

**Using the Multidimensionality-Based DIF Analysis Paradigm to Study
Cognitive Skills that Elicit Group Differences: A Critique**

Mark J. Gierl*

Centre for Research in Applied Measurement and Evaluation
Department of Educational Psychology
University of Alberta

Jeffrey Bisanz

Centre for Research in Child Development
Department of Psychology
University of Alberta

Yuen (Shirley) Y. Li

Centre for Research in Applied Measurement and Evaluation
Department of Educational Psychology
University of Alberta

Paper Presented at the Annual Meeting of the National Council on
Measurement in Education (NCME)

**San Diego, California, U.S.A.
April 13-15, 2004**

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

Using the Multidimensionality-Based DIF Analysis Paradigm to Study Cognitive Skills that Elicit Group Differences: A Critique

Introduction

Camilli and Shepard (1994; also see Roussos & Stout, 1996; Ramsey, 1993; Zieky, 1993) described a three-step approach used by researchers and practitioners attempting to identify biased test items:

First, statistical methods are used to find items for which there are unexpected differences in performance between two groups (e.g., men and women). Second, each potentially biased item is examined for the reasons it is relatively more difficult for a particular group of examinees. Third, an item is considered to be biased if it can be established that the source of the unexpected or "extra" difficulty for one group is not relevant to what the test measures.

(p. xiii)

In other words, this *standard approach* to DIF detection requires that each item is first tested statistically using a conditional DIF detection method (for a review of these methods, see Clauser & Mazor, 1998) and then scrutinized using some form of substantive review to identify the cause of the group difference. This approach has also been described as an exploratory DIF analysis, meaning items that produce unexpected group differences are flagged statistically and then scrutinized by reviewers who attempt to understand why the item may be more difficult for one group of examinees. Exploratory DIF analyses are often conducted when few *a priori* ideas exist about which items elicit group differences or why (Bolt & Stout, 1996; Roussos & Stout, 1996; Stout & Roussos, 1995).

Despite its popularity and frequent use, researchers and practitioners also tend to agree that the standard approach to DIF detection has not increased our understanding about why group differences occur because statistically flagged DIF items are difficult to interpret. For example, Angoff (1993) claimed, more than a decade ago, that: "It has been reported by test developers that they are often confronted by DIF results that they cannot understand; and no amount of deliberation seems to help explain why some perfectly reasonable items have large DIF values" (p. 19). Camilli and Shepard (1994) reported that, in their experience, as many as half of the items with "large" DIF in any one study might not be interpretable. Roussos and Stout (1996),

after reviewing the DIF literature, 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). The authors of the *Standards for Educational and Psychological Testing* (1999), which represent the most widely accepted standards for educational and psychological testing, summarized these concerns more succinctly when they 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)

In short, considerable progress has been made in the development and refinement of statistical methods for identifying items showing DIF, but the development and refinement of substantive methods designed to aid with the interpretation of these items have lagged far behind (Gierl, Bisanz, Bisanz, & Boughton, 2003). Consequently, many DIF studies, as they are currently conducted, do not yield information about the construct-related dimensions that produce group differences, do not produce results that can help test developers modify or refine their procedures and practices, and do not contribute to our understanding about the nature of group differences on educational and psychological tests.

However, two important developments may shed new light on the study of group differences using DIF detection methods. First, Roussos and Stout (1996) developed a *multidimensionality-based DIF analysis paradigm* to bridge the gap between substantive and statistical DIF analyses by linking both to the Shealy-Stout multidimensional model for DIF (Shealy & Stout, 1993). The paradigm requires, first, a substantive analysis where DIF hypotheses are generated and then a statistical analysis where DIF hypotheses are tested. This approach has great promise for bridging the gap between substantive and statistical DIF outcomes so that group differences can be more easily identified *and* interpreted (see, for example, Stout, Bolt, Froelich, Habing, Hartz, &

Roussos, 2003; Gierl & Khaliq, 2001; Gierl et al., 2003; Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001).

Second, research in cognitive psychology is beginning to influence practices in educational and psychological measurement. In fact, some measurement specialists believe that the psychology of test performance must be understood in order to accurately construct, score, and interpret tests because many of those instruments are based on some form of cognitive problem-solving task (e. g., Baxter & Glaser, 1998; Embretson, 1998, 1999; Frederiksen, Glaser, Lesgold, & Shafto, 1990; Frederiksen, Mislevy, & Bejar, 1993; Gierl, Leighton, & Hunka, 2000; Hattie, Jaeger, & Bond, 1999; Irvine & Kyllonen, 2002; Leighton, Gierl, & Hunka, in press; Mislevy, 1996; National Research Council, 2001; Nichols, 1994; Nichols, Chipman, & Brennan, 1995; Nichols & Sugrue, 1999; Pellegrino, Baxter, & Glaser, 1998; Royer, Cisero, & Carlo, 1993; Snow & Lohman, 1989). This important emphasis on the implications of cognitive research for educational and psychological measurement has also influenced the study of group differences using DIF methods. For instance, many researchers either emphasize the potential importance of studying the cognitive factors that produce group differences (e.g., Ackerman, 1992; Bolt & Stout, 1996; Douglas, Roussos, & Stout, 1996; Ercikan & Mendes-Barnett, 2003; Gierl, Rogers, & Klinger, 1999; Gierl & Khaliq, 2001; Roussos & Stout, 1996; Stout, 2002) or attempt to account for these cognitive factors in their study of group differences (Stout et al., 2003; Gierl et al., 2001; Gierl et al., 2003). Thus, the study of group differences using DIF methods could benefit from the merger of cognitive psychology with educational and psychological measurement because researchers are beginning to probe the cognitive bases of these differences.

The purpose of this paper is to evaluate critically this important merger focusing, specifically, on the use of the Roussos and Stout (1996) multidimensionality-based DIF analysis paradigm (herein called the DIF analysis paradigm) to study cognitive skills that elicit group differences. In the first section we summarize the Roussos and Stout (1996) DIF analysis paradigm; we provide an illustration of how the paradigm can be used to study cognitive skills that produce gender differences in mathematics; and we describe how the framework can be used to overcome some of the limitations associated with the standard approach to DIF detection. In the second section

we identify potential problems when using the DIF analysis paradigm to study cognitive skills that elicit group differences, both from a substantive and a statistical perspective.

DIF Analysis Paradigm: An Promising Approach for DIF Detection and Interpretation

Overview

Roussos and Stout (1996) proposed a multidimensionality-based DIF analysis paradigm to link substantive and statistical analyses to the Shealy-Stout multidimensional model for DIF (Shealy & Stout, 1993). The first stage is used to generate DIF hypotheses and the second stage is used to test these hypotheses. By combining substantive and statistical analyses, researchers and practitioners can begin to systematically identify and study the sources of DIF.

