories—three favoring males and three favoring females—the data were collapsed into only one superordinate category for male-favoring items and another for female-favoring items. Thus the analyses are insufficient for evaluating the full taxonomy. Fourth, the finding that male-female differences were less striking on items supposedly favoring females than those supposedly favoring males does not confirm that some items favor females. An alternative hypothesis, that items differ only in the extent to which they favor males and no items favor females, is not precluded by the data. Because males outperformed females on both types of items, these two hypotheses cannot be distinguished with statistical tests based on mean differences. A more sensitive approach would be to use analyses based on differential item functioning procedures, in which gender differences on individual items can be assessed after overall ability is controlled. For these reasons the Gallagher et al. taxonomy has not been tested thoroughly. In the present study we use the taxonomy as the organizing principle to illustrate the substantive component of the DIF analysis framework and, in the process, evaluate whether the content areas and cognitive skills identified by Gallagher et al. differentiates females and males on tests of mathematics achievement.

Statistical Analysis

The second stage in the Roussos-Stout DIF analysis framework is statistically testing the DIF hypotheses. Statistical analyses are used to see whether the data, so structured using the organizing principle, reveal distinct primary and secondary dimensions. Two methods will be used in the current study to evaluate the dimensional structure of the data.

First, SIBTEST was used to quantify the size of DIF and test DIF hypotheses (Shealy & Stout, 1993; Stout & Roussos, 1995). To operationalize SIBTEST, items on the standardized achievement test are divided into the studied subtest and the matching subtest. The studied

subtest contains the items *believed* to measure the primary and secondary dimensions based on the substantive analysis in the first stage, whereas the matching subtest contains the items *believed* to measure only the primary dimension. The matching subtest is used to place females and males into subgroups at each score level so their performance on items from the studied subtest can be compared. The amount of DIF in the studied subtest is reflected in the parameter estimate, $\hat{\mathbf{b}}_{UNI}$. SIBTEST is used to assess this parameter estimate with the test statistic,

$$SIB = \frac{\hat{\mathbf{b}}_{UNI}}{\hat{\mathbf{s}}(\hat{\mathbf{b}}_{UNI})},$$

where $\hat{\mathbf{s}}(\hat{\mathbf{b}}_{UNI})$ is the standard error of $\hat{\mathbf{b}}_{UNI}$. Shealy and Stout (1993) demonstrated that SIB has a normal distribution with mean 0 and variance 1 under the null hypothesis of no DIF. The null hypothesis is rejected if SIB exceeds the 100 $(1 - \alpha / 2)$ percentile point from the standard normal distribution. A technical description of SIBTEST is found in Shealy and Stout (1993).

Second, DIMTEST will be used to test the dimensional structure of the data. Shealy and Stout (1993) attributed DIF to multidimensionality where DIF items are believed to elicit at least one dimension in addition to the primary dimension. SIBTEST is used to estimate and test the DIF effect size measure $\hat{\mathbf{b}}_{UNI}$ based on a substantive analysis of the items believed to measure the primary and secondary dimensions. However, SIBTEST is not a direct test of dimensionality. DIMTEST, by comparison, yields a direct test of dimensionality. DIMTEST, with a refined bias correction method (Froelich, 2000; also see Froelich & Habing, 2001), was used to determine if the studied subtest items were dimensionally distinct from the matching subtest items. The refined bias correction method is based on a bootstrap procedure that eliminates the need for the assessment subtest 2, which was required in the original DIMTEST for bias correction (Stout, 1987). Froelich (2000) demonstrated that the new DIMTEST procedure provides better Type I error control and increases power compared to the original DIMTEST procedure in a variety of simulation conditions. To operationalize DIMTEST, items on the standardized achievement test are divided into the assessment subtest (AT) and the partitioning subtest (PT). In the current study, AT contains the studied subtest items (i.e., subtest containing DIF items believed to be

dimensionality homogeneous with one another but distinct from matching subtest). PT contains the matching subtest items. Under the assumption of unidimensionality, the DIMTEST statistic is,

$$T = \frac{T_L - \overline{T}_G}{\sqrt{1 + 1/N}}$$

where T_L is original DIMTEST test statistic, \overline{T}_G is an estimate of the DIMTEST test statistic averaged across simulations using data generated from a specified examinee ability distribution with the nonparametric item response functions estimated from each item in the dimensionality analysis, and N is the number of simulations. Froelich (2000) demonstrated that T has a normal distribution with mean 0 and variance 1 under the null hypothesis of unidimensionality. The null hypothesis is rejected if T exceeds the 100 $(1 - \alpha / 2)$ percentile point from the standard normal distribution. A technical description of DIMTEST with the new bias correction method is found in Froelich (2000).

