

Running head: USE OF DBF ANALYSES

**Illustrating the Utility of Differential Bundle Functioning
Analyses to Identify and Interpret Group Differences on
Achievement Tests^{**}**

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

Shameem Nyla Khaliq
University of Massachusetts, Amherst

Paper Presented at the Annual Meeting of the
American Educational Research Association (AERA)

Seattle, Washington, USA
April 10-14, 2001

^{**}This paper can also be downloaded from the Centre for Research in Applied Measurement and Evaluation (CRAME) website: <http://www.education.ualberta.ca/educ/psych/crame/>

Illustrating the Utility of Differential Bundle Functioning Analyses to Identify and Interpret Group Differences on Achievement Tests

Fairness is a broad and encompassing topic of importance and consequence in educational and psychological testing. In the 1999 Standards for Educational and Psychological Testing, four characterizations of test fairness are presented. First, fair tests must be free from bias. Bias occurs when tests yield scores or promote score interpretations that result in different meanings for members of different groups. Second, test fairness requires that examinees receive just and equal treatment in the testing process. To achieve this outcome, both the test and the testing context must be considered when scores are interpreted for individuals and groups of examinees. Third, test fairness requires equity in the outcomes of testing. Examinees must be given the chance to demonstrate their proficiency on the construct or constructs the test is designed to measure. If a test has no bias and if examinees receive just and equal treatment during the testing process, then examinees with the same standing on the construct or constructs measured by the test should earn the same score. Fourth, test fairness implies that examinees in the achievement domain have had the opportunity to learn the content covered on the tests. In short, test fairness is a multi-faceted concept with no single technical definition. Moreover, the interpretation of fairness can, and likely will, vary with the testing context.

Differential item functioning (DIF) analyses can yield information about bias. Methods for DIF analyses reflect, in large part, a response to the legal and ethical need to assess examinees without bias, an important consideration for ensuring fair testing practice (Standards of Educational and Psychological Testing, 1999). Generally, examinees are split into two groups, a reference group and a focal group. DIF analysis involves administering a test, matching members of the reference and focal groups 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. DIF can be identified with a variety of statistical methods (see reviews by Millsap & Everson, 1993; Clauser & Mazor, 1998).

If DIF is detected, the possibility exists that the item is biased in some way against one of the groups. At this point it would be possible to delete the item from the test in an effort to reduce a possible source of bias. To do so without understanding why DIF exists would be premature, however. It is entirely possible, for example, that the item is a perfectly relevant measure of the target construct and that the difference between reference and focal groups reflects a true difference in that construct. This outcome is often referred to as item impact. Interpreting the nature and bases of this difference could have implications for understanding differences between the groups in terms of their cognitive skills or educational histories, as well as for adapting instruction for one or both groups. Thus some analysis of the substantive basis for the DIF should be undertaken. If an item is flagged for DIF and an evaluation of the item reveals that the source of DIF is not relevant to the intended purpose of the test, then the item can be considered biased and it may be deleted from the test (Camilli & Shepard, 1994).

The following example, from Dorans and Kulick (1983), illustrates how this process leads to judgments about item bias. 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 a DIF statistical method (the standardization approach), 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 that was extraneous to analogical reasoning: "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)

Using this approach to detect sources of bias requires both adequate statistical procedures and useful substantive analyses that enable valid interpretation of the sources of DIF. Considerable progress has been made in developing and refining statistical methods for flagging items showing DIF (Angoff, 1993, p. 21) but, as might be inferred from the example in the preceding paragraph, methods to aid substantive interpretation have lagged far behind (e.g.,

Bond, 1993; Camilli & Shepard, 1994, Englehard, Hansche, & Rutledge, 1990; Gierl, Rogers, & Klinger, 1999, O'Neill & McPeck, 1993; Standards for Educational and Psychological Testing, 1999; Sudweeks & Tolman, 1993). The typical approach to substantive analysis has been to subject individual DIF items to the scrutiny of "experts" (e.g., curriculum specialists or test developers) for interpretation and judgments about sources of bias. This approach has not been very successful because these judgments tend to be inconsistent with the DIF statistics or unreliable among the judges. Camilli and Shepard (1994) reported that, in their experience, as many as half of the items with "large" DIF might not be interpretable. After reviewing the literature, Roussos and Stout (1996a) concluded that "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). Authors of the 1999 Standards for Educational and Psychological Testing concurred with these findings by stating: "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" (p. 78). This impasse represents one important problem in the study of bias and, therefore, test fairness.

