APPLIED MEASUREMENT IN EDUCATION, *11*(4), 353–369


# Differential Bundle Functioning Using the DFIT Framework: Procedures for Identifying Possible Sources of Differential Functioning

T. C. Oshima
*Department of Educational Policy Studies*
*Georgia State University*

Nambury S. Raju
*Institute of Psychology*
*Illinois Institute of Technology*

Claudia P. Flowers
*Department of Educational Administration, Research, and Technology*
*University of North Carolina at Charlotte*

Jeffrey A. Slinde
*The Psychological Corporation*
*San Antonio, Texas*

A general framework for assessing differential functioning of items and tests (DFIT) was recently proposed. Both unidimensional and multidimensional tests with either dichotomous or polytomous scoring can be handled within this framework. In this study, the DFIT framework is expanded to include differential bundle functioning (DBF) as a mechanism for identifying possible sources of differential functioning. This expansion is illustrated with an empirical (reading comprehension) data set consisting of a gender comparison and a socioeconomic status comparison with 1,000 participants in each group. Bundles on this particular test that exhibited large DBF

Requests for reprints should be sent to T. C. Oshima, Department of Educational Policy Studies, Georgia State University, University Plaza, Atlanta, GA 30303.

include a reading passage that favored boys over girls, possibly due to its context. The theoretical and practical implications of the use of DFIT for identifying possible sources of differential functioning are discussed.

Raju, van der Linden, and Fleer (1992, 1995) introduced a framework for differential functioning of items and tests (DFIT) in which differential item functioning (DIF) and differential test functioning (DTF) are assessed using item response theory (IRT). The DFIT framework offers a general DIF/DTF approach to various data types (unidimensional and multidimensional data with either dichotomous or polytomous scoring). The performance of the DFIT-based DIF/DTF indexes has been empirically evaluated, and promising results have been reported (e.g., Fleer, Raju, & van der Linden, 1995; Oshima, Raju, & Flowers, 1997).

Although most of the existing procedures for detecting differential functioning are defined at the item level (for a thorough review of various DIF indexes, see Millsap & Everson, 1993), several researchers have proposed investigating differential functioning beyond the item level (Longford, Holland, & Thayer, 1993; Shealy & Stout, 1993). A widely accepted concept of DTF by Shealy and Stout is based on nonparametric modeling, whereas DFIT presumes parametric modeling; that is, item and ability parameters are estimated by a calibration program such as BILOG3 (Mislevy & Bock, 1990) prior to the DTF/DIF analysis.

Douglas, Roussos, and Stout (1996) introduced a concept of differential bundle functioning (DBF) in which bundles of items are examined for DIF simultaneously. Using SIBTEST (Shealy & Stout, 1993), they demonstrated that DBF analysis can be used to identify potential sources or causes of DIF. Their examples included an analysis of gender DIF in which bundles of items were created (by content experts) that might favor one group over the other. It should be noted that a bundle of items differs from a testlet, in which adjacent items are bundled together for some organizational purpose. Differential testlet functioning was proposed by Wainer and his colleagues (e.g., Wainer, Sireci, & Thissen, 1991) and its various applications have been reported. In their approach, each testlet was considered to be a polytomous item, and a polytomous DIF analysis using a likelihood ratio test was conducted. Differential testlet functioning via SIBTEST has been recently introduced (Chang, Mazzeo, & Roussos, 1996). Both DBF and differential testlet functioning can be assessed within the DFIT framework.

In test development, tests are typically built using a table of specifications by which some type of cognitive dimensions (e.g., Bloom's [1956] taxonomy) are used to classify items into several levels of response (e.g., knowledge, comprehension, application, analysis, synthesis, evaluation). Usually, however, the test score is considered to reflect the content knowledge (sometimes referred to as the *primary dimension*) but not the levels of response. Levels of response may be thought of as constituting a possible secondary dimension, and it may be useful to bundle items

according to the levels of response and perform a DBF analysis on these bundles for the focal and reference groups of interest. Other possible secondary dimensions include speededness and item formats. It has been shown with simulated data that when a test was speeded, items toward the end exhibited a greater amount of DBF using SIBTEST (Oshima, 1994).

The purpose of this study was to describe how DBF analysis can be conducted within the DFIT framework and to demonstrate the applicability of DBF analysis using real data in an attempt to identify sources of differential functioning. Bundles were created by using the cognitive dimensions or instructional objectives identified in the table of specifications as well as by using reading passages.

