

**Implications of the Multidimensionality-Based DIF Analysis  
Framework for Selecting a Matching and Studied Subtest**

**Mark J. Gierl**

Centre for Research in Applied Measurement and Evaluation  
University of Alberta

**Daniel M. Bolt**

Department of Educational Psychology  
University of Wisconsin, Madison

DRAFT: March 31, 2003

Paper Presented at the Annual Meeting of the National Council on  
Measurement in Education (NCME) at the Symposium entitled,  
“Applications and Practical Considerations of Multidimensional Item  
Response Theory”

**Chicago, Illinois, U.S.A.**

**April 22-24, 2003**

## Abstract

In this paper we describe and illustrate the Roussos-Stout (1996) multidimensionality-based DIF analysis framework, with emphasis on its implication for the selection of a matching and studied subtest for DIF analyses. Standard DIF practice encourages an exploratory search for matching subtest items based on purely statistical criteria, such as a failure to display DIF. By contrast, the multidimensional DIF framework emphasizes a substantively-informed selection of items for both the matching and studied subtest based on the dimensions suspected of underlying the test data. Using two examples, we demonstrate that these two approaches lead to different interpretations about the occurrence of DIF in a test. It is argued that selecting a *valid* matching and studied subtest, as implied by the multidimensional framework, can lead to a more informed understanding of why DIF occurs.

### **Implications of the Multidimensionality-Based DIF Analysis Framework for Selecting a Matching and Studied Subtest**

*Bias* occurs when tests yield scores or promote score interpretations that result in different meanings for members of different groups. Bias is often attributed to construct-irrelevant dimensions that differentially affect the test scores for different groups of examinees (*Standards for Educational and Psychological Testing*, 1999). Group differences can also be attributed to item *impact*. Impact occurs when construct-relevant dimensions differentially affect the test scores for different groups of examinees. In this case, the item is a relevant measure of the target construct and the difference between the groups reflects a true difference on that construct. Differential item functioning (DIF) studies are designed to identify and interpret these construct-related dimensions using a combination of statistical and substantive analyses. The *statistical analysis* involves administering the test, matching members of the reference and focal group on a measure of ability derived from that test, and using statistical procedures to identify group differences on test items. An item exhibits DIF when examinees from the reference and focal groups differ in the probability of answering that item correctly, after controlling for ability. The *substantive analysis* builds on the statistical analysis because DIF items are often scrutinized by expert reviewers (e.g., test developers or content specialists) who attempt to identify the construct-related dimensions that produce group differences. A DIF item is considered biased when reviewers identify some dimension, deemed to be irrelevant to the construct measured by the test, that places one group of examinees at a disadvantage. Conversely, a DIF item displays impact when the dimension that differentiates the groups is judged to be relevant to the construct measured by the test.

Considerable progress has been made in the development and refinement of statistical methods for identifying items showing DIF (see reviews by Clauser & Mazor, 1998; Millsap, & Everson, 1993) but the development and refinement of substantive methods designed to aid with the interpretation of these items have lagged far behind (e.g., Bond, 1993; Camilli & Shepard, 1994; Englehard, Hansche, & Rutledge, 1990; Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001; Gierl, Rogers, & Klinger, 1999; O'Neill & McPeck, 1993; Plake, 1980; Roussos & Stout, 1996; *Standards for Educational and Psychological Testing*, 1999; Sudweeks & Tolman, 1993). The traditional substantive approach--subjecting items flagged with DIF analyses to the scrutiny of

reviewers--has not been successful because the interpretations tend to be inconsistent with the DIF statistics or unreliable among reviewers. This impasse represents a fundamental problem in the study of group differences using DIF methods.

Roussos and Stout (1996) proposed a *multidimensionality-based DIF analysis framework* to bridge the gap between statistical and substantive analyses by linking both to the Shealy-Stout multidimensional model for DIF (Shealy & Stout, 1993). The first stage is a substantive analysis where DIF hypotheses are generated. The second stage is a statistical analysis where the DIF hypotheses are tested. By combining statistical and substantive analyses in a multidimensional framework, researchers and practitioners can begin to systematically identify and study the factors that produce group differences using DIF methods.

The purpose of this paper is threefold: In the first section we describe and critique the traditional approach for DIF detection and interpretation. In the second section we describe the Roussos and Stout (1996) DIF analysis framework, provide an illustration of how the framework can be used in practice, and demonstrate how the framework can be used to overcome some of the limitations associated with the traditional approach. In the third section we highlight some implications of the multidimensional DIF framework for researchers and practitioners who want to identify *and* interpret DIF items.

#### Traditional Approach to DIF Detection and Interpretation: Overview and Critique

##### *Overview*

Camilli and Shepard (1994; also see Roussos & Stout, 1996; Ramsey, 1993; Zieky, 1993) described a three-step approach (i.e., the *traditional* approach) that is widely 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, the traditional approach to DIF detection and interpretation 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 that items producing 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 the researcher or practitioner has few *a priori* ideas 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 tend to agree that the traditional approach has not increased our understand about why group differences occurs because statistically flagged DIF items are difficult to interpret. For example, 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. Angoff (1993) noted: “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). Roussos and Stout (1996) reviewed the DIF literature and claimed, “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 1999 *Standards for Educational and Psychological Testing* concluded:

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, many items identified using statistical methods cannot be interpreted using traditional substantive methods. 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 than can help test developers modify or refine their procedures and practices, do not contribute to our understanding about the nature of group differences and, ultimately, do not shed light on key issues related to fairness in testing.