Fortunately, elements of a successful means for bridging this impasse are beginning to emerge. Because commonly used statistical procedures based on a unidimensional concept of ability can lead to erroneous conclusions about DIF, multidimensional methods have been developed (Ackerman, 1992; Hunter, 1975; Kok, 1988; Shealy & Stout, 1993a). Because interpreting the substantive bases of DIF after an independent statistical analysis has not been very successful, methods of using substantive analysis prior to statistical analysis have been developed (e.g., Roussos & Stout, 1996a). Because sources of DIF may be more apparent in patterns across multiple items rather than in performance characteristics associated with single items, methods have been developed for assessing differential functioning in bundles or sets of items that share some potentially important characteristics (Douglas, Roussos, & Stout, 1996; Oshima, Raju, Flowers, & Slinde, 1998). These advances enable new approaches to the study of differential group performance, approaches that have considerable potential for integrating psychometric and psychological aspects of test performance.

In the first section of this paper we describe these recent advances. In the second section we illustrate some of the ways in which substantive analyses can be integrated with multidimensional, bundles-based, statistical methods using test specifications to organize the analyses. In the third section we highlight some future directions for analyses of DIF. As this approach develops, it could be used to identify the sources of differential group performance and lead to new explanations about group differences, new interpretations about bias and impact and, possibly, new approaches to instruction.

Recent Advances in Multidimensional DIF Analyses

Roussos and Stout (1996a) proposed a two-stage approach to bridge the gap between substantive and statistical analyses by linking both to the Shealy-Stout multidimensional model for DIF (Shealy & Stout, 1993a). The first stage is a substantive analysis in which DIF hypotheses are generated. The second stage is a statistical analysis of the DIF hypotheses. By combining substantive and statistical DIF analyses, researchers can begin to systematically identify and study the sources of DIF. Moreover, the power of this approach can be enhanced when it is used with bundles of items, rather than with individual items.

Shealy-Stout Multidimensional Model for DIF (MMD)

MMD is a framework for understanding how DIF occurs. It is based on the assumption that multidimensionality produces DIF. A dimension is a substantive characteristic of an item that can affect the probability of a correct response. The main construct that the test is intended to measure is the primary dimension. DIF items are believed to measure at least one dimension in addition to the primary dimension (e.g., Ackerman, 1992; Camilli & Shepard, 1994; Hunter, 1975; Kok, 1988; Lord, 1980; Roussos & Stout, 1996a; Shealy & Stout, 1993a). Dimensions that produce DIF are referred to as the secondary dimensions. When primary and secondary dimensions characterize item responses, the data are considered multidimensional. The secondary dimensions are interpreted further. The secondary dimensions are considered auxiliary if they are intentionally assessed as part of the construct on the test. Alternatively, the secondary dimensions are considered nuisance if they are unintentionally assessed as part of the

construct on the test. DIF caused by auxiliary dimensions is benign (reflecting impact) whereas DIF caused by nuisance dimensions is adverse (reflecting bias).

Linking Substantive and Statistical DIF Analyses

The Roussos-Stout DIF analysis paradigm is built on the foundation provided by MMD. The first stage is a substantive analysis in which 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. Roussos and Stout (1996a) contend that MMD can be used to minimize bias in tests if developers can generate accurate DIF hypotheses based on their understanding of the substantive characteristics that may differ for particular subgroups of examinees, and link these characteristics to the underlying dimensional structure of the test data. Substantive characteristics include item content that may appeal, differentially, to one group of examinees (e.g., the supposed bias favoring males in knowledge about hunting and fishing in the example at the beginning of this paper), differential task demands, or groups differences. The link between substantive characteristics and the dimensional structure may be enhanced by studying outcomes from previous DIF analyses, analyzing archival or existing test data for DIF, formulating DIF hypotheses through content reviews by test developers, and testing bundles of items.

The second stage in the Roussos-Stout DIF analysis paradigm is statistically testing the DIF hypotheses. The Simultaneous Item Bias Test (SIBTEST) can be used to test DIF hypotheses and quantify the size of DIF. To operationalize SIBTEST, items on the exam are divided into the studied (or “suspect”) subtest and the matching (or “valid”) subtest. The studied subtest contains the item or bundle believed to measure the primary and secondary dimensions based on the substantive analysis whereas the matching subtest contains the items believed to measure only the primary dimension. The matching subtest places the reference and focal group examinees into subgroups at each score level so their performance on items from the studied subtest can be compared.

The amount of DIF for an item is reflected in a parameter estimate, \hat{b}_{UNI} , that has a standard normal distribution with mean 0 and standard deviation 1 under the null hypothesis of no DIF. A

statistically significant value of \hat{b}_{UNI} that is positive indicates DIF against the focal group, whereas a negative value indicates DIF against the reference group. Research at the Educational Testing Service (ETS) has resulted in guidelines for classifying DIF as negligible, moderate, or large using the Mantel-Haenszel statistical approach (Zieky, 1993, p. 342). Roussos and Stout (1996b, p. 220) adopted the ETS guidelines and applied the results to SIBTEST. They proposed the following guidelines to classify DIF on a single item:

- Negligible or A-level DIF: Null hypothesis is rejected and the absolute value of $\hat{b}_{UNI} < 0.059$,
- Moderate or B-level DIF: Null hypothesis is rejected and $0.059 \leq |\hat{b}_{UNI}| < 0.088$, and
- Large or C-level DIF: Null hypothesis is rejected and $|\hat{b}_{UNI}| \geq 0.088$.