The DIF analysis paradigm is rooted in the Shealy and Stout (1993) multidimensional model for DIF (MMD), which serves as a theoretical basis for understanding how DIF occurs. 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 or *target* dimension. The MMD is based on two assumptions: (a) DIF items elicit at least one secondary dimension, η , in addition to the primary dimension the test is intended to measure, θ , and (b) a difference exists between the two groups of interest in their conditional distributions on the secondary dimension η , given a fixed value on the primary dimension, θ (i.e., $\eta|\theta$). Thus, items that measure the secondary dimension and produce DIF should demonstrate a disproportionate difference between the reference and focal group relative to what should be observed on items that measure only the primary dimension.

Roussos and Stout (1996) interpreted the secondary dimensions further. The secondary dimensions are *auxiliary* if they are intentionally assessed as part of the construct on the test. DIF caused by auxiliary dimensions is *benign*. Alternatively, the secondary dimensions are *nuisance* if they are unintentionally assessed as part of the construct on the test. DIF caused by nuisance dimensions is *adverse* reflecting bias. On a test of social studies achievement, for example, knowledge of social studies might be a primary dimension, critical thinking might be an auxiliary secondary dimension, and testwiseness (i.e., using strategies to select the correct answer based on knowledge of test item characteristics) might be a nuisance secondary

dimension. If a DIF item favors females and this difference can be attributed to the critical thinking auxiliary secondary dimension, when considered in isolation from the social studies primary dimension, then DIF is considered benign. Alternatively, if a DIF item favors males and this difference can be attributed to the testwiseness nuisance dimension, then DIF is considered adverse.

The DIF analysis paradigm is a two-stage procedure built on the foundation provided by the MMD. The first stage is a substantive analysis in which DIF hypotheses are generated. The DIF hypothesis specifies whether a single item or bundle of items designed to measure the primary dimension also measures a secondary dimension, thereby producing DIF. *Organizing principles* are used to identify items or bundles believed to measure secondary dimensions. Organizing principles can be based on content-related properties (e.g., items may be bundled according to curriculum content categories), on psychological characteristics (e.g., items may be bundled according to particular problem-solving strategies), or on any other features deemed relevant for structuring the data to understand the dimensions that differentiate groups (Gierl et al., 2001, pp. 33-34).

The second stage in the DIF analysis paradigm is statistically testing the DIF hypotheses. Statistical analyses are used to see whether the organizing principles reveal distinct primary and secondary dimensions. The Simultaneous Item Bias Test (SIBTEST) can be used to test DIF hypotheses and quantify the size of DIF (Stout & Roussos, 1995). To operationalize SIBTEST, items on the test are divided into the studied subtest and the matching subtest. The studied subtest contains items suspected of measuring the primary and secondary dimensions based on the substantive analysis. Alternatively, the matching subtest contains items believed to measure only the primary dimension. The matching subtest should be an accurate measure of a unidimensional matching criterion because examinees in each subgroup are placed at the same score level so their performance on items from the studied subtest can be compared. Alternatively, if the matching subtest is intended to be a multidimensional composite, then a multidimensional matching subtest is required. In this paper, however, the focus is on a

unidimensional matching subtest because the primary dimension is assumed to be unidimensional.

The amount of DIF in the studied subtest is reflected in the parameter estimate, $\widehat{\beta}_{UNI}$, defined as,

$$\widehat{\beta}_{UNI} = \int B(\theta) f_F(\theta) d\theta ,$$

where $B(\theta) = P(\theta, R) - P(\theta, F)$, the difference in the probabilities of correct response for examinees from the reference and focal groups, respectively, conditional on θ , $f_F(\theta)$ is the density function for θ in the focal group, and d is the width of the scaling interval. $\widehat{\beta}_{UNI}$ is integrated over θ to produce a weighted expected score difference between reference and focal group examinees of the same ability on an item or bundle of items.

SIBTEST is then used to assess this parameter estimate with the test statistic,

$$SIB = \frac{\widehat{\beta}_{UNI}}{\widehat{\sigma}(\widehat{\beta}_{UNI})} ,$$

where $\widehat{\sigma}(\widehat{\beta}_{UNI})$ is the estimated standard error of $\widehat{\beta}_{UNI}$. 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 provided by Shealy and Stout (1993).

Application of the DIF Analysis Paradigm to the Study of Gender Differences in Mathematics

An example helps illustrate this approach. Gierl et al. (2003) used the DIF analysis paradigm to study cognitive dimensions predicted to elicit gender differences in mathematics. The first stage in the DIF analysis paradigm requires generating DIF hypotheses. Gierl et al. used a modified taxonomy of content and cognitive characteristics proposed by Gallagher, De Lisi, Holst, McGillcuddy-De Lisi, Morely, and Cahalan (2000) as the organizing principle to account for gender differences in mathematics. The taxonomy was based on outcomes reported in the educational and psychological literature. Gallagher et al. (2000) predicted that females would

perform better than males on items with contextual characteristics likely to be more familiar to females, on items that require a high level of verbal skill, and on items that require mastery of mathematical content. Conversely, Gallagher et al. predicted that males would perform better than females on items that have contextual characteristics likely to be more familiar to males, on items that place heavy demands on spatial skills, and on items that have multiple solution paths.

Then, Gierl et al. (2003) recruited two reviewers to use the Gallagher et al. (2000) taxonomy for classifying items from a 1996 and 1997 administration of a Grade 9 mathematics achievement test administered in the Canadian province of Alberta. The reviewers were highly qualified to evaluate student performance on the mathematics achievement test by virtue of their tutoring experiences, university education, and mathematics background. 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, 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. Much like the first reviewer, the second reviewer 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. The advantages of using these reviewers include their familiarity with the content, the achievement tests, and, by virtue of their experiences in one-on-one tutoring, the strategies typically used by students to solve mathematics test items.

The reviewers received a training session prior to classifying the 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. The reviewers were instructed to identify the "most salient" characteristic for each item using the problem characteristics in the Gallagher et al. taxonomy. Once the independent classification was complete, the reviewers met to discuss their results with one another and with the authors of the study. All disagreements were discussed, debated, and resolved as a way of ensuring the

categories in the Gallagher et al. 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 were used in the study with little modification. However, the reviewers found the Gallagher et al.'s sixth category, mastery of mathematics content, difficult to apply because it was too inclusive. As a result, this category was split into four mutually exclusive categories: the first involved the application of routine mathematical solutions to new, *unfamiliar* situations; the second involved application of routine mathematical solutions to *familiar* situations; the third involved memorization; the fourth involved symbolic processes. The modified taxonomy, summarized in Table 1, was used for item classification by the reviewers in Gierl et al. (2003).