*Critique of the Traditional Approach to DIF Detection and Interpretation*

At least three factors help account for the problems associated with linking statistical outcomes to substantive interpretations using the traditional approach. First, the traditional approach draws on the exploratory logic for conducting a DIF analysis where researchers and practitioners attempt to interpret statistically flagged items. With this approach,  $n$  items are tested separately for DIF with the remaining  $n - 1$  items (or some combination of the remaining items) serving as the matching subtest. Unfortunately, this approach can lead to inflated Type I errors because a large number of non-specified DIF hypotheses are tested in one analysis. Therefore, the likelihood that the researcher or practitioner will encounter (and may attempt to interpret) a non-DIF item, thought to be a DIF item, increases as the number of test items increase. Moreover, exploratory DIF analyses are not guided by hypotheses or *a priori* judgements about why DIF occurs and there is no framework to facilitate interpretation. Rather, the items are tested separately and those items that produce unexpected group differences are flagged, and then scrutinized by reviewers who attempt to understand why the item may be more difficult for one group of examinees. But many researchers and practitioners have concluded that items identified using exploratory analyses are difficult to interpret.

Second, researchers and practitioners who use the traditional approach to DIF detection and interpretation rely on single-item analyses, meaning that the unit of analysis is one test item. Unfortunately, the conclusions that can be made about group differences are often tenuous using single-item DIF analyses because the unit of analysis is small. Therefore, single DIF items may be difficult to identify because of low statistical power. Single DIF items may also be difficult to interpret because an item only represents a limited, and relatively unreliable, sample of examinee performance on a particular construct-related dimension (Douglas, Roussos, & Stout, 1996; Gierl et al., 2001; Nandakumar, 1993).

Third, the traditional approach to DIF detection and interpretation is guided by an informal multidimensional model, where DIF is attributed to multidimensionality but where the model itself is unspecified. Roussos and Stout (1996, p. 357) claim the informal multidimensional model is based on two assumptions: (a) DIF items measure a secondary dimension in addition to the primary dimension, and (b) the secondary dimension, when considered in isolation, favors one of the two groups being compared. One example presented by Doran and Kulick (1983) helps

illustrate how the informal multidimensional model is be used to interpret DIF. Students were presented with this analogical reasoning item on the Scholastic Assessment Test (SAT):

**DECOY:DUCK:: (A) net:butterfly (B) web:spider (C) lure:fish (D) lasso:rope (E) detour:shortcut.**

Using the standardization DIF statistical method, Dorans and Kulik found that this item was more difficult for females than males when overall ability was controlled. They attributed this outcome to gender-related differences in background knowledge. They also argued that this background knowledge was extraneous to analogical reasoning, thereby producing a potentially biased item: “Inspection of the content of this particular analogy item revealed potential content bias against female candidates, as it required some knowledge of hunting and fishing, two traditionally male-oriented recreational activities” (p. 20). In other words, Dorans and Kulick argued that the DIF item measured two dimensions, analogical reasoning, the primary dimension, and sports terminology, the secondary dimension (assumption a). They also argued that gender-related differences in background knowledge about sports terminology might favor males on the secondary dimension because males, traditionally, have had more experience with these activities (assumption b). Unfortunately, the informal model lacks specificity and, as the Dorans and Kulick example illustrates, it is often used to generate explanations that are *ad hoc* and item specific making it difficult to identify pervasive dimensions that produce group differences.

The informal multidimensional model can also affect DIF statistical detection when it promotes a narrow approach to test score *purification*. When a exploratory single-item DIF analysis is conducted, all items are evaluated by systematically studying each item and then placing the item back into the matching subtest for the next item analysis. Conditioning on the matching subtest is a critical step because it ensures that examinees are matched on a common unidimensional criterion before they are compared. In the majority of DIF applications, the matching subtest is internal to the test, meaning that the test items are used to create the matching subtest (Camilli & Shepard, 1994; Clauser & Mazor, 1998; Holland & Wainer, 1993). However, using an internal matching subtest can create problems because items in the matching subtest might display DIF, implying these items could measure a secondary dimension, thus compromising or confounding the matching dimension. Hence, steps must be taken to ensure the matching subtest provides an

appropriate measure of the primary dimension when an internal criterion is used for DIF analyses, and excluding items that measure secondary dimensions.

Some researchers suggest an iterative purification procedure in which DIF items are flagged, these flagged items are removed from the matching subtest using statistical criteria, and then the data are re-analyzed using the purified (i.e., DIF-free) matching subtest to flag additional DIF items (e.g., Allalouf, Hambleton, & Sireci, 1999; Camilli & Shepard, 1994; Dorans & Holland, 1993, pp. 60-61; Holland & Thayer, 1988; Lord, 1980). That is, iterative purification is an attempt to refine the matching subtest by automatically removing items flagged for DIF using purely statistical criteria because it is assumed, under the informal multidimensional model, that these items also measure a secondary dimension to the disadvantage of one group. Little or no attempt is made to interpret DIF before items are removed from the matching subtest when iterative purification is used. As a result, some of the limitations associated with the traditional approach to DIF detection may affect iterative purification. For example, iterative purification relies on exploratory DIF analyses. Yet these analyses are prone to inflated Type I errors when a large number of unspecified DIF hypotheses are tested in a single analysis causing the needless removal of non-DIF items from the matching subtest. Iterative purification focuses on single DIF items. Yet these items may be difficult to identify because of low statistical power causing, perhaps, the needless inclusion of DIF items in the matching subtest. Iterative purification is conducted with the assumption that DIF items measure the secondary dimension to the disadvantage of one group (i.e., the secondary dimension produces biased test items). Yet there is no substantive analysis undertaken during the purification process to evaluate this assumption. This assumption may be inaccurate because DIF can be attributed to bias or impact producing three possible outcomes that could affect the composition of the matching subtest. If the item is biased due to a secondary dimension, then the item should be removed from the matching subtest. If the item displays impact due to a secondary dimension that is not required for matching, then the item should be removed from the matching subtest but treated separately from the biased item (i.e., biased and impact items presumably measure different secondary dimensions; see Roussos & Stout, 1996). If the item displays impact due to a secondary dimension that is required for matching, meaning the matching subtest is judged to be multidimensional composite, then DIF analyses with multiple matching scores should be