A complete technical description of SIBTEST is found in Shealy and Stout (1993a).

Differential Bundle Functioning

Douglas et al. (1996) used the Shealy-Stout (1993a) MMD to study the effects of differential bundle functioning (DBF). A bundle is a set of test items that are presumed to measure a common secondary dimension. Bundle analysis, recall, was one of the substantive methods suggested by Roussos and Stout in their DIF analysis paradigm for understanding the nature of differential group performance. Statistical methods for flagging bundles are more powerful than methods for flagging individual items. Consequently, this approach is often preferred, especially when small performance differences on single items combine to produce a large performance difference across all the items in a bundle (Nandakumar, 1993). A bundle is created according to an organizing principle. Douglas et al. proposed two general methods for organizing items to create bundles. The first is a confirmatory approach in which experts, typically content specialists, use their judgements to identify and group items that are believed to measure a common secondary dimension. These bundles are then tested to see if they differ from the matching subtest. The second is an exploratory approach in which statistical procedures are used to identify distinct dimensions. Many statistical procedures are available to bundle items including factor analysis, cluster analysis, and multidimensional scaling, to name but a few (for a

review, see Hattie, 1985). The dimensions identified with these procedures are used to form bundles that are interpreted and then tested. Using SIBTEST, for example, \hat{b}_{UNI} reflects the amount of DBF for a bundle, just as it reflects the amount of DIF for a single item. Unlike DIF, however, no guidelines exist for classifying \hat{b}_{UNI} on bundles as negligible, moderate, or large.

Using Test Specifications to Identify and Interpret Group Differences

When implementing a confirmatory approach, test specifications (also called table of test specifications or the test blueprint) could serve as a convenient organizing principle to bundle items. This approach to the study of DBF has not been carefully documented in the psychometric literature (see, however, Oshima et al., 1998), yet it represents a viable tool for identifying and interpreting items that function differently across groups.

Test specifications are used to outline the achievement domain and provide a guideline for obtaining a representative sample of items from the domain during test construction. A common structure for these specifications is to create a two-way matrix in which the rows represent the content areas to be measured on the test and the columns represent the cognitive skills required to solve items. Items in each cell of the matrix are then created by test developers who attempt to anticipate the cognitive skills used by the examinees to solve items in each content area (Millman & Green, 1989). Content areas reflect the curricular domain, which is often tied to the examinees' program of studies. By adequately representing the content areas in the test specifications, test developers are attempting to establish the validity of test score inferences based on content (Ackerman, 1994). Categories of cognitive skill are intended to reflect the thought processes used by the examinees to solve test items. These skills are often based on the Taxonomy of Educational Objectives (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956). This taxonomy 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). By adequately representing the cognitive skills in the test specifications, test developers are attempting to establish the validity of the test score inferences based the examinees' cognitive skills.

As an illustration, the test specifications from a 1997 and 1998 Grade 6 science achievement test used in Canada are presented in Table 1. The science test contained 50 multiple-choice items, each with four options. Items on the test were based on concepts, topics, and facts from the examinees' program of studies. Test developers classified items into five content areas: evidence and investigation, air and aerodynamics, sky science, observation and inference, and trees and environment. Items were also classified into two cognitive categories: knowledge and skills. The test specifications are a function of crossing the five content areas with the two cognitive levels to produce a 10-cell matrix. A similar format is used in other large-scale testing programs such as the School Achievement Indicators Project in Canada and the National Assessment of Educational Progress in the United States. Assuming that item classification is both reliable and valid, test specifications can serve as the organizing principle for bundling items in the study of differential group performance because they provide the developers' representation of the content and cognitive domain, they are readily available for researchers and practitioners, and they are easy to use.

To illustrate the use of test specifications as the organizing principle to study gender differences in science, two analyses were conducted. In the first analysis, we used content area as the organizing principle, and in the second we used cognitive category. Comparing these two analyses illustrates the usefulness of these two organizing principles for studying group differences. These analyses were conducted on response data from 12000 Grade 6 students (6000 males and 6000 females) who wrote the 1997 administration of the Science achievement test developed and administered in the Canadian province of Alberta. The stability of the 1997 results were also evaluated by using the same test specifications as the organizing principle to study gender differences using data from a 1998 administration. A complete description of these tests is available from the Alberta Learning internet site (http://ednet.edc.gov.ab.ca/k_12/testing/). Large samples were used to produce stable results, although SIBTEST can yield adequate results even with samples as small as 250 examinees per group (Roussos & Stout, 1996b). We conclude this section by comparing the interpretability of a single-item DIF analysis with that of the DBF analyses.