The second stage in the DIF analysis paradigm requires statistically testing the DIF hypotheses. Statistical analyses were used to evaluate whether the modified Gallagher et al. (2000) organizing principle revealed distinct primary and secondary dimensions when comparing females and males. Gierl et al. (2003) used a four-step procedure to identify dimensions that elicited gender differences. 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, $\hat{\beta}_{UNI}$, for each item. Data from the 1996 and 1997 administrations of the Grade 9 mathematics achievement test were used. Each test contained 55 dichotomously-scored items and the analyses were conducted using the responses from 6000 females and 6000 males. Second, items were grouped by the modified Gallagher et al. categories using the reviewers' classification, and the $\hat{\beta}_{UNI}$ values for these items were plotted by content and cognitive category. The category bundles for the 1996 and 1997 mathematics achievement tests are shown in Figure 1. Third, bundles were identified by visually examining the graphs and looking for interpretable patterns where a group of items *consistently* favored females or males. From the results in Figure 1, Gierl et al. identified two bundles, one for spatial skills favoring males and a second for memorization skills favoring females. These two bundles

served as the studied subtests. Items for the remaining five categories were evenly distributed, for the most part, between the two groups revealing no systematic gender differences. These items served as the matching subtest because they did not systematically favor either group. Fourth, the interpretable bundles were tested using SIBTEST.

As shown in Table 2a, males performed better than females on items requiring spatial skills. Conversely, females performed better than males on items requiring memorization skills. Gierl et al. (2003) concluded males performed better than females on items that require significant spatial processing, a finding consistent with previous research (e.g., Casey, Nuttall, Pezaris, & Benbow, 1995; Halpern, 1997), and females performed better than males on items require significant memorization skills. Support was less apparent for other sources of gender differences in the Gallagher et al. (2000) taxonomy.

Advantages of the Multidimensionality-Based DIF Analysis Paradigm

The DIF analysis paradigm, as just described and illustrated, helps overcome some of the limitations associated with the standard approach to DIF detection. Rather than using exploratory logic, the DIF analysis paradigm draws on confirmatory logic for DIF detection and interpretation. The confirmatory approach begins with a substantive analysis to generate DIF hypotheses. It is followed with a statistical analysis where the DIF hypotheses are tested. Skilled reviewers use categories in the organizing principle to structure items on the studied and matching subtests. Then, SIBTEST is used to test the items in the studied subtests. Each DIF analysis, therefore, provides a test of the proposed hypotheses. A confirmatory approach provides better Type I error control than an exploratory approach because only a comparatively small number of DIF hypotheses are tested. A confirmatory approach also has great potential to enable researchers and practitioners to systematically identify the sources of DIF so a body of *confirmed* DIF hypotheses can be created which, when accumulated over studies, may lead to a better understanding of why DIF occurs (Stout & Roussos, 1995).

Rather than focusing only on single items, the DIF analysis paradigm can be used to evaluate single items and bundles of items. DIF hypotheses are specified and tested to determine whether items designed to measure the primary dimension also measure a secondary dimension, thereby

producing group differences. Moreover, a single item may not yield an adequate measure of the secondary dimension that produces DIF (Douglas et al., 1996; Gierl et al., 2001; Nandakumar, 1993). However, a bundle of items provides a broader sample of examinee performance over a larger number of items. When these bundles tap a secondary dimension, they may *amplify* and detect group differences leading to a more powerful statistical analysis even when the same items tested separately show no statistically significant effects (Nandakumar, 1993). Results from bundle analyses may also be more interpretable than results from item analyses because the bundle represents a larger sample of the secondary dimension.

Rather than using a model-less approach, the DIF analysis paradigm is guided by a formal multidimensional model for understanding how DIF occurs. This model emphasizes that a careful study of the underlying dimensions of a test is needed to identify and interpret group differences. Essentially, the emphasis is placed on a distinction between the primary, or intended *target* dimension of a test, and the secondary, or unintended *auxiliary* and *nuisance* dimensions of a test. Items that measure the secondary dimension should demonstrate a disproportionate difference between the reference and focal group relative to what should be observed on items that measure the primary dimension. Moreover, this distinction between the primary and secondary dimensions is required for selecting the matching subtest and testing the studied subtest.

To summarize, differential item functioning studies are designed to identify and interpret construct-related dimensions that elicit group differences. Considerable progress has been made in developing statistical methods for identifying DIF items, but procedures to aid with the substantive interpretations of these items have lagged far behind. To overcome this problem, Roussos and Stout (1996) proposed a multidimensionality-based DIF analysis paradigm. With this approach, data are structured using an organizing principle to produce substantively meaningful hypotheses which, in turn, are tested statistically. The unification of substantive and statistical approaches to DIF detection, as outlined in the DIF analysis paradigm, yields more interpretable DIF results and, hence, new insights into the nature of group differences on tests.

Challenges in Using the DIF Analysis Paradigm to Study the Cognitive Bases of Group Differences

The DIF analysis paradigm has great promise for bridging the gap between substantive and statistical DIF outcomes so that group differences can be more easily identified and interpreted. However, the study of cognitive factors that elicit group differences will only lead to a better understanding of these differences if cognitive performance can be modeled using measurement procedures. In the next section we evaluate key assumptions about how cognition is modeled and tested using the DIF analysis paradigm. First, we focus on challenges related to the substantive analyses. Then, we focus on a challenge related to the statistical analysis.

Substantive Analysis

The first stage of the DIF analysis paradigm requires generating DIF hypotheses, which specify whether items measure distinct primary and secondary dimensions. To decide whether the data elicit these dimensions, organizing principles are used to identify items believed to share certain characteristics outlined in the organized principle. Typically, reviewers classify items into distinct categories specified in the organizing principle. However, two significant obstacles facing researchers and practitioners are identifying appropriate cognitive organizing principles and classifying test items accurately and appropriately.

Identifying Cognitive Organizing Principles

Test Specifications. Test specifications provide an obvious and readily available organizing principle for identifying and studying cognitive skills across groups. These specifications are used during test construction to outline the achievement domain and help the developers obtain a representative sample of items from this domain. The specifications also guide item writing and help structure the test based on the content and cognitive domain that the test is designed to measure. A thorough analysis of the content areas measured by the test and the cognitive skills required by the examinees to solve items may also help identify subsets of items that measure distinct dimensions associated with these content areas and cognitive skills (e.g., Ackerman, Gierl, & Walker, 2003; Gierl et al., 2001; Oshima, Raju, Flowers, & Slinde, 1998; Zhang & Stout, 1999).