conducted (e.g., Mazor, Kanjee, & Clauser, 1995; Stout, Li, Nandakumar, & Bolt, 1997). Iterative purification procedures do not distinguish among these three possible outcomes. In short, statistical outcomes cannot guide purification efforts; adequate statistical procedures are required to identify the DIF items *and* useful substantive analyses are required to interpret the dimensions producing differences before the matching subtest can be created.

In 1993 William Angoff claimed, "Out of the methodological efforts of the last 25 years has emerged an arsenal of methods for the analysis and identification of items with DIF characteristics" (p. 21). Despite the deployment of this statistical arsenal, many researchers and practitioners agree that interpreting DIF items is difficult. The traditional approach to DIF detection and interpretation, where statistically identified items are subjected to the scrutiny of reviewers, has generally not been successful at identifying the underlying causes of DIF, improving test development practices, increasing our understanding of group differences, generating a body of hypotheses that can guide future DIF studies, or providing new insights into the nature of fairness in testing. The failure to link statistical outcomes and substantive interpretations with the traditional approach may be attributed, in part, to the exploratory analysis of single items where DIF is linked to multidimensionality without specifying a model to operationalize this important link.

Fortunately, researchers are developing new methods and theoretical models to overcome the limitations with the traditional approach to DIF detection and interpretation. For instance, methods where substantive analyses are conducted prior to statistical analyses have been developed and applied to the study of group differences (e.g., Abbott, 2003; Bolt, Froelich, Habing, Hartz, Roussos, & Stout, in press; Gierl et al., 2001; Gierl & Khaliq, 2001; Gierl, Bisanz, Bisanz, & Boughton, in press; Vandenberghe & Gierl, 2001) because interpreting the substantive bases of statistically-flagged DIF items has not been successful. Methods have been developed for assessing and interpreting differential bundle functioning (DBF; Douglas, Roussos, & Stout, 1996; Gierl et al., 2001; Oshima, Raju, Flowers, & Slinde, 1998; Stout & Roussos, 1995) because sources of DIF may be more apparent across bundles of items rather than single items. Theoretical models have been developed to specify how multiple dimensions can cause DIF (e.g., Ackerman, 1992; Hunter, 1975; Kok, 1988; Shealy & Stout, 1993) because DIF is attributed to multidimensionality. These advances have led to the development of a new approach to DIF

detection and interpretation that has the potential to integrate substantive and statistical analyses in a multidimensional framework so researchers and practitioners can begin to systematically identify and study the sources of DIF. These advances are described and illustrated in the next section.

#### DIF Analysis Framework: An Alternative Approach for DIF Detection and Interpretation

##### *Overview*

Roussos and Stout (1996) proposed a multidimensionality-based DIF analysis framework 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 framework 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 are assumed to elicit at least one secondary dimension in addition to the primary dimension and (b) a difference is assumed to exist between the two groups of interest in their conditional distributions on the secondary dimension, given a fixed value on 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* (reflecting impact). 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 mathematics achievement, for example, knowledge of mathematics 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 mathematics 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 Roussos-Stout DIF analysis (1996) framework 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 with specific characteristics. 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 that appear to elicit particular problem-solving strategies may be bundled), or on any other features deemed relevant for understanding dimensions that differentiate groups (Gierl et al., 2001, pp. 33-34).

The second stage in the Roussos-Stout DIF analysis framework 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 standardized test are divided into the studied (or suspect) subtest and the matching (or valid) 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 subgroups are placed at the same score level so their performance on items from the studied subtest can be compared.

Like many other DIF procedures, SIBTEST uses differences in the expected scores conditional on ability across groups to test for DIF. The method can be applied using either dichotomously-scored items (Shealy & Stout, 1993) or polytomously-scored items (Chang, Mazzeo, & Roussos, 1995). Since the approach under both item scoring conditions is basically the same, we describe the more general case as it applies to polytomous items.