For the SIBTEST and DIMTEST analyses, problems were characterized by a single category in the Gallagher et al. (2000) taxonomy. That is, a problem may be classified as requiring spatial skills or verbal skills, but not both. However, mathematical problems often contain several different characteristics (e.g., Mabbott & Bisanz, submitted) that may elicit diverse strategies. To the extent that no single characteristic is dominant in every problem, conventional DIF analyses may be misleading. To determine if the results from SIBTEST and DIMTEST were insensitive to the influence of multiple features within single problems, multiple regression was used. In particular, the salience of each category from the Gallagher et al. taxonomy was used to predict gender differences across problems, as indexed with \hat{b}_{UNI} .

Method

Student Sample and Achievement Tests

Data from the 1996 and 1997 administrations of a Grade 9 mathematics achievement test were analyzed in this study. The achievement tests were developed and administered in the Canadian province of Alberta by the government ministry in-charge of education, *Alberta Learning*. Data from 6000 females and 6000 males were selected randomly from the 1996 and

1997 databases, and analyzed in this study. Each database contained, approximately, 36000 students.

Each test contained 45 multiple-choice items and 10 numeric response items. All items were scored dichotomously. In addition, twenty-six multiple-choice and six numeric-response items were identical for the 1996 and 1997 administrations. Items were classified into four curricular content areas--number, patterns and relations, shape and space, and statistics and probability.

The tests were developed using a comprehensive process with many quality control checks designed to screen out content-related gender differences. Items were developed by teachers nominated to serve on item-writing committees. All items are based on concepts, topics, and facts from the province-wide curriculum (Alberta Learning, 1996). The items undergo an internal review before they are field tested. Members of the internal review committee are from the achievement testing unit at Alberta Learning. The purpose of the internal review is to examine the field test items 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. Once reviewed, the field test items are administered throughout the province to participating schools. Teachers and students are encouraged to comment on the field test items. Field test results are then evaluated using a classical item analysis. These sources of information—comments of the internal review committee, teachers', and students' along with the statistical results—are considered by the test developers during the selection of the final test items. The final form of the mathematics test is then created. A second internal review is conducted after the test developer has made his or her final selection of test items. The committee for this review includes the achievement test developers and five teachers. 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 balance the gender-related content by equalizing the number of pictures or references to males and females. 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. The strengths of this test development process were documented and recognized in a review by the General Accounting Office (General Accounting Office, 1993).

Substantive Analysis

The study began with a comprehensive review of the 1996 and 1997 mathematics items. Two reviewers with diverse skills were recruited specifically for this study. The two reviewers were not only familiar with the content area and achievement tests but also with the way students solve problems through their tutoring experiences and educational training. The first reviewer was a third-year female undergraduate engineering student who completed her secondary education in Alberta. She had extensive coursework in mathematics, she was familiar with the provincial achievement testing program (having written the exams as a secondary school student), and she had extensive experience tutoring secondary school students in mathematics. The second reviewer was a second-year male graduate student in educational measurement who completed his undergraduate degree in mathematics education. He also completed his secondary school education in the province of Alberta. Much like the first reviewer, he had extensive experience with mathematics, he was familiar with the provincial achievement testing program, and he tutored secondary school students in mathematics. He was also familiar with the provincial mathematics curriculum through his undergraduate teacher training.

To begin, the reviewers were told that females and males may have different mathematical problem-solving skills and, consequently, may solve the same item in different ways; that specific item characteristics may be differentially appealing to females and males, thereby producing different responses; and that researchers had identified some of the item characteristics believed to produce gender differences in mathematics. However, the reviewers were *not* told which content areas and cognitive skills in the Gallagher et al. (2000) taxonomy were believed to favor females or males. Then, the reviewers received a training session prior to classifying the 1996 and 1997 test items in which they worked independently and applied the Gallagher et al. (2000) taxonomy to a sample of items from previously administered mathematics achievement tests to practice the classification task. Once the independent classification was complete, the reviewers

met to discuss their results with one another and with the authors of this study. All disagreements were discussed, debated, and resolved as a way of ensuring the categories in the Gallagher et al. (2000) taxonomy were interpreted in the same manner by each reviewer.