However, this cognitive organizing principle has important weaknesses. For example, if the test specifications are a poor representation of the achievement domain, then this organizing principle will yield inadequate results. Further, it must be assumed that the items accurately measure the content areas and cognitive skills outlined in the test specifications. This assumption is, sometimes, incorrect (e.g., Ballator, 1996; Silver & Kenney, 1993). But the most serious weakness stems from the simplistic assumption about the psychology of test performance inherent to most test specifications. Typically, cognitive skills are intended to reflect the thought processes used by examinees to solve test items as outlined in *Bloom's Taxonomy of Educational Objectives* (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956). This taxonomy, which contains six different levels of thinking ranging from knowledge (i.e., recall of specific information) to evaluation (i.e., the ability to judge the value of materials and methods for a given purpose), may provide a convenient framework to classify items from a test developer's representation of cognition. However, Bloom's taxonomy does not adequately represent the scope or complexity of applied cognitive processes as described in contemporary research on cognition (e.g., Durso, Nickerson, Schvaneveldt, Dumais, Lindsay, & Chi, 1999; Leighton & Sternberg, 2004; Snow & Lohman, 1989; Sternberg, 1994) and it may not even represent how students actually solve items on tests (Gierl, 1997a, 1997b). Instead, the emphasis during test development is often on curricular features such as content coverage (Emmerich, 1989) and predictive features such as student classification (Embretson, 1985). Cognitive features are often poorly evaluated because item writers typically are not trained to identify the cognitive processes required to solve test items, and item writers rarely have an adequate model of cognition (aside from Bloom's taxonomy) to guide their inferences about student performance (e.g., Bejar, 1993; Embretson, 1999; Hattie, Jaeger, & Bond, 1999; Mislevy, 1996; Nichols, 1994; Nichols & Sugrue, 1999). Therefore, a more sophisticated representation of cognition is needed in most test specifications, the representation should reflect what is currently known about the psychology of test performance, and the representation should be validated using data obtained from students.

Research-Based Cognitive Organizing Principles. Outcomes from research in developmental and cognitive psychology can also be used to create cognitive organizing principles. As was

illustrated previously, Gierl et al. (2003) used a modified taxonomy containing two content areas and seven cognitive skills identified by Gallagher et al. (2000) to study gender differences in mathematics. Unfortunately, this type of cognitive organizing principle is rare. To our knowledge, there is no comparable taxonomy in other content areas—like science, social studies, or language arts—outlining the cognitive skills that elicit group differences on tests. Therefore, a host of cognitively-based, content-specific, organizing principles must still be developed.

When cognitive organizing principles are developed, validation studies should be conducted to ensure the cognitive characteristics can be linked to particular group differences on tests. For example, Gierl et al. (2003) concluded that males perform better than females on items that require significant spatial processing and that females perform better than males on items requiring memorization. Support was less apparent for the other five cognitive skills predicted to elicit gender differences in the modified Gallagher et al. (2000) taxonomy, as these skills either were rarely observed in the items or were consistently unrelated to gender differences in performance.

Therefore, the outcomes from Gierl et al. (2003) raise questions about the adequacy of the Gallagher et al. (2000) organizing principle for understanding the cognitive bases of gender differences in mathematical achievement. To evaluate the generalizability of the initial findings, we recently replicated the Gierl et al. study using the Foundation Skills Assessment (FSA)—Numeracy, which is a curriculum-based achievement exam administered to all Grade 10 students in the Canadian province of British Columbia. Many factors were consistent between the Gierl et al. study and the FSA—Numeracy replication study. For example, the mathematics curriculum in Alberta and British Columbia is the same because the two provinces share a common curriculum framework as part of the Western Canadian Protocol for Collaboration (WCP) in Basic Education. The WCP is an agreement between the western Canadian provinces and northern territories to develop and implement a common curriculum for all students (for more details, see the Western Canadian Protocol website at <http://www.wcp.ca/>). As a result, the test specifications used to design the tests were very similar (see Appendix A). In addition to using tests with the same specifications, both studies used the same organizing principle to classify items (see Table

1); the same two reviewers conducted the item classification; the same methods were used to identify item bundles; and the same analyses were used to test the identified bundles.

Data from the 2001 and 2002 administrations of the FSA—Numeracy exams were used in the replication study. The 2001 and 2002 exam contained 32 and 28 dichotomously-scored items, respectively. Analyses were conducted using the responses from a random sample of 2000 females and 2000 males. Bundles for the 2001 and 2002 exams are shown in Figure 2. From these results, three bundles are apparent: a shortcut, an application of routine mathematical solutions to unfamiliar situations, and a symbolic process bundle, all favoring males. Consistent with expectations based on the modified Gallagher et al. (2000) taxonomy, males outperformed females on items eliciting shortcut strategies. However, inconsistent with expectations from the modified taxonomy, males outperformed females on the application of routine mathematical solutions to unfamiliar situations items and the symbolic processing items. The FSA—Numeracy results also differ from Gierl et al. (2003): Neither the spatial bundle nor the memorization bundle favored males or females, respectively, on the FSA examinations (refer to Table 2b).

The inconsistent results across the two studies highlights two potential problems with applying research-based cognitive organizing principles to different tests, even when these principles are available. First, generalization is precarious. In the current study, the results from Gierl et al. (2003) were compared to results from the FSA—Numeracy exam. Although students differed in the two testing programs, the studies shared many important similarities including the use of the same test specifications, the same organizing principle (see Table 1), the same content reviewers, the same methods for identifying bundles, and the same analyses for testing the bundles. Yet, despite these important similarities, the results differed both in relation to one another (Gierl et al. identified spatial bundles favoring males and memorization bundles favoring females whereas the FSA—Numeracy exams produced shortcut, application of routine mathematical solutions to unfamiliar situations, and symbolic processing bundles, all favoring males) and to key predictions outlined in the modified Gallagher et al. taxonomy (2000) (only two and one of the predictions specified in the modified Gallagher et al. taxonomy were found using the Alberta and British Columbia data, respectively). The nature of these diverging outcomes is

unclear. It could reflect differences in the cognitive characteristics specific to the items on the two tests. But it could also reflect the actual cognitive bases of gender differences in mathematics, which may be minimal, different in the two testing programs, and different from the predictions outlined in the Gallagher et al. taxonomy. More generally, however, these divergent outcomes suggest that it may be difficult to create a body of confirmed DIF hypotheses, as suggested by Stout and Roussos (1995), because outcomes from research-based cognitive organizing principles are slow to accumulate over studies.

Second, the distribution of items is uneven across cognitive categories in the organizing principle. In addition to identifying the “most salient” characteristic for each item using the modified Gallagher et al. (2000) taxonomy, our reviewers were also asked to rate “all salient” characteristics associated with each item using Table 1. This summary allowed us to classify each item according to one or more of the nine categories and evaluate the cognitive complexity of the items on the tests. As shown in Table 3, the item frequency for each content or cognitive category outlined in the modified Gallagher et al. (2000) taxonomy ranged from 0% to 74% across the two test administrations evaluated in Gierl et al. (2003) and from 0% to 93% across the two test administrations for the FSA—Numeracy exams. Moreover, the item frequency for four of the nine categories was 12% or less (i.e., male content; female content; memorization; symbolic) in the Gierl et al. study, meaning items on the two mathematics achievement tests rarely or never measured two of the content areas or elicited two of the cognitive processes outlined in the modified taxonomy. In other cases, item frequency varied by category across the test administrations. For instance, the FSA—Numeracy exam containing items in the application of routine mathematical solutions to unfamiliar situations and memorization categories for 2002 but not 2001, making it difficult to evaluate the generalizability of any gender effects in these two categories. The shortcuts dimension on the FSA—Numeracy exam is another example of this problem: It contained two items favoring males producing a significant statistical outcome on the 2001 administration but only one item favoring males producing a nonsignificant statistical outcome on the 2002 administration. Thus, it is difficult to evaluate the effect of the spatial dimension on gender differences because the outcome is inconsistent across the two test

administrations due, in part, to the small number of items measuring this dimension. This uneven distribution of items in cognitive categories both within and across administrations detracts from the usefulness of the cognitive organizing principle because some cognitive skills are rarely or never observed and, therefore, gender differences cannot be evaluated for these skills.