## THE DFIT FRAMEWORK

Within unidimensional dichotomous IRT, an examinee's true score on a $k$-item test can be expressed as

$$T_s = \sum_{i=1}^{k} P_i(\theta_s), \tag{1}$$

where $P_i(\theta_s)$ is the probability of success on item $i$ for examinee $s$ with ability $\theta_s$. In the DFIT framework, each examinee in the focal group has a true score ($T_{sF}$) and an additional true score obtained as if he or she were in the reference group ($T_{sR}$). If $T_{sR}$ and $T_{sF}$ are equal for an examinee, then the examinee's true score is independent of group membership. The greater the difference between $T_{sR}$ and $T_{sF}$, the greater the differential functioning of a test. A measure of DTF at the examinee level may be defined as $(T_{sF} - T_{sR})^2$. Therefore, an overall measure of DTF across examinees may be defined as

$$DTF = \underset{F}{\varepsilon}(DTF_s) = \underset{F}{\varepsilon}(T_{sF} - T_{sR})^2, \tag{2}$$

where the expectation ($\varepsilon$) is taken over the focal group. This definition of DTF is similar to but somewhat different from DTF as defined by Shealy and Stout (1993). Interested readers are referred to Equations 8 and 9 in Shealy and Stout.

Raju et al. (1995) defined two types of differential functioning at the item level: compensatory DIF (CDIF) and noncompensatory DIF (NCDIF). The sum of the CDIF indexes across all items on the test is equal to DTF. The NCDIF index is

considered to be a special case of the CDIF index in which the assumption is made that all items except the studied item are DIF-free. Significance tests for DTF, CDIF, and NCDIF indexes can be performed. According to Raju et al., a chi-square test with $N_F - 1$ degrees of freedom (*df*), where $N_F$ is the number of examinees in the focal group, can be used for testing the significance of the observed DTF. Similarly, NCDIF is tested by a chi-square test with $N_F - 1$ *df*. Fleer (1993) found that a chi-square test for the NCDIF index was overly sensitive for large sample sizes. Based on a Monte Carlo study, he thus proposed an alternative cutoff value of .006 for declaring significant NCDIF. Significant CDIF items are those removed from the test to achieve nonsignificant DTF.

## DBF

Within the DFIT framework, DBF analysis starts by calibrating all items in a given test with an appropriate IRT model for two groups of interest. After item parameters are put on a common scale, items are then classified into appropriately defined bundles (e.g., classifying items into different skill categories, into categories favoring different race or gender groups, or into Bloom's levels of response). Once the bundles are created, two types of DBF can be defined (bundle CDIF and bundle NCDIF) in the DFIT framework. A summary of CDIF, NCDIF, bundle CDIF, bundle NCDIF, and DTF is displayed in Table 1. Note that the DTF value defined in the context of CDIF is equivalent to the DTF value defined in the context of NCDIF.

### Bundle CDIF

Bundle CDIF is an extension of CDIF. The CDIF values can be simply added for items in each bundle to obtain a value of bundle CDIF. Due to the additive nature of CDIF, the sum of bundle CDIF values is the DTF value for the total test. This approach will be useful for test developers whose task is to inspect the impact of removing a certain bundle (e.g., a passage in reading test) on DTF.

### Bundle NCDIF

Bundle NCDIF is an extension of NCDIF. Unlike CDIF, a value of bundle NCDIF is not the sum of NCDIF values. The DFIT analysis is carried out separately for each bundle, resulting in a DTF value for each bundle. Whereas with bundle CDIF other bundles are taken into account in calculating DTF for the studied bundle, this is not the case with bundle NCDIF. Another point of interest is that estimates of ability parameters ($\theta_s$) are based on all items in the test and are only obtained once. Therefore, what is investigated in this approach is whether a certain bundle of items is responded to differently by two groups of interest matched by ability as measured