During the training session, some noteworthy changes were made to the Gallagher et al. (2000) taxonomy. Gallagher et al. identified six categories in their taxonomy that were expected to produce gender differences. Five of the Gallagher et al. categories (A1-3, B1-2 in Table 1) were used in this study with little modification.

However, our reviewers found Gallagher et al.'s (2000) sixth category, mastery of mathematics content, difficult to apply because it was too inclusive. As a result, this category was split into three mutually exclusive categories: the first involved the application of routine mathematical solutions to new, *unfamiliar* situations; the second involved application of routine mathematical solution to *familiar* situations; the third involved symbolic processes. The reviewers also identified the need for a new cognitive category, memorization, for items that required examinees to recall key information. The modified taxonomy, summarized in Table 1, was used for item classification.

Next, all items from the 1996 and 1997 tests were classified by the reviewers. For analyses with SIBTEST and DIMTEST, reviewers were asked to identify the *most salient* characteristic for each item using the modified Gallagher et al. (2000) taxonomy. This summary allowed us to classify each item into one of the nine categories. For the multiple regression analyses, reviewers were asked to rate *all salient* characteristics associated with each item on a 4-point scale rating from 1-Not at All Salient to 4-Very Salient. This summary allowed us to classify each item according to one or more of the nine categories. The reviewer rating form is presented in Appendix A.

Disagreements between reviewers were infrequent (11 of 110 items) and resolved in consultation with the authors. Six and five items from the 1996 and 1997 administrations, respectively, were scrutinized by the two reviewers and the four authors. Consensus was not reached for three items from the 1996 and two items from the 1997 administrations, and these items were excluded.

Statistical Analysis

Once the substantive reviews of the math items were completed, the items were sorted into each category using the reviewers' rating of the *most salient* characteristic and the items were tested using SIBTEST and DIMTEST. A four-step procedure was used to identify dimensions that elicited gender differences (see Gierl et al., 2001). First, all DIF items were identified with SIBTEST using a single item analysis (i.e., studying one item at a time and using the remaining items as the matching subtest) to obtain the DIF effect size measure, \hat{b}_{UNI} , for each item. Second, items were grouped by the nine cognitive categories using the reviewers' classification and the \hat{b}_{UNI} values for these items were graphed by category. Third, bundles were identified by visually examining the graphs and looking for interpretable patterns. Fourth, the interpretable bundles were tested. For the SIBTEST analysis, the studied subtest contained the bundle associated with a specific category that produced a gender difference. The matching subtest contained bundles with no identifiable systematic source of gender difference. For the DIMTEST analysis, AT contained the bundles associated with a specific category that produced a gender difference. PT contained bundles with no identifiable systematic source of gender difference. That is, AT and PT in the DIMTEST analysis contained the same bundles as the studied and matching subtests in the SIBTEST analysis.

A multiple regression analysis was also conducted using the reviewers' ratings of *all salient* characteristics associated with each item. In this analysis, \hat{b}_{UNI} served as the dependent variable and the reviewers' saliency rating for each category served as the independent variables.

Results

Psychometric Characteristics of the Achievement Tests

The psychometric characteristics on the mathematics achievement tests are presented in Table 2 to provide the reader with a brief summary of the achievement tests used in this study. Four general outcomes are highlighted. First, mean test performance was comparable within and between groups for the 1996 and 1997 administrations. Males outperformed females in both years, although the differences were small. Second, skewness and kurtosis were similar

between the two groups for each test across years, indicating that the shape of the distributions were comparable between females and males. Third, overall item difficulty and discrimination values were similar for both groups across years. Fourth, internal consistency was also comparable within groups across years. Thus, the test developers were successful at minimizing gender differences, at least at the level of test scores.

Analyses Using Rating of Most Salient Characteristics

Frequency of most salient characteristics. Table 3 contains the frequency of the most salient characteristics. For the items on both the 1996 and 1997 tests, the most salient characteristic was application of routine mathematical solutions to familiar situations (42% and 40%, respectively) followed by verbal skills (23% and 26%). No items were judged to have contexts favoring males or females. Thus test developers were successful in eliminating gender-based differences in content. More generally, the uneven distribution of items across categories can be expected when analyses are based on organizing principles not used in the development of the test.