Yet, this outcome could, perhaps, be anticipated whenever the organizing principle is not used to guide test development. That is, organizing principles developed independently of the test development process and applied to existing test items will likely produce an uneven distribution of items across the cognitive categories because the exams are not developed to test hypotheses about the cognitive bases of group differences. In fact, it is often challenging to evaluate the cognitive bases of student performance, in general, because cognitive skills may not be well represented on tests, given the impact of cognitive theory on test design is minimal (Embretson, 1998; National Research Council, 2001; Pellegrino, Baxter, & Glaser, 1999). Embretson (1994), in particular, believes that test developers have been slow to integrate cognitive theory into measurement practice because developers lack a framework for using cognitive theory to develop tests. This outcome suggests that cognitive organizing principles applied to existing tests will invariably produce a tenuous fit because the tests were not designed from a cognitive model or framework. Embretson (1998) also argued that cognitive theory is not likely to impact measurement practice until its role can be clearly established in test design. However, until this role is established more firmly, it may be difficult to evaluate how cognitive skills influence student performance using cognitive organizing principles applied *post hoc* to existing test items.

Item Classification

Item classification is another important step in the substantive analysis. Reviewers must classify each item according to the categories in the organizing principle. Thus, the cognitive categories outlined in the organizing principle must be comprehensive in order to adequately describe student performance over a relatively large number of test items. The cognitive categories must also be distinct to ensure the items can be classified in a consistent manner by the reviewers. Gierl et al. (2003), for example, used an organizing principle containing two content areas and seven cognitive skills to describe the cognitive basis of gender differences in

mathematics, based on the taxonomy presented by Gallagher et al. (2001). Gierl et al. argued that their reviewers were particularly well-suited for the classification task by virtue of their knowledge of the mathematics content, their familiarity with the achievement tests, and their insight into student problem-solving strategies resulting from their one-on-one tutoring experiences. Moreover, the reviewers received a training session prior to classifying items to ensure the categories in Table 1 were interpreted consistently. The two reviewers also classified the test items independently prior to meeting with one another to discuss their ratings. Items where no consensus was reached were dropped from the study. Despite these precautions designed to ensure that item classification was *reliable*, there were no data collected to ensure that item classification was *valid*. Instead, Gierl et al. relied on the experiences and expertise of the two reviewers, and assumed item classification was accurate. Of course, this assumption may be incorrect. Therefore, additional data could be collected from students using verbal reports to evaluate the accuracy of reviewers' item classification (cf. Baxter & Glaser, 1998; Embretson & Gorin, 2001; Gierl, 1997a, 1997b; Hamilton, Nussbaum, & Snow, 1997; Leighton, Rogers, & Maguire, 1999; Norris, 1990).

Once the items are classified, differential item and bundle functioning analyses are conducted. Dimensional homogeneity produced by grouping similar items within a category using an organizing principle containing distinct categories should yield unique subtests that can be detected statistically (Stout et al., 2003). However, to produce dimensionally homogeneous categories each item must be associated with *one category only* in the organizing principle. If items were placed in multiple categories, then dimensionally heterogeneous categories would be produced because the same item characteristics would influence different categories. Therefore, the reviewers were instructed to identify the "most salient" characteristic for each item using the problem characteristics in Table 1. The reviewers' ratings from the item classification task were then used to create Figures 1 and 2.

In addition to identifying the "most salient" characteristic for each item using the modified Gallagher et al. (2000) taxonomy, our reviewers were also asked to rate "all salient" characteristics associated with each item using Table 1 (as previously discussed). This summary

allowed us to classify each item according to one or more of the nine categories thereby evaluating the cognitive complexity of the items and the adequacy and completeness of the “most salient” classification required for the DIF analyses. For both the Alberta and British Columbia exams, the use of one mutually exclusive content or cognitive category per item oversimplified the cognitive complexity elicited by most test items. For example, Gierl et al. (2003) demonstrated that mathematics items contained multiple content and cognitive characteristics, as items elicited an average of 2.5 salient categories across two test administrations (see Table 3a). The FSA—Numeracy items had similar characteristics, as items elicited an average of 3.9 salient categories across two test administrations (see Table 3b). This finding strongly suggests item classification based on mutually exclusive cognitive categories is likely to be an oversimplification of the content and cognitive characteristics these items actually elicit.

Item classification is also based on the assumption that examinees all use the *same strategy* to solve each item. However, our reviewers revealed that their ratings were based on the strategy *most likely* to be used by students. The reviewers claimed this inference was required to classify items into one category even though they sometimes identified multiple strategies that could be used to correctly solve some items. To illustrate this point, consider the example in Appendix B. The reviewers in Gierl et al. (2003) believed this item could be solved by memorizing the concept or by drawing a diagram. If the reviewers believed that memorization is the more likely strategy (i.e., the examinee remembered that a negative z-score is always below the mean), then they would classify the item as memorization. Alternatively, if the reviewers believed 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 judgment about strategy use or students can use multiple strategies. When two strategies are considered to be mutually exclusive, the consequence is that the final item classification does not accurately represent the diversity of cognitive skills used by students. The extent of this problem is not clear

when the DIF analysis paradigm is used, and neither is its impact on the analysis and the interpretation. However, measurement specialists must recognize that strategy diversity is unavoidable when students solve items on tests. In fact, cognitive researchers who study mathematical problem solving report that multiple strategies are used commonly by students to solve problems and that a student may even apply a different strategy to the same problem because, at any one point in time, a student possesses a repertoire of problem-solving strategies (e.g., Kuhn, 1995; Kuhn, Garcia-Mila, Zohar, & Andersen, 1995; LeFevre, Sadesky, & Bisanz, 1996; Mabbott & Bisanz, 2003; Siegler, 1988; Siegler, 1998, pp. 282-298; Siegler & Shipley, 1995). Therefore, complex test-taking performance should not be simplified by describing this performance with a single strategy or process unless it can be demonstrated that an item only elicits a single strategy or process. Instead, we should attempt to identify the dominant cognitive processes or strategies that characterize complex test performance, we should evaluate the testing conditions that elicit these diverse processes and strategies, and we should study the how these processes and strategies differ between groups of examinees.