TABLE 1
Summary of CDIF, NCDIF, Bundle CDIF, Bundle NCDIF, and DTF

| *CDIF* | *NCDIF* |
|---|---|
| Does NOT assume that all items but the studied item are DIF free. | Assumes that all items but the studied item are DIF free. |
| Additive | Nonadditive |
| Compensatory (i.e., positive CDIF and negative CDIF can cancel each other when summed.) | Noncompensatory |
| *Bundle CDIF* | *Bundle NCDIF* |
| $\sum_{i=1}^{n} CDIF,$ | $\left[\frac{1}{N_F}\sum_{s=1}^{N_F}\left(\sum_{i=1}^{n} P_{iF}(\theta_s) - \sum_{i=1}^{n} P_{iR}(\theta_s)\right)^2\right] / n,$ |
| where $n$ is the number of items in the bundle. | where $n$ is the number of items in the bundle and $N_F$ is the number of examinees in the focal group. |
| Can address the impact of removing a certain bundle. | Can be used to investigate the possible sources of differential functioning. |
| *DTF* | |
| $\sum_{i=1}^{N_I} CDIF,$ | $\frac{1}{N_F}\sum_{s=1}^{N_F}\left(\sum_{i=1}^{N_I} P_{iF}(\theta_s) - \sum_{i=1}^{N_I} P_{iR}(\theta_s)\right)^2$ |
| where $N_I$ is the total number of items. | where $N_I$ is the total number of items and $N_F$ is the number of examinees in the focal group. |

*Note.* CDIF = compensatory differential item functioning; NCDIF = noncompensatory differential item functioning; DTF = differential testing functioning.

by the entire test. When each bundle contains a different number of items, the direct comparison of DTF across different bundles is not recommended because DTF is a function of the number of items (i.e., the larger the number of items, the larger the DTF). In this case, DTF can be divided by the number of items in the bundle. This average DTF for the bundle will serve as bundle NCDIF. It is hoped that bundle NCDIF will lead to a possible identification of sources of differential functioning.

## METHOD

The achievement test used in this study is the Grade 4 reading comprehension test from the Metropolitan Achievement Tests, Elementary 2, Form S (Psychological Corporation, 1993). It consists of 55 multiple-choice items (four options) that are dichotomously scored. The same cognitive classifications and objectives used in

TABLE 2
Table of Specification

| *Objectives* | *Knowledge/Recognition* | *Understanding* | *Thinking Skills* |
|---|---|---|---|
| Mode of comprehension | | | |
| Initial understanding | | | |
| Demonstrate the ability to construct meaning with specific details and relations in reading selections of various types, length, and levels of difficulty. | | | |
| Specific detail | 2, 5, 7, 10, 21, 30, 39, | | |
| Understand explicit details. | 44, 46, 47, 53 | | |
| Action/reason/sequence | 8, 13, 14, 17, 22, 26, 33, | | |
| Understand explicitly stated actions, reasons, and sequences. | 37, 42, 45, 48 | | |
| Interpretation | | | |
| Demonstrate the ability to construct meaning by inferring and interpreting the meaning of ideas, events, and relations in reading selections of various types, lengths, and levels of difficulty. | | | |
| Inference/drawing conclusions | | 1, 3, 4, 11, 12, | 6, 36, 51, 52 |
| Infer implicit ideas and draw conclusions. | | 25, 28, 31 | |
| Extending meaning | | 9, 16, 24, 32, | 50 |
| Understand and interpret implicit ideas, events, and relations. | | 35, 41 | |
| Critical analysis | | | 15, 18, 34, 38, |
| Demonstrate the ability to synthesize and evaluate explicit and implicit information in a variety of reading selections. | | | 43 |
| Metacognition | | | |
| Demonstrate the ability to determine and use text factors and reader strategies with reading selections of various types, lengths, and levels of difficulty. | | | 19, 20, 23, 27, 29, 40, 49, 54, 55 |

Note. From Balow, Farr, & Hogan, 1993. Metropolitan Achievement Tests: Seventh Edition. Copyright © by Harcourt Brace & Company. Reproduced by permisssion. All rights reserved.

the development of the reading comprehension test were also used in forming item bundles. The test publisher's table of specifications showing these cognitive classifications and objectives is shown in Table 2 (Balow, Farr, & Hogan, 1993, pp. 40–41).

The reader should be reminded that the purpose of this study is to demonstrate how to conduct a DBF analysis using the DFIT framework. Therefore, the following descriptions of creating bundles and of choosing focal and reference groups of

interest are simply examples. There are, of course, many other ways to create bundles or contrasting groups.