SIBTEST results. The category bundles for the 1996 and 1997 mathematics achievement tests are shown in Figure 1. The x-axis shows the nine categories and the y-axis represents the \hat{b}_{UNI} value for each item. Positive \hat{b}_{UNI} values favor males, whereas negative values favor females. The items from the 1996 administration are open circles whereas the items from the 1997 administration are solid circles. From this analysis, bundles associated with spatial and memorization skills are apparent and favor males and females, respectively. Items for the remaining five cognitive categories are evenly distributed, for the most part, between the two groups revealing no systematic gender differences.

Table 4 contains the outcomes from the statistical tests of the two interpretable bundles using the 1996 and 1997 results. In these analyses, the two studied subtests contained items associated with spatial skills and memorization, respectively. The matching subtest contained items from the remaining bundles. All hypotheses were tested with a directional test using an alpha level of 0.05. Males in Grade 9 performed better than females with comparable

mathematics test scores on items requiring spatial skills. Females in Grade 9 performed better than males with comparable mathematics test scores on items requiring memorization skills.

The stability of this outcome was also evaluated by comparing the 32 common items across the 1996 and 1997 results. Recall, both 1996 and 1997 tests were created using the same test specifications and development process. The results are shown in Figure 2. Items from 1996 are shown as open circles and items from 1997 are shown as solid circles. Bundles associated with spatial and memorization skills, again, are apparent and favor males and females, respectively, whereas common items in the remaining three cognitive categories are evenly distributed between the two groups. These results reveal that differential bundle functioning (DBF) is stable over a two-year period using a different group of examinees, and they support the interpretation that males and females systematically differ in two categories.

DIMTEST results. SIBTEST was used to estimate and test the DIF effect size measure, \hat{b}_{UNI} , using the modified Gallagher et al. (2000) taxonomy where the data were structured and tested according to primary and secondary dimensions. SIBTEST, however, is not a direct test of dimensionality. DIMTEST, by comparison, yields a direct test of dimensionality. Therefore, to evaluate this assumption about dimensionality, DIMTEST was used. The dimensional distinctiveness of the spatial and memorization subtests was evaluated against the remaining items that formed the matching subtest. The results are presented in Table 5. The spatial subtest was dimensionally distinct from the matching subtest for both 1996 and 1997. The memorization subtest, on the other hand, was not dimensionality distinct from the matching subtest in either year. Thus DIMTEST only confirmed one of the dimensions implied by the SIBTEST analyses.

Analyses Using Rating of All Salient Characteristics

Frequency of all salient characteristics. SIBTEST and DIMTEST are based on the assumption that each item can be classified in a single, dominant salient category. Test items can be associated with more than one salient category, however. Table 6 contains the frequency of all

were verbal skills (28% in 1996 and 30% in 1997) and application to familiar situations (22% and 26%). When categories other than the most salient are counted, the frequency of problems with shortcuts becomes notably larger. Items with female content still do not exist, and only 3% of problems have male content that is judged to be salient.

Multiple regression results. Finally, to evaluate whether the reviewers' rating of all salient characteristics would predict \hat{b}_{UNI} , multiple regressions were conducted. The reviewers' saliency rating for each item in each category, ranging from 1-Not at All Salient to 4-Very Salient, served as the independent variables. As evident in Table 6, the distribution of items across categories was uneven and some categories were present rarely or not at all. Consequently, categories that were infrequently represented (fewer than 10% of the total number of judgments) were deleted from analyses. In all cases, the categories excluded contained items that were not related to \hat{b}_{UNI} (r 's ranged from -0.16 to 0.17, p 's > 0.20). In the analysis of 1996 data, five independent variables (shortcuts, spatial, verbal, applications unfamiliar, applications familiar; see Table 6) accounted for 17% of the sums of squares in \hat{b}_{UNI} , $F(5, 46) = 1.87$, $p = .12$. Spatial ability was the only significant predictor, $t(1, 46) = 2.81$, $p < .05$, and it favored males. The four independent variables used for the 1997 analysis (shortcuts, spatial, verbal, applications familiar) accounted for 14% of the sums of squares, $F(4, 48) = 2.01$, $p = .11$. Males performed better than females on items requiring spatial ability, $t(1, 48) = 2.13$, $p < .05$, and verbal ability, $t(1, 48) = 2.09$, $p < .05$.