As a first step, SIBTEST estimates  $ES_R(\mathbf{q})$  and  $ES_F(\mathbf{q})$ , the expected score for a studied item conditional on a target ability level  $\mathbf{q}$  for a reference and focal group, respectively. However, in place of ability, SIBTEST uses total scores for a matching subtest of items believed to contain no DIF. Then the expected item scores are estimated as

$$ES_R(t) = \sum_{k=1}^m kP_{Rk}(t) \quad \text{and} \quad ES_F(t) = \sum_{k=1}^m kP_{Fk}(t),$$

where  $P_{gk}(t)$  denotes the empirical proportion of examinees in group  $g$  that obtain score  $k$  on the studied item and have matching subtest score  $t$ .  $ES_R(t)$  and  $ES_F(t)$  contain bias due to measurement error in the matching subtest. The defining feature of SIBTEST is its use of a regression correction procedure that determines adjusted expected item scores  $ES_R^*(t)$  and  $ES_F^*(t)$ . These adjusted scores more accurately reflect examinees of equal ability levels across groups and, thus, are more meaningful for comparing group differences on the studied item. SIBTEST uses a weighted average difference of these adjusted scores (weighted by the proportion of examinees obtaining matching subtest score  $t$ ) to estimate a DIF index denoted as  $\mathbf{b}_{UNI}$ , where

$$\widehat{\mathbf{b}}_{UNI} = \sum_{t=1}^T ([ES_R^*(t) - ES_F^*(t)] \frac{N_R(t) - N_F(t)}{N}).$$

In this formula,  $T$  is the maximum score on the matching subtest,  $N_R(t)$  and  $N_F(t)$  are the numbers of examinees obtaining matching subtest score  $t$  from the reference and focal groups, respectively, and  $N$  is the total number of examinees. For large samples,  $\widehat{\mathbf{b}}_{UNI}$  is approximately normal assuming a null hypothesis of no DIF, and the standard error of  $\widehat{\mathbf{b}}_{UNI}$  is given as

$$\widehat{\mathbf{s}}_{\widehat{\mathbf{b}}_{UNI}} = \left[ \sum_{t=0}^T \left( \frac{N_R(t) - N_F(t)}{N} \right)^2 \left( \frac{\widehat{\mathbf{s}}_{Rt}^2}{N_{Rt}} + \frac{\widehat{\mathbf{s}}_{Ft}^2}{N_{Ft}} \right) \right]^{1/2}.$$

The test statistic  $SIB = \frac{\widehat{\mathbf{b}}_{UNI}}{\widehat{\mathbf{s}}_{\widehat{\mathbf{b}}_{UNI}}}$  is evaluated against a standard normal distribution, and a null

hypothesis of no DIF is rejected whenever  $|SIB| > z_{1-\frac{\alpha}{2}}$ . For additional details on this approach,

see Chang et al. (1996), or Shealy and Stout (1992) for the application to dichotomously-scored items.

*Applying the DIF Analysis Framework to the Study of Gender Differences in Mathematics*

An example helps illustrate this approach. Gierl et al. (in press) used the DIF analysis framework to study cognitive dimensions predicted to cause gender differences in mathematics. The first stage in the DIF analysis framework requires generating DIF hypotheses. Gierl et al. used a modified taxonomy of content and cognitive characteristics proposed by Gallagher, De Lisi, Holst, McGillicuddy-De Lisi, Morely, and Cahalan (2000) as the organizing principle to account for gender differences in mathematics. 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, males would perform better than females on items that have contextual characteristics likely to be more familiar males, on items that are likely to place heavy demands on spatial skills, and on items that have multiple solution paths.

Gierl et al. recruited two reviewers to use the Gallagher et al. taxonomy for classifying items from a 1996 and 1997 administration of a Grade 9 mathematics achievement test. The reviewers were highly qualified to evaluate student performance on Grade 9 mathematics achievement test by virtue of their tutoring experiences, university education, and mathematics background. Each reviewer worked independently. 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.

The second stage in the DIF analysis framework involved statistically testing the DIF hypotheses. Statistical analyses were used to evaluate whether the Gallagher et al. (2000) organizing principle revealed distinct primary and secondary dimensions when comparing females and males. Gierl et al. (in press) 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,  $\hat{b}_{UNI}$ , for each item. Data from the 1996 and 1997

administrations of a 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 selected randomly from the 1996 and 1997 databases. Second, items were grouped by the Gallagher et al. cognitive categories using the reviewers' classification, and the  $\widehat{b}_{UNI}$  values for these items were plotted by cognitive category. The category bundles for the 1996 and 1997 mathematics achievement tests are shown in Figure 1<sup>1</sup>. Negative  $\widehat{b}_{UNI}$  values favor females and positive  $\widehat{b}_{UNI}$  values favor males. 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 bundle for spatial skills favoring males and a second bundle 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. As a result, Gierl et al. used these items as the matching subtest because the items did not systematically favor either group and, therefore, were presumed to be dimensionally homogeneous. Fourth, the interpretable bundles were tested. Gierl et al. used two methods to test the bundles. SIBTEST was used to test the DBF hypotheses. As shown in Table 1, males in Grade 9 performed better than females on items requiring spatial skills whereas females performed better than males on items requiring memorization skills. DIMTEST was also used to test the dimensional structure of the bundles because DIF is attributed to multidimensionality. DIMTEST, with a refined bias correction method (Froelich, 2000; Froelich & Habing, 2001), was used to determine whether the studied subtests were dimensionally distinct from the matching subtest. Table 2 contains the outcomes from the DIMTEST analysis. The spatial subtest was dimensionally distinct from the matching subtest for both 1996 and 1997 but the memorization subtest was not dimensionality distinct from the matching subtest in either year. Thus DIMTEST only confirmed one of the dimensions implied by the SIBTEST analyses. Gierl et al. concluded that males performed better than females on items that require significant spatial

---

<sup>1</sup> The knowledge and skills expected to favor males included item context favoring males, short-cuts/multiple solution paths, and spatial skills. The knowledge and skills expected to favor females included item context favoring females, verbal skills, application of routine mathematical

processing, a finding consistent with previous research (e.g., Casey, Nuttall, Pezaris, & Benbow, 1995; Halpern, 1997), but that support was less apparent for other sources of gender differences in the Gallagher et al. (2000) taxonomy.