Statistical Analysis

DBF Interpretative Guidelines

Potenza and Dorans (1995) noted that, to be used effectively, a DIF detection procedure needs an interpretable effect size measure. Guidelines exist for interpreting many DIF effect size measures *at the item level* (e.g., Dorans, 1989, p. 226; Jodoin & Gierl, 2001; Shealy & Stout, 1993, p. 181; Zieky, 1993, p. 342). Unfortunately, no guidelines exist for interpreting $\hat{\beta}_{UNI}$ *at the bundle level* and, as a result, there is no agreement on how to distinguish statistical from practical significance in differential bundle functioning (DBF) research. Creating interpretative DBF guidelines is further complicated by the potential for amplification (Nandakumar, 1993). 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. Gierl et al. (2003) reported that males performed better than females on items requiring spatial skills whereas females performed better than males on items requiring

memorization skills. The statistical outcomes from their analyses revealed that spatial and memorization items differed between gender. However, the magnitude and, hence, the interpretation implied by the effect size measures differs for the spatial and memorization items (see $\widehat{\beta}_{UNI}$ column in Table 2). For instance, males had an expected score 0.25 and 0.29 points higher on the spatial items compared to females whereas females have an expected score 0.04 and 0.03 points higher on the memorization items compared to males, conditional on the matching subtest for the 1996 and 1997 administrations, respectively. Clearly, the effect size for the memorization items is small relative to the spatial items, but how small is too small for understanding the nature of group differences in DBF research? The answer to this question is unclear, in part, because statistical outcomes must be anchored to substantive interpretations. Unfortunately, these substantive interpretations are difficult to create. Cohen (1992, p. 156), for example, provides this advice for distinguishing small, medium, and large effect sizes: A small effect size is noticeably smaller than medium but not so small as to be necessarily trivial, a medium effect size is likely to be visible to the naked eye of a careful observer, and a large effect size is the same distance above medium as small is below it. These general requirements indicated why research is sorely needed to identify and evaluate supplemental effect size criteria to promote DBF interpretations that can guide the study of group differences on tests. Cohen's requirements also reveal why effect size guidelines are difficult to create.

Conclusions

Procedures to aid with the substantive interpretations of DIF items have lagged behind the development of statistical methods for identifying these items. To address this problem, Roussos and Stout (1996) developed a multidimensionality-based DIF analysis paradigm in which substantive analyses are used to develop DIF hypotheses and statistical analyses are used to test the hypotheses. This method has great promise for bridging the gap between substantive and statistical DIF outcomes so that group differences can be more easily identified *and* interpreted. The merger of cognitive psychology with educational and psychological measurement may also play an important role in bridging the gap between substantive and

statistical DIF outcomes because cognitive factors could illuminate the substantive bases of group differences on tests thereby guiding future DIF studies.

The substantive step in the DIF analysis paradigm directs the study of group differences. Organizing principles are used to identify items or bundles believed to measure secondary dimensions with specific characteristics deemed relevant for understanding dimensions that differentiate groups. Reviewers use the categories in an organizing principle to structure the dimensions. Thus, an important step in implementing the DIF analysis paradigm is identifying an appropriate organizing principle and classifying items according to the categories in the organizing principle. We identified several key problems with conducting this step using a cognitive organizing principle. For example, few cognitive organizing principles exist to account for student performance on tests. When cognitive organizing principles are implemented, they should be validated using different tests and student samples. We found, however, that validation studies can raise more questions than answers when results do not generalize across tests or student samples. Validation studies also highlight limitations with a particular cognitive organizing principle, especially when these principles are applied *post hoc* to existing tests and when the development of the organizing principle is not coordinated with the test construction process. These problems with the cognitive organizing principle may be overcome by orchestrating the test development and analysis procedures and by creating more cognitively-based, content-specific, organizing principles to promote a broader study of group differences on tests.

We also noted that reviewers classify items using the cognitive organizing principle. Item classification is assumed to be accurate by virtue of the experience and expertise of the reviewers. However, this assumption may be incorrect and, as a result, item classification could be supplemented with student verbal reports to evaluate the accuracy of the classification outcomes. Reviewers must also classify each item into one category, as the organizing principle should contain dimensionally homogeneous item sets. However, this constraint may oversimplify the cognitive complexity of test performance because items often elicit skills found in more than one cognitive category. Item classification is also based on the assumption that all examinees

use the same strategy to solve an item on the test. However, this assumption is inconsistent with contemporary psychological research indicating that students use a variety of strategies to solve mathematics items and that a student may even use different strategies to solve the same item from one time to the next. We contend that these problems with item classification pose the greatest threat to the veracity of the cognitive inferences when the DIF analysis paradigm is used. Both assumptions—item performance characterized by one mutually exclusive cognitive category and student performance characterized by one dominant strategy—oversimplify the psychology of complex test performance. Unfortunately, the implications of these assumptions on either the analyses or the interpretations are unclear. Therefore, these assumptions, and their implications, should be evaluated further when using the DIF analysis paradigm. Alternatively, if researchers want to identify or develop new methods to model group differences in cognition, then these methods should permit analyses that are *consistent with the assumptions* underlying student problem solving as outlined in contemporary psychological research.

Once the data are structured using the organizing principle, DIF hypotheses are tested statistically. Effect size measures are used to increase the interpretability of items flagged with DIF detection procedures. Unfortunately, no guidelines exist to interpret the effect size measure for bundles and, therefore, researchers and practitioners have no standards from which to differentiate statistical from practical significance. This problem with the statistical analysis can be overcome by identifying and evaluating supplemental effect size criteria. Thus, research is needed to identify and evaluate effect size guidelines for interpreting differential bundle functioning.

A Final Thought

To this point we have described several steps that must be taken if the gap between substantive and statistical DIF outcomes is to be bridged. These steps seem reasonable and necessary but, at the same time, we are not entirely comfortable with our recommendations. We are concerned that fundamental incompatibilities may exist between assumptions underlying cognitive theories of the type needed for substantive analyses and assumptions underlying DIF statistical analyses. If so, then these incompatibilities may undermine all attempts to bridge the

gap. The value of using theories and constructs from cognitive psychology to understand differences in performance on educational tests is obvious, in one sense, because these theories are directly relevant for understanding how people use knowledge to answer questions and solve problems. In another sense, however, the match may be less than ideal. In contemporary cognitive psychology, the focus is on the individual problem solver who brings an array of solution procedures to each problem and who selects an approach from these procedures as a result of a complex interaction between task and personal characteristics. This selection process is not easy to predict, and certainly the solution procedures need not be determined reliability by a single item characteristic. Moreover, accuracy often is not viewed as a particularly good indicator of cognitive processing because the same cognitive processes can yield either correct or incorrect answers, and different cognitive processes can yield the same answers. In contrast, DIF-based measurement methods are based on assumptions about the agency of items rather than individuals, the centrality of an accurate solution product rather than solution process, and the assumption of person-item consistency rather than inconsistency. Thus, cognitive theories, as they exist currently, may not provide a good substantive basis for generating DIF hypotheses to be tested with DIF statistical methods. Conversely, DIF statistical methods, in their current form, may not be adequate for testing DIF hypotheses generated from cognitive analyses. We are not sure yet whether these apparent incompatibilities are intractable or whether they can be handled with adjustments to the cognitive and/or statistical analyses. However, these nagging concerns must be resolved to determine whether the cognitively-based substantive and statistical DIF analyses we described can be usefully integrated or whether they must, by necessity, remain separate and distinct.