Three different sets of bundles were created for this study. The first set was based on the cognitive classifications (i.e., the level of response) as indicated in Table 2. The three types of bundles were knowledge (22 items), understanding (14 items), and thinking (19 items). The second set was based on instructional objectives, also indicated in Table 2. The six objectives were specific detail (11 items); action, reason, and sequence (11 items); inference and drawing conclusions (12 items); extending meaning (7 items); critical analysis (5 items); and metacognition (9 items). The last set of bundles was based on the reading passages. There were 10 reading passages in this test, each containing four to seven questions.

Two sets of contrasting groups were used. One was boys versus girls and another was high socioeconomic status (SES) versus low SES. From the 5,945 fourth-grade students included in the research database, random samples of 1,000 boys (reference group), 1,000 girls (focal group), 1,000 high-SES students (reference group) and 1,000 low-SES students (focal group) were selected. SES was defined by three levels in these test data. High SES was sampled from the highest third (Level 3) and low SES was sampled from the lowest third (Level 1). Means and standard deviations for the performance of each group on this 55-item test are shown in Table 3.

The dichotomously scored items were calibrated separately for each group with BILOG3 (Mislevy & Bock, 1990) using the three-parameter logistic IRT model. Then, IPLINK (Lee & Oshima, 1996) was used to put item parameters from both groups on a common scale. The linking method used was a modified version of the test characteristic curve (TCC) method. In traditional TCC linking (Stocking & Lord, 1983), the difference between two TCCs is minimized over a range of ability values, $\theta_j$. For example, $\theta_j$ can be the $j$ (say, 11 or 21) equally spaced values between –4 and 4. The modified TCC method minimizes the difference between the two TCCs over all the theta points for the examinees in the focal group (say, 1,000 thetas if $N_F$ = 1,000), which is consistent with the definition of DTF shown in Equation 2.

A two-stage linking procedure was used. In the first stage, after the item parameters were placed on a common scale using all the items on the test, the DFIT program (Raju, 1995) identified items with large DIF (i.e., NCDIF > .006). Then, these items were deleted from the linking in the second stage. The linking coeffi-

TABLE 3
Descriptive Statistics for the 55-Item Test by Group

| *Group* | *N* | *M* | *SD* |
|---|---|---|---|
| Boys | 1,000 | 35.40 | 12.46 |
| Girls | 1,000 | 38.51 | 11.09 |
| High SES | 1,000 | 39.72 | 10.72 |
| Low SES | 1,000 | 35.02 | 12.24 |

*Note.* SES = socioeconomic status.

cients obtained from the second stage were used for the subsequent DFIT and DBF analyses. Note that, in accordance with the definition of DTF, the item parameters from the reference group were put on the scale of the item parameters from the focal group in this study.

Finally, the DFIT program was used to calculate DBF indexes. For bundle CDIF, a standard DIF/DTF program was run once and each value for bundle CDIF was obtained by simply summing the item-level CDIF values for the items in each bundle. For bundle NC–DIF, a standard DIF/DTF program was run for each bundle as if the bundle was the whole test. Although this process had to be repeated as many times as the number of bundles, the DFIT input file made it easy to choose only certain items to be submitted to the DIF/DTF program.

## RESULTS

First, the entire test was submitted to the DFIT program. For the gender comparison, the two items removed to achieve nonsignificant DTF (i.e., significant CDIF items) were Items 22 and 25. Those are also the items identified as having significant DIF by NCDIF using the .006 criterion. For the SES comparison, no item was identified as having significant DIF by either CDIF or NCDIF.

Then, CDIF values were summed for each bundle to obtain a value for bundle CDIF. Tables 4 and 5 show bundle CDIF for three different categories (cognitive classifications, objectives, and passages) for the gender comparison and the SES comparison, respectively. Because the sum of the bundle CDIF values equals the total DTF value, the contribution of each bundle to the total DTF as well as the cancellation effect, can be inspected in these tables. In this sense, CDIF and bundle CDIF are similar to SIBTEST, which detects the phenomena of simultaneous DIF amplification and cancellation (Nandakumar, 1993).