*Using the DIF Analysis Framework to Enhance the Identification and Interpretation of DIF*

The multidimensional DIF framework, as described and illustrated in this paper, helps overcome some of the limitations associated with the traditional approach to DIF detection and interpretation. Rather than using exploratory logic, the DIF analysis framework 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. Organizing principles [e.g., Gallagher et al. (2000) taxonomy of content areas and cognitive characteristics believed to produce gender differences in mathematics] can be used to identify and skilled reviewers can be used to structure the studied items. Then, SIBTEST is used to test the studied items. Each DIF study, 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 and study the sources of DIF so a body of *confirmed* DIF hypotheses can be created (Stout & Roussos, 1995). These confirmed hypotheses, accumulated over studies, may lead to a better understanding of why DIF occurs.

Rather than focusing on single items, the DIF analysis framework can be used, more broadly, 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 measures a secondary dimension, thereby producing DIF. In some cases, however, 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), but a bundle of items provide 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

---

solutions to new, unfamiliar situations, application of routine mathematical solutions to familiar situations, memorization skills, and symbolic processes. These skills are labeled on the x-axis.

(Nandakumar, 1993). Results from DBF analyses may also be more interpretable than results from differential item functioning analyses because the bundle represents a larger sample of the secondary dimension. Although virtually unexplored, the use of organizing principles in a substantive review to create and then test DIF hypotheses greatly enhances the potential of the DIF analysis framework for helping researchers and practitioners identify and interpret group differences on tests (Stout, 2002)

Rather than using an informal multidimensional model, the DIF analysis framework is guided by a formal multidimensional model for understanding how DIF occurs. This framework emphasizes that a careful study of the underlying dimensions of a test is needed. Essentially, the model emphasizes 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 what might be thought of as a disproportionate difference between the reference and focal group relative to what should be observed on items that measure the primary dimension.

The multidimensional model can also guide the formation of the matching and studied subtests. The matching subtest contains items believed to measure only the primary dimension, whereas the studied subtest contains the item or bundle believed to measure the primary and secondary dimensions based on the substantive analysis. The matching subtest in the DIF analysis framework is clearly related to the matching subtest derived from iterative purification in the traditional approach. Specifically, items that are "pure" measures of the primary dimension should not display DIF and thus would be identified as candidate items in the traditional approach. However, there are some subtle but important differences. For instance, non-DIF items would be included in the matching subtest following the iterative purification procedure but should not automatically be regarded as candidates for the matching subtest in the confirmatory DIF analysis approach. Indeed, if it is known or suspected that an item is not measuring the primary dimension by studying outcomes from previous DIF analyses, analyzing archival or existing test data, or formulating DIF hypotheses from content reviews, then this item may be excluded from the matching subtest, regardless of the magnitude of DIF. Also, DIF items are automatically excluded from the matching subtest when exploratory purification is used because statistical outcomes supersede insights based on substantive reviews. However, some DIF items

that are strongly believed to measure the primary dimension may be tolerated in the multidimensional framework.

These differences highlight the importance of conducting statistical and substantive analyses when deciding which items should be included or excluded from the matching and studied subtests. Because the DIF framework incorporates a two-stage approach, it can be used to overcome some of the limitations inherent to iterative purification in the traditional approach. For example, exploratory single-item DIF analyses can result in inflated Type I errors. Therefore, a substantive analysis should supplement the statistical analysis so the DIF items slated for deletion from the matching subtest can, in fact, be attributed to a secondary dimension and not a Type I error. Single DIF items may be difficult to identify and interpret. Therefore, DBF analyses should be considered when attempting to differentiate the primary and secondary dimensions because item bundles may be easier to identify statistically and interpret substantively compared to single items. DIF can be attributed to auxiliary or nuisance secondary dimensions (i.e., secondary dimensions causing impact or bias, respectively). Therefore, substantive analyses should be undertaken to identify and differentiate items associated with these dimensions. In sum, adequate statistical procedures and useful substantive analyses are required to identify and interpret the dimensions producing differences before the matching and studied subtest can be created.

#### *Identifying the Matching and Studied Subtest: A Comparison of Two Approaches*

Identifying a valid matching subtest is imperative for accurate DIF detection because examinees are matched on this primary dimension before they are compared. The traditional approach often draws on a statistically-based iterative purification procedure to identify the matching subtest which, in turn, is used to identify DIF items on the studied subtest. The DIF analyses framework, by comparison, uses both statistical and substantive procedures to identify the matching and studied subtest. These two approaches can lead to different statistical and substantive outcomes as illustrated in the following two examples.