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.
- Ackerman, T. A., Gierl, M. J., & Walker C. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*, 37-53.
- 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.
- Ballator, N. (1996). *The NAEP guide: A description of the content and methods of the 1994 and 1996 assessments*. Washington, DC: National Council on Educational Statistics.
- Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice, 17*, 37-45.
- Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323-358). Hillsdale, NJ: Erlbaum.
- Bloom, B. S. (Ed.), Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *A taxonomy of educational objectives: Handbook I Cognitive Domain*. New York: Logmans, Green.
- Bolt, D., & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika, 23*, 67-95.
- 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.
- Cohen, J. (1992). A power primer. *Psychological Bulletin, 11*, 155-159

- 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.
- Douglas, J., Roussos, L., & Stout, W., (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. *Journal of Educational Measurement*, 33, 465-484.
- Durso, F. T., Nickerson, R. S., Schvaneveldt, R. W., Dumais, S. T., Lindsay, D. S., & Chi, M. T. S. (Eds.), *Handbook of Applied Cognition*. New York: Wiley.
- Embretson, S. E. (1985). Introduction for the problem of test design. In S. E. Embretson (Ed.). *Test design: Developments in psychology and psychometrics* (pp. 3-17). New York: Academic Press.
- Embretson, S. E. (1994). Application of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107-135). New York: Plenum Press.
- Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3, 380-396.
- Embretson, S. E. (1999). Cognitive psychology applied to testing. In F. T. Durso, R. S. Nickerson, R. W., Schvaneveldt, S. T. Dumais, D. S. Lindsay, & M. T. H. Chi (Eds.), *Handbook of Applied Cognition* (pp. 629-660). New York: Wiley.
- Embretson, S. E., & Goring, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, 38, 343-368.
- Emmerich, W. (1989). *Appraising the cognitive features of subject tests* (Research Rep. No. RR-89-53). Princeton, NJ: Educational Testing Service.
- Ercikan, K., & Mendes-Barnett, S. (2003). *Disentangling sources of DIF: Interpretation of results from SIBTEST*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Frederiksen, N., Glaser, R. L., Lesgold, A. M., & Shafto, M. G. (1990). *Diagnostic monitoring of skills and knowledge acquisition*. Hillsdale, NJ: Erlbaum.
- Frederiksen, N., Mislevy, R. J., & Bejar, I. I. (1993). *Test theory for a new generation of tests*. Hillsdale, NJ: Erlbaum.

- 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.
- Gierl, M. J. (1997a). Comparing the cognitive representations of test developers and students on a mathematics achievement test using Bloom's taxonomy. *Journal of Educational Research, 91*, 26-32.
- Gierl, M. J. (1997b). An investigation of the cognitive foundation underlying the rule-space model (Doctoral dissertation, University of Illinois at Urbana-Champaign, 1996). *Dissertation Abstracts International, 57* (08), 5351-B. (University Microfilms No. AAC 97-02524)
- Gierl, M. J., Bisanz, J., Bisanz, G., & Boughton, K. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the DIF analysis paradigm. *Journal of Educational Measurement, 40*, 281-306.
- 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.
- Gierl, M. J., Leighton, J. P., & Hunka, S. (2000). Exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educational Measurement: Issues and Practice, 19*, 34-44.
- Halpern, D. F. (1997). Sex differences in intelligence: Implications for education. *American Psychologist, 52*, 1091-1102.
- Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10*, 181-200.
- Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent methodological questions in educational testing. *Review of Research in Education, 24*, 393-446.

- Irvine, S. H., & Kyllonen, P. C. (Eds). (2002). *Item generation for test development*. Mahwah, NJ: Erlbaum.
- 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.
- Kuhn, D. (1995). Microgenetic study of change: What has it told us? *Psychological Science, 6*, 133-139.
- Kuhn, D., Garcia-Mila, M., Zohar, A., & Andersen, C. (1995). Strategies for knowledge acquisition. *Monographs of the Society for Research in Child Development, 60*, Serial No. 245.
- LeFevre, J., Sadesky, G. S., & Bisanz, J. (1996). Selection of procedures in mental addition: Reassessing the problemsize effect in adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 216-230.
- Leighton, J. P., Gierl, M. J., & Hunka, S. (in press). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*.
- Leighton, J. P., & Sternberg, R. J. (Eds.). (2004). *The nature of reasoning*. Cambridge, UK: Cambridge University Press.
- Leighton, J. P., Rogers, W. T., & Maguire, T. O. (1999). Assessment of student problem solving on ill-defined tasks. *Alberta Journal of Educational Research, 45*, 409-427.
- Mabbott, D. J., & Bisanz, J. (2003). Developmental change and individual differences in childrens multiplication. *Child Development, 74*, 1091-1107.
- Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*, 379-416.
- Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 16*, 159-176.
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. J. Pellegrino, N.

- Chudowsky, and R. Glaser (Eds.). Board on Testing and Assessment, Center for Education. Washington, DC: National Academy Press.
- Nichols, P. (1994). A framework of developing cognitively diagnostic assessments. *Review of Educational Research, 64*, 575-603.
- Nichols, P. D., Chipman, S. F., & Brennan, R. L. (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum.
- Nichols, P., & Sugrue, B. (1999). The lack of fidelity between cognitively complex constructs and conventional test development practice. *Educational Measurement: Issues and Practice, 18*, 18-29.
- Norris, S. P. (1990). Effect of eliciting verbal reports of thinking on critical thinking test performance. *Journal of Educational Measurement, 27*, 41-58.
- Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. A. (1998). Differential bundle functioning using the DFIT framework: Procedures for identifying possible sources of differential functioning. *Applied Measurement in Education, 11*, 353-369.
- Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practices. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education* (pp. 307-353). Washington, DC: American Educational Research Association.
- Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 367-388). Hillsdale, NJ: Erlbaum.
- Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.
- Royer, J.M., Cisero, C.A., & Carlo, M. S. (1993). Techniques and procedures for assessing cognitive skills. *Review of Educational Research, 63*, 201-243.
- Siegler, R. S. (1988). Individual differences in strategy choice: Good students, not-so-good students, and perfectionists. *Child Development, 59*, 833-851.
- Siegler, R. S. (1998). *Children's thinking* (3rd edition). Upper Saddle River, NJ: Prentice-Hall.
- Siegler, R. S., & Shipley, E. (1995). Variation, selection, and cognitive change. In G. Halford &