Analysis #1: Content Area as the DBF Organizing Principle

Five content areas from the test specifications shown in Table 1 were used to organize items into bundles in the first analysis. The items in each content area were classified by test developers. The analysis was conducted in four steps. First, all DIF items were identified with SIBTEST using a single item analysis (i.e., studying one item at a time and using the remaining items as the matching subtest) to obtain the DIF effect size measure, \hat{b}_{UNI} , for each item.

Second, items were grouped by the five content areas from the test specifications and the \hat{b}_{UNI} s for these items were graphed. Third, bundles were identified by visually examining the graphs and looking for interpretable patterns. Other, more numerical methods could be used to identify bundles in this step (e.g., Roussos, Stout, & Marden, 1998) but simple graphical analysis serves well in this case for illustrating the process of identifying bundles. Fourth, the bundles identified in step 3 were tested using the remaining items (i.e., items that did not form interpretable bundles) as the matching subtest after deleting anomalous items that displayed C-level (large) DIF. C-level DIF items were removed from the matching subtest to ensure that it was a homogenous measure across the two groups.

The bundles from the analysis of the 1997 Science achievement test data are shown in Figure 1 as circles. The x-axis shows the five content areas and the y-axis represents the \hat{b}_{UNI} value for each item. Positive \hat{b}_{UNI} values favor males, whereas negative \hat{b}_{UNI} values favor females. A-level (negligible) DIF items are solid black circles, whereas open white circles represent B-level (moderate) and C-level (large) DIF items. In this example, bundles associated with content area 2, air and aerodynamics, and content area 4, observation and inference, are apparent and favor males and females, respectively. Items for the remaining three content areas are evenly distributed, for the most part, between the two groups. Table 2 contains the outcome from the statistical test of each bundle. The bundles were tested using the remaining items as the matching subtest except for those items displaying C-level or large DIF. All hypotheses were tested with a directional test at a conservative alpha level of 0.01. Males in Grade 6

systematically differ from females with comparable science test scores on their knowledge and skills in the content area 2, air and aerodynamics. Females in Grade 6 systematically differed from males with comparable science test scores on their knowledge and skills in the content area 4, observation and inference.

The stability of this outcome was evaluated by comparing the 1998 results to the 1997 results. The 1998 Science test was created using the same test specifications and development process. Bundles from the 1998 analysis are shown as diamonds in Figure 1. Similar patterns emerged for bundles associated with content area 2 and 4. Table 2 contains the outcome from the statistical test of each bundle. These results indicate that the DBF outcomes are stable over a two-year period using a different cohort of examinees, and they support the interpretation that males and females systematically differ in these two content areas.

To summarize, test specifications frequently include diverse content areas. The results of these analyses confirm that content areas can be used effectively as an organizing principle for identifying and interpreting gender differences. As a next step, the nature of the secondary dimension would be determined by interpreting each bundle as either an auxiliary or a nuisance dimension given the purpose of the test. Our objective is not to make this interpretation but rather to illustrate how DBF analyses can be used to identify bundles of items that elicit group differences, which in turn require interpretation.

Analysis #2: Cognitive Category as the DBF Organizing Principle¹

The cognitive categories outlined in Table 1 were used as the organizing principle in the second analysis. Grade 6 Science achievement test items were classified into two cognitive categories, knowledge and skills. Knowledge is described as understanding the concepts and processes of science by the developers for this test. Skills refer to the application of knowledge. Both components are based on the cognitive categories described in Bloom et al. (1956). The same procedure was used as in Analysis #1. The graph for this analysis is presented in Figure 2. The x-axis is the cognitive level and the y-axis is the \hat{b}_{UNI} value for each item produced by SIBTEST. In this example, no clear patterns appear for the 1997 or 1998 Science test. Instead,

the items are evenly distributed across the two cognitive levels with only one item clearly favoring females in knowledge on the 1997 test.

The failure of this analysis to yield evidence of DBF can lead to at least two conclusions. First, the organizing principle may be valid and there simply may be no gender differences in either aspect of cognition. Second, the organizing principle may not be valid or, at least, not sufficiently sensitive to capture aspects of cognition in which the genders differ. The second possibility requires some consideration.