In the first example, both approaches were compared using data from the Gierl et al. (in press) gender difference study, as described earlier. Iterative purification was first used to identify a matching subtest. To purify the data, items were analyzed for DIF using a single-item analysis

and any items with large DIF (i.e.,  $\hat{b}_{UNI} \geq 0.050$  and  $p < 0.05$ ) were removed. Then, the purified matching subtest was used for another DIF analysis. The results are presented in Table 3. The matching subtest contained 45 items for the 1996 administration and 44 items from the 1997 administration. The matching subtest, in turn, was used to test ten studied items from the 1996 administration and 11 studied items from the 1997 administration. The studied items identified with iterative purification are also shown in Figures 2a and b for the 1996 and 1997 data, respectively, along with the entire item set reported by Gierl et al. using the Gallagher et al. organizing principle. A different and less interpretable result appears when results from iterative purification are compared with the results from DIF analysis framework. For example, both procedures identify spatial items that favor males. However, the iterative purification approach only flagged the subset of spatial items with large DIF values. Iterative purification also identified a subset of verbal items favoring males and females. Yet when the verbal items were considered as a bundle, as in the DIF analysis approach, no systematic gender differences were apparent. Similarly, the iterative purification approach identified a subset of items requiring the application of routine mathematical solutions to familiar situations favoring males in 1996 and both females and males in 1997. Yet, again, when the items in this category are considered as a bundle, there is no systematic gender difference. The memorization items that elicited small but systematic gender differences in the DIF analysis approach were not identified using iterative purification.

In the second example, data from the Psychopathy Checklist-Revised (PCL-R; Hare, 1991) were used. The PCL-R is a 20-item instrument containing a list of attributes intended to describe the prototypical psychopath. Each attribute is scored by a trained rater following an interview with the respondent and a review of the respondent's file history. A total score of 30 is recommended as a suitable cutoff for distinguishing psychopaths from non-psychopaths (Hare, 1991). The instrument is commonly used in prisons, and has come to play an important role in decisions related to sentencing, parole, and prescription of therapeutic interventions. Each item is scored on a 0-2 scale, with 0 indicating the absence of the attribute, 1 the possible presence of the attribute, and 2 definite presence of the attribute. Table 4 contains the PCL-R items.

To illustrate the potential problems with an iterative purification approach, we used data from English and North American prisoners to evaluate DIF in the PCL-R. Item response patterns for

3282 North American and 849 English prisoners were compared. In the first analysis, an iterative purification approach was used to identify items for the matching subtest. For this purpose, Poly-SIBTEST was used (Chang, Mazzeo, & Roussos, 1996). To purify the data, items were analyzed for DIF using a single-item analysis and any items with large DIF (i.e.,  $\hat{b}_{UNI} \geq 0.100$  or  $p < 0.05$ )<sup>2</sup> were removed. Then, the purified matching subtest was used for another DIF analysis. Results are presented under the heading *Exploratory* in Table 5. The matching subtest contained eight items. Because limited amounts of DIF were found in these items, their use in the matching subtest would appear to be supported statistically. This matching subtest, in turn, was used to test 12 studied items. Taken together, the studied items imply an expected PCL-R score 0.143 points lower for North American prisoners compared to English prisoners conditional on the matching subtest score.

What has been ignored, however, is the validity of the items included in the matching subtest, as they reflect the intended construct of psychopathy. As has been emphasized in this paper, such decisions should be based on substantive and statistical outcomes, especially considering that iterative purification is not guaranteed to produce a meaningful partitioning of the matching and studied subtest. In practice, items on the PCL-R are frequently categorized as affective/personality items and behavioral items, where the affective/personality items (Items 1, 2, 4, 5, 6, 7, 8, and 16) are believed to lie closer to the core of psychopathy (Hare, 1991). Indeed, in cross-cultural comparisons, it is the behavioral items that are expected to perform most differentially across group (Hare, 1991). This finding could be used to guide any cross-cultural PCL-R DIF study, where the affective/personality items could justifiably serve as the matching subtest. (Note that the matching subtest identified using statistical outcomes in the iterative purification approach was a mix of both affective/personality and behavioral items.) Table 5 (under the heading *Confirmatory*) illustrates how the affective/personality items perform when studied as a stand-alone matching subset using Poly-SIBTEST. Although there is more DIF in the matching subtest relative to the iterative purification approach, these items are more conceptually appealing because they measure and match examinees on a well-understood

---

<sup>2</sup> Slightly different criteria were used for identifying DIF items in the Poly-SIBTEST analysis when compared with the SIBTEST analysis, reported in the previous example, because the PCL-R items are scored on a three-category scale.

dimension closely related to psychopathy. Table 5 also contains the DIF results for each behavioral item. Each  $\hat{b}_{UNI}$  now represents the expected score difference for an English and North-American prisoners matched according to their levels on the affective/personality dimension of psychopathy. For example, these results suggest that Item 19, "Revocation of conditional release" elicits a much higher score for North American prisoners, conditional on the affective/personality matching subtest, compared to the English prisoners. Perhaps more interesting is the overall conclusion regarding the cumulative effects of DIF: The iterative purification procedure indicates that DIF favors the English prisoners ( $\hat{b}_{UNI} = -0.143$ ) whereas the theory-driven procedure indicates that DIF favors the North Americans prisoners ( $\hat{b}_{UNI} = 0.401$ ), and by a much larger difference.