- T. Simon (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp. 31-76). Hillsdale, NJ: Erlbaum.
- Silver, E. A., & Kenney, P. A. (1993). An examination of relationships between the 1990 NAEP mathematics items for grade 8 and selected themes from the NCTM standards. *Journal for Research in Mathematics Education*, *24*, 159-166.
- 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.
- Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331). New York: American Council on Educational, Macmillian.
- Standards for Educational and Psychological Testing*. (1999). Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.
- Sternberg, R. J. (Ed.). (1994). *Encyclopedia of human intelligence*. New York: Macmillian.
- Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika*, *67*, 485-518.
- Stout, W., Bolt, D., Froelich, A. G., Habing, B., Hartz, S., & Roussos, L. (2003). Development of a SIBTEST bundle methodology for improving test equity with applications for GRE test development. (ETS Research Report 03-06). Princeton, NJ: Educational Testing Service.
- Stout, W., & Roussos, L. (1995). *SIBTEST manual*. University of Illinois: Department of Statistics, Statistical Laboratory for Educational and Psychological Measurement.
- Zhang, J., & Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 231-249.
- 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 A

The test specifications for the Alberta Mathematics Achievement Tests and the British Columbia FSA—Numeracy Exams.

Alberta—Content Area	Percentage of Items Across Two Administrations
<p>Number</p> <ul style="list-style-type: none"> • Explain and illustrate the structure and the interrelationship of the sets of numbers within the rational number system • Develop a number sense of powers with integral exponents and rational bases • Use a scientific calculator or a computer to solve problems involving rational numbers • Explain how exponents can be used to bring meaning to large and small numbers, and use calculators or computers to perform calculations involving these numbers 	25-29%
<p>Patterns and Relations</p> <ul style="list-style-type: none"> • Generalize, design, and justify mathematical procedures by using appropriate patterns, models, and technology • Solve and verify linear equations and inequalities in one variable • Generalize arithmetic operations from the set of rational numbers to the set of polynomials 	31-38%
<p>Shape and Space</p> <ul style="list-style-type: none"> • Use trigonometric ratios to solve problems involving a right triangle • Describe the effects of dimension changes in related 2-D shapes and 3-D objects in solving problems involving area, perimeter, surface area, and volume • Specify conditions under which triangles may be similar or congruent, and use these conditions to solve problems • Use spatial problem solving in building, describing, and analyzing geometric shapes • Apply coordinate geometry and pattern recognition to predict the effects of translations, rotations, reflections, and dilatations on 1-D lines and 2-D shapes 	22-28%

Appendix A (con't)

<p>Statistics and Probability</p> <ul style="list-style-type: none"> • Collect and analyze experimental results expressed in two variables; use technology, as required • Explain the use of probability and statistics in the solution of complex problems 	13-15%
--	--------

British Columbia—Content Area	Percentage of Items Across Two Administrations
<p>Number</p> <ul style="list-style-type: none"> • Students solve problems involving numbers, including rational and irrational numbers. • They perform basic operations on the real number system and apply these skills in various practical, real-life, or technical contexts. 	25-35%
<p>Patterns and Relations</p> <ul style="list-style-type: none"> • Students use patterns to solve problems. • They simplify and manipulate algebraic expressions and make connections between algebraic and graphical representations. 	20-30%
<p>Shape and Space</p> <ul style="list-style-type: none"> • Students use trigonometry to analyze real-life situations. • They use geometric relationships among shapes to solve problems. 	25-35%
<p>Statistics and Probability</p> <ul style="list-style-type: none"> • Students interpret, draw inferences, and communicate statistical information. • They use probability terminology and determine permutations and combinations in possible events. 	10-20%

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.

Table 1

The Modified Gallagher et al. (2000) Taxonomy Outlining the Content and Cognitive Skills Expected to Elicit Gender Differences in Mathematics, as used in Gierl et al. (2003)

A. Knowledge and Skills Favoring Males

- 1) Item Content 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 Content Favoring Females
Solving the problem requires material more likely to be familiar to females (e.g., items requiring knowledge about traditionally female 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 Situations
 - 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 Solutions to Familiar Situations
 - 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 2a

Differential Bundle Functioning Results for the Alberta Mathematics Achievement Tests Reported in Gierl et al. (2003)

	Bundle	No. of Items	$\hat{\beta}_{UNI}$	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 five categories, with the exclusion of three items in 1996 and two items in 1997 that were not classified by the reviewers. A positive $\hat{\beta}_{UNI}$ favors males.

Table 2b

Differential Bundle Functioning Results for the British Columbia FSA—Numeracy Exams

	Bundle	No. of Items	$\hat{\beta}_{UNI}$	Favors
2001				
	Shortcuts	2	0.07*	Males
	Application Unfamiliar	0	---	---
	Symbolic	2	0.05*	Males
2002				
	Shortcuts	1	0.02	---
	Application Unfamiliar	3	0.11*	Males
	Symbolic	2	0.11*	Males

* $p < .05$.

Note. The matching subtest used in each year was created by combining items from the remaining categories. A positive $\hat{\beta}_{UNI}$ favors males.

Table 3a

Frequency for All Content and Cognitive Characteristics Identified on the Alberta Mathematics Achievement Test Items, as Reported in Gierl et al. (2003)

	Male Content	Short-Cuts	Spatial	Female Content	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

Note. Frequency counts were based on 52 items from the 1996 administration and 53 items from the 1997 administration. The frequency counts should be interpreted, as follows: 4—'Male Content' means that male content was noted by the reviewers on 4 different test items. Similarly, 38—'Verbal' means that verbal skills, according to the reviewers, were required to solve 38 different test items.

Table 3b

Frequency for All Content and Cognitive Characteristics on the British Columbia FSA—Numeracy Exam Items

	Male Content	Short-Cuts	Spatial	Female Content	Verbal	Application Unfamiliar	Application Familiar	Memorization	Symbolic	Total
2001	1	12	10	0	24	4	27	18	25	121
2002	1	11	9	0	20	9	17	12	26	105
Total	2	23	19	0	44	13	44	30	51	226

Note. Frequency counts were based on 32 items from the 2001 administration and 28 items from the 2002 administration.

Figure Captions

Figure 1. Gender differences for all items in the 1996 and 1997 Alberta mathematics achievement tests, using the modified Gallagher et al. (2000) taxonomy (see Table 1), as reported in Gierl et al. (2003).

Figure 2. Gender differences for all items in the 2001 and 2002 British Columbia FSA—Numeracy exams, using the modified Gallagher et al. (2000) taxonomy (see Table 1).