As previously noted, items for each cell in the test specifications are created by writers who categorize the items into content areas and who try to anticipate the cognitive processes examinees use to answer the questions correctly. Consequently, the table of specification provides a test developer's representation of the content areas and cognitive categories. Items are readily classified into content areas based on the intended purpose of the test and on the development process that supports this purpose (i.e., tests are developed from items that are designed to cover specific content areas). However, it is much more difficult to classify items by cognitive categories because it requires that test developers anticipate how students solve items (cf. Bisanz, Bisanz, & Lefevre, 1984). Moreover, the emphasis during test construction 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. Indeed, in many or most cases the cognitive processes used by students are not known with certainty. Item writers usually are content specialists working from test specifications that have no formal relation to current psychological theory (e.g., Bejar, 1993; Embretson, 1985; Hattie, Jaeger, & Bond, 1999; Mislevy, 1996; Nichols, 1994; Nichols & Sugrue, 1999; Snow & Lohman, 1989; Snow & Peterson, 1985). From a psychological perspective, for example, "content area" as described in Table 1 seems highly redundant with "knowledge", conceived of as understanding both concepts and processes of science. Although little is known about the validity of using Bloom's taxonomy for classifying students' cognitive processes, the available evidence suggests that the taxonomy is inadequate. Gierl (1997), for example, found that Bloom's taxonomy did not provide an accurate

model to guide item writers for anticipating the cognitive processes used by students on a large-scale achievement test in mathematics. The cognitive processes expected by item writers matched the processes actually used by students, as identified with think-aloud protocols, in only 54% of the cases outlined in the test specifications.

The failure to find gender differences related to either cognitive category may be attributable, in part, to an inadequate representation of cognition. For example, only two cognitive skills were used to describe student performance across five content areas for the 50-item test in our study. This scheme simply does not reflect the complexity of processes and knowledge found in contemporary cognitive research (e.g., Sternberg, 1994). Many other large-scale assessments also present a limited view of cognition. For example, the table of test specifications for 1992 NAEP mathematics assessment only specifies three cognitive categories—conceptual knowledge, procedural knowledge, and problem solving—across five content areas (Ballator, 1996). This outcome may also be attributed to the incorrect identification of cognitive skills (i.e., actual cognitive skills used by students may differ from the expected cognitive skills outlined in the test specifications for some of the items; see Gierl, 1997). Clearly there is a need for development of more sophisticated cognitive organizing principles and for greater evidence on the extent to which students actually use specified processes and knowledge when answering test items.

Comparing DIF and DBF

Researchers and practitioners have reported that DIF is difficult to interpret. A comparison of DIF and DBF results helps to account for this finding. In Figure 1, we illustrated how content areas 2 and 4 systematically favored males and females, respectively, using a graphical approach. We also demonstrated that these two bundles differ statistically from a matching subtest across two different test administrations. A very different and less interpretable result appears when we rely on a more traditional single-item analysis. Seven items from the 1997 administration and eight items from the 1998 administration displayed DIF at the B- and C-level. As previously noted, these items are shown in Figure 1 for the 1997 and 1998 administrations with open circles and diamonds, respectively. When the results from the two administrations were considered together, no clear patterns emerge. For example, at least one DIF item was

found in each content area and consistency across administrations was minimal. With the DBF approach, however, a stable and consistent pattern of results is found.

The researcher or practitioner who is attempting to interpret these outcomes—either to make a decision about the DIF items on the test or to identify possible causes of DIF—is presented with different information from these two analyses. At the item level, 15 DIF items were flagged across the 100 items in the two administrations, no pattern is apparent, and one is left to make decisions and inferences about DIF from a very small number of items. At the bundle level, two clusters of items were consistently identified across the two administrations. These results illustrate the advantage of moving from an item to a bundle classification system: When we focus on systematic clusters of items using an organizing principle to structure the data, the interpretability of results is enhanced.

The reason for this outcome is clear: The item can be a poor level of analysis because an item represents a small, relatively unreliable, sample of behavior. By moving to the bundle level, we have a pattern of results representing a larger sample of behavior. Moreover, as Nadakumar (1993) and Douglas et al. (1996) have demonstrated, a single item may not yield an adequate measure of the secondary dimension that produces DIF. Hence, this dimension will not be detected. Bundling items that tap a common secondary dimension amplifies and effectively increases the sensitivity of the analysis for detecting group differences. Because bundles provide a broader sample of the secondary dimensions than any single item, they should be easier to interpret substantively, leading to better explanations about the nature of group differences. This approach should also yield fewer Type I errors because only a relatively small number of DIF hypotheses are tested.

Summary and Future Directions for DIF Research

An “arsenal” of statistical methods exist for identifying DIF (Angoff, 1993, p. 21). Despite the availability of these methods, many researchers agree that interpreting items that display DIF is difficult. To address this problem, Roussos and Stout (1996a) proposed the DIF analysis paradigm that unifies substantive and statistical approaches to DIF detection, thereby providing a framework for linking psychology and psychometrics. This method is based on the

multidimensional model of DIF proposed by Shealy and Stout (1993a). The DIF analysis paradigm is a two-stage approach in which the first stage is a substantive analysis designed to generate DIF hypotheses and the second stage is a statistical analysis designed to test the DIF hypotheses. By combining these analyses, researchers can identify the causes of DIF and create a body of confirmed DIF hypotheses that yield a better understanding of why DIF occurs².