Conclusions and Discussion

In the present study we illustrated an application of the Roussos-Stout (1996) DIF analysis framework to the study of gender differences in mathematics. The DIF analysis framework was used to study content and cognitive skills believed to produce gender DIF in mathematics, as described by Gallagher et al. (2000). The DIF analysis framework requires a two-stage approach. The first stage is a substantive analysis. In the current study, two reviewers used a *modified* version of the Gallagher et al. (2000) taxonomy to classify the content and cognitive characteristics of items taken from two large-scale standardized achievement tests in mathematics. The reviewers' judgments were based on their knowledge of the content area, the

our conclusion must be tempered. We used mathematics test that were designed with the explicit goal of minimizing obvious, content-related gender differences. Consequently, the tests in our study had an uneven distribution of items across categories which is to be expected when the organizing principle was not used to guide test development. Different results might be found using tests designed without the explicit goal of minimizing gender differences.

Using the Roussos-Stout Framework

Substantive analysis. The tests of differential bundle functioning and dimensionality used in this study require that each item be associated with only one category in the Gallagher et al. (2000) taxonomy. It should be noted, however, that this coding requirement tends to oversimplify the cognitive complexity elicited by the test items. We demonstrated, for example, that the mathematics items contained multiple content and cognitive characteristics, as items elicited an average of 2.6 salient categories. This finding strongly suggests item classification that requires a mutually exclusive rating is an oversimplification of the categories actually needed to characterize mathematics items. To overcome this limitation, the reviewers were asked to identify all characteristics associated with each item and rate the salience of each characteristic on a 4-point scale. In this case, classifying items according to single or multiple categories did not substantially affect our conclusions about gender differences.

A far more problematic outcome in the substantive analysis stemmed from the fact that item coding is often strategy dependent. Our reviewers revealed that ratings were based on the strategy *most likely* to be used by students. This inference was required to classify items into one category even though the reviewers identified multiple strategies that could be used to correctly solve some items. To illustrate this point, an example item is presented in Appendix B. Our reviewers believed this item could be solved by memorizing the concept or by drawing a diagram. If the reviewers believe that memorization is the more likely strategy, (i.e., the examinee remembered that a z-score of 2.0 is always two standard deviations below the mean), then they would classify the item as memorization. Alternatively, if the reviewers believe that a spatial strategy is more likely to be used (i.e., the examinee sketched a diagram of the normal distribution and shaded the region of the distribution associated with a z-score of -2.0), then they

would classify the item as spatial. Thus, coding depends on the reviewer's judgment about which strategy is used more frequently. This approach is problematic because the reviewers can make an erroneous judgement about strategy use or students can use both strategies. The consequence in either case is that the final code misrepresent the characteristics of the items when solved by students using different strategies. This problem is not overcome by using multiple codings per item because different strategies can be linked with different categories. For instance, when the reviewers deemed the memorization strategy to be more likely for the Appendix B item, the salience categories were verbal skills, application of routine mathematical solutions to familiar situations, and memorization. However, when the reviewers believed that a spatial strategy was more likely, the salient categories were spatial ability, verbal skills, and application of routine mathematical solutions to familiar situations. This outcome shows that any rating can misrepresent the content and cognitive skills elicited by an item when the item can be solved correctly using different strategies.

Statistical Analysis. The value of using multiple statistical methods is demonstrated by the finding that different procedures yielded converging evidence about spatial ability but diverging evidence about memorization and verbal skills. The use of multiple methods is particularly important when clear criteria are lacking for interpreting statistical results, as is currently the case when SIBTEST is used for DBF analyses.