#### Summary

Differential item functioning studies are designed to identify and interpret construct-related dimensions that elicit group differences. The traditional approach, where statistically identified items are substantively interpreted, generally has not increased our understanding for why DIF occurs despite its frequent use. The failure to link statistical and substantive outcomes with the traditional approach may be attributed to the exploratory analysis of single items where DIF is linked only superficially to the multidimensional data structure. To address this problem, Roussos and Stout (1996) proposed the multidimensionality-based DIF analysis framework to unify the substantive and statistical approach to DIF detection and interpretation. The first stage is a substantive analysis where DIF hypotheses are generated. The DIF hypothesis specifies whether an item or bundle designed to measure the primary dimension also measures a secondary dimension, thereby producing DIF. To decide whether the data contain distinct dimensions, organizing principles are used to identify items or bundles of items that share specific characteristics. The second stage 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. By combining substantive and statistical analyses, researchers can identify the sources of DIF, thereby increasing our understanding of *why* DIF occurs. Moreover, the DIF analysis framework helps overcome the limitations of the traditional approach: It is based on the confirmatory logic of hypothesis testing, which increases the

interpretability of items that display DIF; it can be used to identify and interpret single items and bundles of items; and it is guided by a formal multidimensional model for understanding how and why DIF occurs.

#### Implications for Practice

Often, researchers and practitioners focus on the studied subtest (i.e., DIF items) without considering carefully the matching subtest. For example, the traditional approach applies statistical and substantive methods to the evaluation of items on the studied subtest but only statistical methods--typically, using iterative purification--to the evaluation of items on the matching subtest. Little or no attempt is made to interpret DIF before items are removed from the matching subtest. But, as we emphasized, iterative purification is not guaranteed to produce a meaningful partitioning of the matching and studied subtest. In our two examples, iterative purification was used to identify studied items with large DIF statistical values relative to a matching subtest with items displaying small DIF statistical values.

The DIF analysis framework, by comparison, guides the development of the matching and studied subtest. The matching subtest serves as a unidimensional criterion designed to place examinees at the same score level so their performance on the studied subtest can be compared. Conversely, the studied subtest contains items suspected of measuring a multidimensional criterion based on the substantive analysis. A DIF analysis is then conducted to compare the studied subtest relative to matching subtest to determine whether the secondary dimension can be detected. If the secondary dimension is detected, it can be attributed to the interpretable characteristics identified in the substantive analysis. For instance, Gierl et al. (in press) identified and tested a spatial studied subtest believed to favor males and a memorization studied subtest believed to favor females relative to a matching subtest containing items that showed no systematic gender differences. This item partitioning provided a test of two different cognitive skills believed to produce gender differences in mathematics relative to items that did not appear to elicit any systematic gender differences. The PCL-R data were used to test 12 behavioral items on the studied subtest relative to an eight-item, theoretically-based, affective/personality matching subtest. This item partition provided a test of the studied subtest score difference between English and North-American prisoners on the behavioral dimension of psychopathy relative to those prisoners with the same score on the affective/personality dimension of

psychopathy. These examples are intended to demonstrate that researchers and practitioners who want to identify and interpret DIF items should attend to the matching and studied subtest because both subtests must be empirically validated as a meaningful comparative criterion in the study of group differences. This form of validation requires comprehensive statistical and substantive analyses, like the approach described by Roussos and Stout (1996) in their multidimensional DIF framework.

## References

- Abbott, M. (2003). Gender equity in Alberta's Social Studies 30 diploma examinations. *Alberta Journal of Educational Research*. Revised manuscript resubmitted for review.
- 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.
- Allalouf, A., Hambleton, R., & Sireci, S. (1999). Identifying the causes of translation DIF on verbal items. *Journal of Educational Measurement*, 36, 185-198.
- 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.
- Bolt, D., Froelich, A. G., Habing, B., Hartz, S., Roussos, L., & Stout, W. (in press). *An applied and foundational research project addressing DIF, impact, and equity: With applications to ETS test development* (ETS Technical Report). Princeton, NJ: Educational Testing Service.
- Bolt, D. & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika*, 23, 67-95.
- Bond, L. (1993). Comments on the O'Neill and McPeck paper. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 277-279). Hillsdale, NJ: Lawrence Erlbaum.
- 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.
- Chang, H.H., Mazzeo, J. & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, 33, 333-353.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, 17, 31-44.
- Dorans, N. J., & Kulick, E. M. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach* (ETS Technical Report RR-83-9). Princeton, NJ: ETS.
- Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and Standardization. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

- 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.
- 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.
- 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., Bisanz, J., Bisanz, G., & Boughton, K. (in press). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the DIF analysis framework. *Journal of Educational Measurement*.
- 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, April). *Consistency between statistical procedures and content reviews for identifying translation DIF*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Quebec, Canada.
- Halpern, D. F. (1997). Sex differences in intelligence: Implications for education. *American Psychologist*, 52, 1091-1102.
- Hare, R.D. (1991). *Manual for the Revised Psychopathy Checklist (Second Ed.)*. Toronto: Multi-Health Systems.
- Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.
- Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum
- Hunter, J. F. (1975, December). *A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items*. A paper presented at the National Institute of Education Conference on Test Bias. Annapolis, MD.

- 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.
- Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement*, *32*, 131-144.
- 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.
- Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297-334.
- 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.
- 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.
- 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.
- 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.
- Shealy, R., & Stout, W. F. (1993a). 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. (2002). Psychometrics: From practice to theory and back. *Psychometrika*, *67*, 485-518.
- Stout, W., Li, H. H, Nandakumar, R., & Bolt, D. (1997). MULTISIB: A procedure to investigate DIF when a test is intentionally two-dimensional. *Applied Psychological Measurement*, *21*, 195-213.
- 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.
- Vandenberghe, C., & Gierl, M. J. (2001). *Differential bundle functioning on mathematics and science achievement tests: A comparison of non-aboriginal and aboriginal students*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.
- 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.
- 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.