Instead of focusing on group differences at the item level, Douglas et al. (1996) studied group differences at the bundle level. A bundle is a set of test items that tap a common secondary dimension. A bundle is created according to an organizing principle. In the current study, test specifications were used as an organizing principle for bundling items to illustrate the advantages of this approach. Five content areas from a Grade 6 Science achievement test served as the organizing principle in analysis #1. Items from content area 2, air and aerodynamics, consistently favored males and items from content area 4, observation and inference, consistently favored females. These outcomes were found across two test administrations using different groups of examinees. Cognitive level served as the organizing principle in analysis #2. Items from the knowledge and skills cognitive level did not produce an interpretable pattern of results within or across years. This outcome may be attributed to the inadequate representations of cognition currently used in test development.

The DIF analysis paradigm also helps explain why group differences at the item level may be hard to interpret. The item can be a poor level of analysis for DIF because it represents a small and unreliable sample of behavior. Hence, single-item DIF analyses often produce uninterpretable results. By moving to the bundle level, we have a pattern of results representing a larger sample of behavior. Moreover, bundling items that tap a common secondary dimension amplifies this effect and can lead to the detection of group differences. Because bundles provide a broader sample of the secondary dimension, they should be easier to interpret substantively and therefore should lead to better explanations about the nature of group differences. DBF analysis can lead to an interpretable and stable pattern of results, as illustrated in this study.

Focus On Organizing Principles

The examples presented in this paper demonstrate the utility of a bundle analysis compared to a single-item analysis. One challenge ahead, however, is identifying appropriate principles and their strengths and weaknesses. Currently, four general organizing principles can guide substantive analyses in the DIF analysis paradigm.

- Test Specifications: As illustrated in this paper, test specifications can be used to form bundles. Test specifications outline the achievement domain and help developers obtain a representative sample of items. The specifications guide item writing and help structure the final form of the test. This approach has many strengths: The specifications provide the developer's representation of the content and cognitive domain; they are readily available for researchers and practitioners; they are easy to use; and they can guide interpretation. This approach also has weaknesses. For example, if the test specifications are a poor representation of the achievement domain, then this approach will yield inadequate results. Further, it must be assumed that item classification is both reliable and valid. This assumption is, sometimes, incorrect (Ballator, 1996; Paul Nichols, personal communication, June 14, 2000; Silver & Kenney, 1993). Finally, if the test specifications represent items accurately and if members of both the reference and focal groups had equal opportunity to learn, then it would be very difficult to detect adverse DBF with this approach.
- Content Analysis: Content specialists can review items and identify dimensions based on item content. A content analysis is, in the pure sense, guided by the professional experience of the reviewers. This analysis is usually conducted after the test has been created. No statistical results are used to identify dimensions. Two forms may exist: (a) content specialists may use their experience and judgment to identify primary and secondary dimensions during an item review or (b) content-based judgments can be found in the DIF literature. For example, Walter and Young (1997) undertook a content analysis of the multiple-choice section of Social Studies Grade 12 Diploma examination administered in the Canadian province of Alberta. Overall mean scores on this exam frequently favor males. The purpose of their study was to address the gender differences frequently found on this

exam from a qualitative, feminist perspective. Walter and Young concluded that the content of the exam overrepresented the male perspective by focusing on power relations and control, emphasized the conflict model over the nurturing model through word choice and constructions that convey the struggle for power and domination, and highlighted male role models. From their review, they hypothesized that males would outperform females in the examination content areas of Economics, Politics, History, and Control Tactics. They also expected females to outperform males in the area of Peace Initiatives and Internationalism. However, these hypotheses were not tested using empirical data. Thus, the hypotheses about gender differences proposed by Walters and Young could be used to evaluate and statistically test the content areas believed to differentiate males and females in social studies.

- Empirical Analysis: Statistical methods can be used to aid in identifying the test dimensions. The outcomes from these statistical approaches are then interpreted. This approach is substantive to the extent that the bundles identified with the empirical procedures are, in fact, interpretable. Empirical approaches include, but are not limited to, factor analysis, cluster analysis, and multidimensional scaling. For example, Douglas et al. used HCA, a cluster analysis procedure, and DIMTEST, a dimensionality test for sets of items, to evaluate the dimensional structure for history items from the National Assessment of Educational Progress exam. They found a six-item bundle tapping a secondary dimension interpreted as "knowledge of some important documents in early American history". When these items were evaluated as a bundle by gender using SIBTEST, the bundle displayed DBF in favor of males. In this example, exploratory statistical procedures were used to identify a bundle, the bundle was interpreted, and then tested statistically for DBF.
- Psychological Analysis: A psychological analysis involves the use of any hypothesized item structure that can be formulated from a psychological perspective. For example, a cognitive theory may provide some predictions about gender differences. These outcomes may be operationalized using items and tested as part of the studied subtest to evaluate the theory. Gallagher, De Lisi, Holst, McGillcuddy-De Lisi, Morely, and Cahalan (2000), for example,