Potenza and Dorans (1995) note that, to be used effectively, a DIF detection procedure needs an interpretable effect size measure. Effect size measures are commonly used to increase the interpretability of DIF analyses and guidelines exist for interpreting many DIF effect size measures (e.g., Dorans, 1989, p. 226; Jodoin & Gierl, 2001; Shealy & Stout, 1993, p. 181; Zieky, 1993, p. 342). Unfortunately, no guidelines exist for interpreting \hat{b}_{UNI} on a bundle of items and, as a result, there is no agreement on how to distinguish statistical from practical significance in DBF research. Creating interpretative DBF guidelines is further complicated by the potential for amplification. Amplification occurs when small but systematic performance differences on single items combine to produce a large performance difference for a bundle of items. This outcome implies that small item-level differences, which may go unnoticed, can be magnified when the

same difference is evaluated with a bundle (Nandakumar, 1993). To overcome this limitation, we supplemented the SIBTEST analysis using DIMTEST to evaluate the dimensional structure of the data that is merely assumed in SIBTEST. These results were consistent across statistical methods for the spatial but not the memorization dimension. This inconsistency might be attributed to lack of adequate effect size criteria for interpreting the DBF analyses using SIBTEST. To overcome the reliance on statistical significance testing, research is needed to identify and evaluate supplemental effect size criteria to promote DBF interpretations that are substantively meaningful.

References

- Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.
- Alberta Education. (1996). *Program of Studies: Mathematics*. Edmonton, AB: Curriculum Standards Branch, Alberta Education.
- Bielinski, J., & Davison, M. L. (2001). A sex difference by item difficulty interaction in multiple-choice mathematics items administered to national probability samples. *Journal of Educational Measurement, 38*, 51-77.
- Boughton, K., Gierl, M. J., & Khaliq, S. (2000, May). *Differential bundle functioning on mathematics and science achievement tests*. Paper presented at the annual meeting of the Canadian Society for Studies in Education, Edmonton, Alberta, Canada.
- Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Newbury Park: Sage.
- Casey, M. B., Nuttall, R., Pezaris, E., & Benbow, C. P. (1995). The influence of spatial ability on gender differences in mathematics college entrance test scores across diverse samples. *Developmental Psychology, 31*, 697-705.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*, 31-44.
- Douglas, J., Roussos, L., and Stout, W., (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. *Journal of Educational Measurement, 33*, 465-484.
- DeMars, C. E. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. *Applied Measurement in Education, 11*, 279-299.
- Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2*, 217-233.

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.

Froelich, A. G. (2000). *Assessing the unidimensionality of test items and some asymptotics of parametric item response theory*. Unpublished Doctoral Dissertation. University of Illinois at Urbana-Champaign, Department of Statistics.

Froelich, A. G., & Habing, B. (2001). *Refinements of the DIMTEST methodology for testing unidimensionality and local independence*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Gallagher, A. M. (1998). Gender and antecedents of performance in mathematics testing. *Teachers College Record, 100*, 297-314.

Gallagher, A. M., & De Lisi, R., (1994). Gender differences in Scholastic Aptitude Test-Mathematics problem solving among high-ability students. *Journal of Educational Psychology, 86*, 204-211.

Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-De Lisi, A. V., Morely, M., & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology, 75*, 165-190.

Garner, M., & Engelhard, G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education, 12*, 29-51.

General Accounting Office (1993). *Educational testing: The Canadian experience with standards, examinations, and assessments* (GAO/PEMD-93-11). Washington, DC: U.S. General Accounting Office.

Gierl, M. J., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*, 26-36.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests. *Journal of Educational Measurement, 38*, 164-187.

- Gierl, M. J., Rogers, W. T., & Klinger, D. (1999). Using statistical and substantive reviews to identify and interpret translation DIF. *Alberta Journal of Educational Research, 45*, 353-376.
- Halpern, D. F. (1997). Sex differences in intelligence: Implications for education. *American Psychologist, 52*, 1091-1102.
- Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent methodological questions in educational testing. *Review of Research in Education, 24*, 393-446.
- Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349.
- Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine and J. Rost (Eds.), *Latent trait and latent class models*, (pp. 263-274). New York: Plenum Press.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Mabbott, D., & Bizanz, J. (submitted). *Developmental change and individual differences in children's multiplication*.
- McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Erlbaum.
- McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24*, 99-114.
- Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 16*, 159-176.
- 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.
- Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the validation process. *Educational and Psychological Measurement, 40*, 397-404.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23-37.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.

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.

Standards for Educational and Psychological Testing. (1999). Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

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

Stout, W. & Roussos, L. (1995). *SIBTEST manual*. University of Illinois: Department of Statistics, Statistical Laboratory for Educational and Psychological Measurement.