Table 1

*Differential Bundle Functioning Results for the Grade 9 Mathematics Achievement Tests*

	Bundle	No. of Items	$\hat{b}_{UNI}$	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.

Table 2

*Dimensionality Assessment Results for the Grade 9 Mathematics Achievement Tests*

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 five categories, as in the differential bundle functioning analysis in Table 1.

Table 3

*SIBTEST Results for the Math 9 1996 and 1997 Items Using Iterative Purification*

1996		1997	
Matching	Studied	Matching	Studied
Item	$\hat{b}_{UNI}$	Item	$\hat{b}_{UNI}$
1	0.004	2	0.067 *
3	0.003	4	0.091 *
5	-0.037 *	7	0.055 *
6	-0.033 *	9	0.057 *
8	0.026 *	14	0.046 *
10	0.029 *	20	0.057 *
11	0.032 *	33	0.084 *
12	0.020 *	37	0.055 *
13	0.005	47	-0.058 *
15	0.000	52	-0.089 *
16	-0.007	<b>Total</b>	<b>0.448</b>
17	0.024 *		
18	0.007		
19	-0.017 *		
21	0.037 *		
22	0.005		
23	0.004		
24	0.001		
25	0.016 *		
26	-0.009		
27	-0.035 *		
28	0.012		
29	0.009		
30	-0.001		
31	0.018 *		
32	-0.042 *		
34	-0.016 *		
35	0.043 *		
36	0.027 *		
38	0.003		
39	-0.004		
40	-0.018 *		
41	-0.043 *		
42	-0.027 *		
43	-0.043 *		
44	-0.015 *		
45	-0.012		
46	0.046 *		
48	0.042 *		
49	-0.046 *		
50	-0.023 *		
51	0.032 *		
53	-0.003		
54	0.000		
55	-0.008		
		1	-0.001 *
		3	-0.002
		5	-0.048 *
		6	0.044 *
		7	0.041 *
		8	0.043 *
		9	-0.009
		10	-0.013
		11	0.024 *
		12	-0.029 *
		13	0.009
		15	0.001
		16	0.013 *
		17	0.015
		18	-0.013 *
		19	-0.036 *
		20	-0.042 *
		22	-0.009
		23	-0.022 *
		24	0.007
		25	-0.010
		26	-0.029 *
		27	0.027 *
		28	0.033 *
		30	-0.002
		31	0.041 *
		32	-0.025 *
		33	-0.028 *
		34	-0.015
		36	-0.029 *
		37	0.006
		39	-0.029 *
		40	0.045 *
		41	-0.006
		42	-0.018 *
		43	-0.029 *
		45	0.016 *
		46	0.049 *
		47	-0.026 *
		48	0.022 *
		49	-0.005
		52	0.033 *
		54	0.004
		55	-0.002
		<b>Total</b>	<b>0.189 *</b>

\*  $p < 0.05$

Table 4.

*Items on the Psychopathy Checklist-Revised*

---

1. Glibness/superficial charm
  2. Grandiose sense of self-worth
  3. Need for stimulation/proneness to boredom
  4. Pathological lying
  5. Conning/manipulative
  6. Lack of remorse or guilt
  7. Shallow affect
  8. Callous/lack of empathy
  9. Parasitic lifestyle
  10. Poor behavioral controls
  11. Promiscuous sexual behavior
  12. Early behavior problems
  13. Lack of realistic, long-term goals
  14. Impulsivity
  15. Irresponsibility
  16. Failure to accept responsibility
  17. Many short-term marital relationships
  18. Juvenile delinquency
  19. Revocation of conditional release
  20. Criminal versatility
-

Table 5.

*Poly-SIBTEST Results for the Psychopathy Checklist-Revised Items Using Exploratory and Confirmatory Purification Procedures*

<i>Exploratory</i>		<i>Confirmatory</i>	
<i>Matching</i>	<i>Studied</i>	<i>Matching</i>	<i>Studied</i>
Item	$\hat{b}_{UNI}$	Item	$\hat{b}_{UNI}$
6	-0.080*	1	0.280*
7	0.073*	2	0.131*
8	-0.040	3	-0.081*
11	-0.025	4	0.224*
14	-0.022	5	-0.148*
15	0.052*	9	-0.179*
16	-0.033	10	-0.176*
18	-0.018	12	-0.235*
		13	-0.127*
		17	-0.179*
		19	0.246*
		20	0.101*
		<b>Total</b>	<b>-0.143</b>
		1	0.167*
		2	-0.016
		3	0.067
		4	0.170*
		5	-0.207*
		6	-0.156*
		7	0.040
		8	-0.025
		9	-0.101*
		10	-0.026
		11	-0.015
		12	-0.070
		13	-0.088*
		14	0.075*
		15	0.090*
		16	-0.115*
		17	-0.145*
		18	0.067
		19	0.372*
		20	0.177
		<b>Total</b>	<b>0.401</b>

Figure Caption

*Figure 1.* Gender differences for all items in the 1996 and 1997 tests, using the Gallagher et al. (2000) taxonomy as reported in Gierl et al. (in press).

*Figure 2a.* SIBTEST results using iteration purification and the DIF analysis framework for the Mathematics 9 1996 test.

*Figure 2b.* SIBTEST results using iteration purification and the DIF analysis framework for the Mathematics 9 1997 test.