developed a taxonomy for categorizing mathematics problems on standardized tests to predict gender differences. The taxonomy was based on cognitive characteristics likely to produce gender differences in mathematics, as identified by Halpern (1997), Casey, Nuttall, and Pezaris (1997), and Gallagher (1998). The Gallagher et al. (2000) taxonomy differentiated math items by item context, verbal and spatial demands, content mastery, and strategy selection. They used the taxonomy to empirically demonstrate how cognitive task requirements could differentially affect the performance of males and females on GRE-Q items, thereby producing gender differences favoring males. However, classification outcomes were not reported by cognitive category. As a result, the method presented in this paper could be used to evaluate and statistically test each cognitive category in the Gallagher et al. taxonomy in order to study those cognitive processes that are believed to differentiate males and females in mathematics—we are currently pursuing this line of research.

Aside from these approaches, other organizing principles could be used to structure the data and evaluate group differences, including approaches that combine two or more of the four just described. In fact, many different organizing principles could be used to structure the same items, resulting in interpretable outcomes because groups typically vary on numerous dimensions. This possibility could have important implications for the study of nuisance dimensions that produce bias. It could also have important implications for the study of auxiliary dimensions that produce group differences on variables such as instructional efficacy, curricular change, and opportunity to learn. Thus, organizing principles can have a dramatic impact on the interpretation of group differences identified with DBF analyses.

Conclusion

DBF analyses, like DIF analyses, have limitations. For example, both approaches rely on an internal estimate of ability to match examinees. Therefore, it must be assumed that the total test score (or some estimate of it) is unbiased and that group differences at the bundle or item level relative to the total score provide a meaningful indicator of differential performance. This outcome may not occur if bias is pervasive, affecting all items on the test. In addition, both approaches compare a reference to a focal group. These groups should be well-defined and reasonably

homogenous relative to the construct or constructs measured by the test. When groups are broadly defined such as males versus female, Caucasian versus African-American, or English-speaking versus French-speaking, within-group differences may erode the meaningfulness of between-group differences. Indeed, all DIF detection procedures have requirements and limitations that must be acknowledged (see Clauser & Mazor, 1998, pp. 36-38).

The logic for evaluating DBF also has important advantages over the conventional method of testing a single item. These advantages include improved statistical power for detecting group differences, especially when bundles are used, and reduced Type I error when a small number of hypotheses are tested. Possibly, the greatest advantage of this approach comes from structuring the data using an organizing principle to produce substantively meaningful DBF hypotheses which, in turn, are tested statistically. This approach could help us identify sources of differential group performance and lead to new explanations about the nature of group differences. It is embedded within a paradigm that unifies substantive and statistical approaches to DIF detection, which may yield new methods for bridging the gap between psychology and psychometrics.

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. (1994). Using multidimensional item response theory to understand what items and tests are measuring. Applied Measurement in Education, *7*, 255-278.
- 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.
- 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.
- Bisanz, J., Bisanz, G. L., & Lefevre, J. (1984). Interpretation of instructions: A source of individual differences in analogical reasoning. Intelligence, *8*, 161-177.
- 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.
- 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. (1997). Mediators of gender differences in mathematics college entrance test scores: A comparison of spatial skills with internalized beliefs and anxieties. Developmental Psychology, *33*, 669-680.
- 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.

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.

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.

Emmerich, W. (1989). Appraising the cognitive features of subject tests (Research Rep. No. RR-89-53). Princeton, NJ: Educational Testing Service.

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

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. (1997). 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., 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.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. Applied Psychological Measurement, 9, 139-164.

Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent methodological questions in educational testing. Review of Research in Education, 24, 393-446.

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.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), Educational measurement, (3rd ed., pp. 335-366). New York: American Council of Education, MacMillan.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement, *17*, 297-334.

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.

Nichols, P. (1994). A framework of developing cognitively diagnostic assessments. Review of Educational Research, *64*, 575-603.

Nichols, P., & Sugrue, B. (1999). The lack of fidelity between cognitively complex constructs and conventional test development practice. Educational Measurement: Issues and Practice, *18* (2), 18-29.

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.

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

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

Roussos, L. A., Stout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. Journal of Educational Measurement, *35*, 1-30.

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.

Shealy, R., & Stout, W. F. (1993b). An item response theory model for test bias and differential test functioning. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 197-239). Hillsdale, NJ: Lawrence 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.

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.

Snow, R. E. & Peterson, P. L. (1985). Cognitive analyses of tests: Implications for redesign. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics (pp. 149-166). New York: Academic Press.