Sudweeks, R. R., & Tolman, R. R. (1993). Empirical versus subjective procedures for identifying gender differences in science test items. *Journal of Research in Science Teaching, 30*, 3-19.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF analysis paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement, 38*, 147-163.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Erlbaum.

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.

Appendix B

A sample mathematics item that could be solved using a memorization or a spatial strategy.

2. The results from a Biology examination are normally distributed. A student is told that his mark on the exam corresponds to a z-score of -2.0 but not told the mean, standard deviation, or actual mark. Which conclusion is **always** true in this situation?
- A. The student scored above 50%.
 - B. The student scored below 50%.
 - C. The student scored above the mean.
 - D. The student scored below the mean.

Author Note

Mark J. Gierl, Centre for Research in Applied Measurement and Evaluation, Department of Educational Psychology, 6-110 Education North, Faculty of Education, University of Alberta, Edmonton, Alberta, Canada, T6G 2G5

Jeffrey Bisanz, Centre for Research in Child Development, Department of Psychology, University of Alberta, Edmonton, AB, Canada, T6G 2E9

Gay L. Bisanz, Centre for Research in Child Development, Department of Psychology, University of Alberta, Edmonton, AB, Canada, T6G 2E9

Keith A. Boughton, Educational Testing Service, Mailstop O3T, Princeton, New Jersey, U.S.A., 08541

This research was supported with funds awarded to the first three authors from the Social Sciences and Humanities Research Council of Canada. Please address correspondence to Mark J. Gierl, Centre for Research in Applied Measurement and Evaluation, 6-110 Education Centre North, Faculty of Education, University of Alberta, Edmonton, Alberta, Canada, T6G 2G5.
Email: mark.gierl@ualberta.ca

Table 1

The Modified Gallagher et al. (2000) Outlining the Content and Cognitive Skills Expected to Produce Gender Differences in Mathematics

A. Knowledge and Skills Favoring Males

- 1) Item Context Favoring Males

Solving the problem requires material more likely to be familiar to males (e.g., items requiring knowledge about traditionally males activities like race cars or football).
 - 2) Short-cuts/ Multiple solution paths
 - a) Multiple solution paths meaning more than one solution path leads to a correct answer. The quick solution may be imaginative or insightful (but does not involve drawing a picture). The slower solution may be more systematic and planful. Significant time savings to the solution is the key feature for this category.
 - b) Test-taking skills can contribute to the faster or more accurate solution. By test-taking skills, we mean that examinees use the characteristics and formats of the items to improve their score by, for example, using other items to find clues, definitions, or algorithms for solutions to the current item.
 - c) The context looks like a familiar one, but the solution is not one that is generally associated with the context (e.g., on the first glance the problem appears to deal with averages, but to solve it one needs to use a rate of growth).
 - 3) Spatial
 - a) Requires the conversion of a word problem to a spatial representation (i.e., generation of spatial format). Spatial representation is an important part of the problem.
 - b) Requires using a given spatial representation (e.g., convert it to a mathematical expression or extract information to be used in solving a problem). Spatial representation is an important part of the problem.
 - c) Requires the transformation of information presented in a spatial format to a different spatial format (e.g., a given parabola has to be modified according to some rules). The change has to be produced.
 - d) Spatial information must be maintained in “working memory” while other spatial information is being transformed (e.g., maintain a particular shape in working memory so that it can be compared with a transformed shape). Working memory refers to the information we activate and use when solving problems. Working memory can become overloaded, resulting in errors, when there simply are too many pieces of information to keep track of simultaneously. Also, information can be lost from working memory over time.
 - e) Multiple solution paths meaning more than one solution path leads to a correct answer. One or more of the likely solutions involves drawing or using a picture.
-

Table 1 (con't)