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.

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.

Walter, C., & Young, B. (1997). Gender bias in Alberta Social Studies 30 examinations. Canadian Social Studies, 31, 83-86,89.

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.

Author Note

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

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

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

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

Shameem Nyla Khaliq, Research and Evaluation Methods Program, Room 160 Hills South, University of Massachusetts, Amherst, MA 01003-4140

This research was supported with funds from the Social Sciences and Humanities Research Council of Canada (SSHRC). We would like to thank W. Todd Rogers, Paul Nichols, Louis Roussos, and Jeri Benson for their comments. Please address all correspondence to Mark J. Gierl, Centre for Research in Applied Measurement and Evaluation, 6-110 Education Centre North, Faculty of Education, University of Alberta, Edmonton, Alberta, Canada, T6G 2G5.
Email: mark.gierl@ualberta.ca

Footnotes

¹Outcomes from content area-by-cognitive level interaction (i.e., the items in each cell of the Table 1 matrix) could also be graphed and analyzed on tests with a large number of items. In the current study, some cells for the science test had few or no items for specific content area-by-cognitive level combinations. Consequently, this approach was not used.

²It might be apparent, at this point, that the acronym DIF has two interpretations. DIF can refer to the body of research devoted to the study of methods for identifying and interpreting group differences on psychometric instruments. DIF can also refer to differential item functioning as distinct from DBF or DTF (differential test functioning, as described by Shealy & Stout, 1993b). At this point in the discussion, we are referring to the former. But as researchers and practitioners begin to study and implement DBF and DTF analyses, we foresee a time when the acronym DIF may result in confusion because it has two distinct interpretations.

Table 1

Test Specifications for a 1997 and 1998 Grade 6 Science Achievement Test

CONTENT AREA	COGNITIVE CATEGORIES	
	<p><u>KNOWLEDGE:</u> Understanding the concepts and processes of science.</p>	<p><u>SKILLS:</u> Application of knowledge.</p>
<p><u>EVIDENCE AND INVESTIGATION:</u> Work cooperatively with others to design and carry out an investigation in which variables are identified and controlled; and recognize the importance of accuracy in observation and measurement, and apply suitable methods to record, compile, interpret, and evaluate observations and measurements gathered by self and group; and work cooperatively with others in designing and carrying out an investigation of practical problem and in developing a possible solution.</p>	<p><u>1997</u> 0 Items <u>1998</u> 1 Item</p>	<p><u>1997</u> 11 Items <u>1998</u> 7 Items</p>
<p><u>AIR AND AERODYNAMICS:</u> Describe properties of air, and the interactions of air with objects in flight and construct devices that move through air, and identify adaptations for controlling flight.</p>	<p><u>1997</u> 8 Items <u>1998</u> 9 Items</p>	<p><u>1997</u> 5 Items <u>1998</u> 5 Items</p>
<p><u>SKY SCIENCE:</u> Observe, describe, and interpret the movement of objects in the sky, and identify pattern and order in these movements.</p>	<p><u>1997</u> 3 Items <u>1998</u> 3 Items</p>	<p><u>1997</u> 3 Items <u>1998</u> 5 Items</p>
<p><u>OBSERVATION AND INFERENCE:</u> Apply observation and inference skills to recognize and interpret patterns, and to distinguish a specific pattern from among a group of similar patterns, and apply a knowledge of the properties and interactions of materials to the investigation and identification of a material sample.</p>	<p><u>1997</u> 1 Item <u>1998</u> 1 Item</p>	<p><u>1997</u> 9 Items <u>1998</u> 8 Item</p>
<p><u>TREES AND ENVIRONMENT:</u> Describe characteristics of trees and the interaction of trees with other living things in the local environment.</p>	<p><u>1997</u> 6 Items <u>1998</u> 6 Items</p>	<p><u>1997</u> 4 Items <u>1998</u> 5 Items</p>

Table 2

Differential Bundle Functioning Results for the Grade 6 1997 and 1998 Science AchievementTest

	Bundle	No. of Items	Beta-Uni	Favors
1997				
	Content Area 2: Air and Aerodynamics	13	.423*	Males
	Content Area 4: Observation and Inference	9	-.440*	Females
1998				
	Content Area 2: Air and Aerodynamics	14	.500*	Males
	Content Area 4: Observation and Inference	9	-.324*	Females

* $p < .01$.

Note. The matching subtest used in each year was created by combining items from content areas 1, 3, and 5 with the exception of items displaying C-level or large DIF.

Figure Caption

Figure 1. Bundles for the five content areas from the 1997 and 1998 Grade 6 Science achievement test. A-level (negligible) DIF items are solid whereas B-level (moderate) and C-level (large) DIF items are open.

Figure 2. Bundles for the cognitive levels from the 1997 and 1998 Grade 6 Science achievement test.