B. Knowledge and Skills Favoring Females

1. Item Context Favoring Females
Solving the problem requires material more likely to be familiar to females (e.g., items requiring knowledge about traditionally females activities like the cost of family care or interpersonal relationships).
 2. Verbal
 - a) Requires the conversion of a word problem to an algebraic expression. These items require the conversion only. This category does not include items where a mathematical expression is generated as a step in arriving at a solution to the problem.
 - b) Verbal information must be maintained in working memory while additional information is being processed; primarily used for items with heavy verbal load.
 - c) Reading math (e.g., using a newly defined function or understanding the properties of an algebraic expression).
 3. Application of Routine Mathematical Solutions to New, Unfamiliar Situation
 - a) Requires labeling the problem as a specific type of problem and/or retrieving a formula or routine that should be known from memory, but is not immediately apparent.
 - b) The problem is multi-step and requires accuracy and a systematic approach. For example, two successive calculations must be done and the second calculation uses information from the first calculation in a new, unfamiliar situation
 4. Application of Routine Mathematical Solution to Familiar Situation
 - a) The context is a familiar one, frequently seen in mathematics course work; the solution path is one that is generally associated with the context.
 - b) The problem is multi-step and requires accuracy and a systematic approach. For example, two successive calculations must be done and the second calculation uses information from the first calculation but in a familiar situation
 5. Memorization
Recall of definitions, terms, formulas, and mathematical facts necessary to solve the problem. For example, the item requires that the examinee know the properties of an arithmetic sequence, the eccentricity of a parabola, the radius of a circle, or the properties of conics.
 6. Symbolic Processes
 - a) Solution requires pure algebraic manipulation or calculation.
 - b) Questions where two mathematical expressions or quantities must be compared and the values of the two are equal (this type of problem has no verbal element).
-

Table 2

Psychometric Characteristics for the Grade 9 Mathematics Achievement Tests

Characteristic	<u>1996</u>		<u>1997</u>	
	Males	Females	Males	Females
No. of Examinees	6000	6000	6000	6000
No. of Items	55	55	55	55
Mean	34.70	33.14	32.71	31.64
SD	10.90	11.02	11.30	11.17
Skewness	-.34	-.18	-.16	-.08
Kurtosis	-.73	-.89	-.87	-.89
Mean Item Difficulty	.63	.60	.59	.58
SD Item Difficulty	.15	.16	.15	.16
Mean Item Discrimination ^a	.53	.52	.54	.53
SD Item Discrimination	.13	.13	.13	.14
Internal Consistency ^b	.92	.92	.92	.92

^aBiserial correlation^bCronbach's alpha coefficient

Table 3

Frequency of Most Salient Content and Cognitive Characteristics on Grade 9 Mathematics Achievement Tests

	Male Context	Short-Cuts	Spatial	Female Context	Verbal	Application Unfamiliar	Application Familiar	Memorization	Symbolic	Total
1996	0	2	6	0	12	6	22	4	0	52
1997	0	0	8	0	14	5	21	2	3	53
Total	0	2	14	0	26	11	43	6	3	105 ^a

^aThree items were dropped from the 1996 and two items were dropped from the 1997 administration because the reviewers could not agree on the classification of the item.

Table 4

Differential Bundle Functioning Results for the Grade 9 Mathematics Achievement Test

	Bundle	No. of Items	Beta-Uni	Favors
1996				
	Spatial	6	0.25*	Males
	Memorization	4	-0.04*	Females
1997				
	Spatial	8	0.29*	Males
	Memorization	2	-0.03*	Females

* $p < .05$.

Note. The matching subtest used in each year was created by combining items from the remaining seven categories, with the exclusion of three items in 1996 and two items in 1997 that were not classified.

Table 5

Dimensionality Assessment Results for the Grade 9 Mathematics Achievement Test

Assessment Subtest	No. of Items	T
1996		
Spatial	6	2.67*
Memorization	4	0.68
1997		
Spatial	8	1.97*
Memorization	2	1.34

* $p < .05$.

Note. The partitioning subtest (PT) used in each year was created by combining items from the remaining seven categories, as in the differential bundle functioning analysis.

Table 6

Frequency of All Salient Content and Cognitive Characteristics on Grade 9 Mathematics Achievement Tests

	Male Context	Short-Cuts	Spatial	Female Context	Verbal	Application Unfamiliar	Application Familiar	Memorization	Symbolic	Total
1996	4	25	20	0	38	14	31	6	0	138
1997	4	15	25	0	39	8	34	4	3	132
Total	8	40	45	0	77	22	65	10	3	270

Figure Caption

Figure 1. Gender differences for items in the 1996 and 1997 tests, organized into bundles using the nine categories in Table 1.

Figure 2. Gender differences for common items in the 1996 and 1997 tests, organized into bundles using the nine categories in Table 1.