Next, each bundle was submitted to the DFIT program to obtain bundle NCDIF. Figures 1 and 2 show bundle NCDIF graphically for boys versus girls by cognitive classifications, objectives, and passages. In each figure, "All Items" indicates the entire test. Several observations can be made from these figures. First, there appears to be more differential functioning due to passages than due to cognitive classifications or objectives for this particular sample of fourth graders. The most obvious DBF is associated with Passage 5. Passage 5 is titled "The Roadrunner: A Strange Bird," in which the roadrunner was described, including what it eats (snakes, etc.) and how fast (18 miles per hr) it runs. Examining the mean difference of item probabilities over examinees revealed that this bundle favored boys. It is also noteworthy that the two NCDIF items (Items 22 and 25) identified on the entire test were in Passage 5.

Understanding showed the highest bundle NCDIF value among the three levels of cognitive classifications, and inference and drawing conclusions and critical

TABLE 4
Bundle CDIF for Each Bundle (Boys vs. Girls)

| *Bundle* | *Item* | *ΣCDIF* |
|---|---|---|
| (a) Cognitive classifications | | |
| Knowledge/recognition | 2, 5, 7, 8, 10, 13, 14, 17, 21, 22, 26, 30, 33, 37, 39, 42, 44, 45, 46, 47, 48, 53 | −.005 |
| Understanding | 1, 3, 4, 9, 11, 12, 16, 24, 25, 28, 31, 32, 35, 41 | .039 |
| Thinking skills | 6, 15, 18, 19, 20, 23, 27, 29, 34, 36, 38, 40, 43, 49, 50, 51, 52, 54, 55 | .014 |
| | Total (DTF) | .048 |
| (b) Objectives | | |
| Specific details | 2, 5, 7, 10, 21, 30, 39, 44, 46, 47, 53 | −.027 |
| Action, reason, and sequences | 8, 13, 14, 17, 22, 26, 33, 37, 42, 45, 48 | .022 |
| Inference and drawing conclusions | 1, 3, 4, 6, 11, 12, 25, 28, 31, 36, 51, 52 | .036 |
| Extending meaning | 9, 16, 24, 32, 35, 41, 50 | .028 |
| Critical analysis | 15, 18, 34, 38, 43 | −.025 |
| Metacognition | 19, 20, 23, 27, 29, 40, 49, 54, 55 | .014 |
| | Total (DTF) | .048 |
| (c) Passages | | |
| Passage 1 | 1–5 | −.006 |
| Passage 2 | 6–9 | .018 |
| Passage 3 | 10–13 | .008 |
| Passage 4 | 14–19 | −.031 |
| Passage 5 | 20–25 | .070 |
| Passage 6 | 26–32 | −.035 |
| Passage 7 | 33–37 | .008 |
| Passage 8 | 38–44 | −.028 |
| Passage 9 | 45–50 | .016 |
| Passage 10 | 51–55 | .028 |
| | Total (DTF) | .048 |

Note. CDIF = compensatory differential item functioning; DTF = differential test functioning.

analysis showed the two highest bundle NCDIF values among the six levels of objectives. However, differences of the magnitudes of bundle NCDIF among these bundles may not be of practical significance.

Figures 3 and 4 display the same analyses previously described, but this time applied to the SES comparison. Again, there were several interesting trends. In contrast to the gender comparison, some of the DBF due to cognitive classifications or objectives appeared large for the SES comparison. The direct comparison of Figures 1 and 3 as well as Figures 2 and 4 should be avoided because all DBF analyses are affected by how well the two scales are put on a common scale (i.e., linking). For the current data set, the fit of the TCCs from the two groups was better for the gender comparison than for the SES comparison.

There was no particularly large bundle NCDIF value for passages in the SES comparison. The largest value of bundle NCDIF for the SES bundles was Passage 2. Passage 2 is titled "Are You Eating Healthy Foods?" and describes how eating fat is not necessarily a bad thing if one chooses to eat fats from fruits, vegetables, and nuts as opposed to fats from animal foods. For both the gender and the SES comparisons, the most neutral passage was Passage 7. Passage 7 describes the ice merchant in horse buggies who sold ice door to door in early times before the invention of the refrigerator. Perhaps Passage 7 is neutral because it is distant from all fourth graders' daily lives.

TABLE 5
Bundle CDIF for Each Bundle (High SES vs. Low SES)

| *Bundle* | *Item* | *ΣCDIF* |
|---|---|---|
| (a) Cognitive classifications | | |
| Knowledge/recognition | 2, 5, 7, 8, 10, 13, 14, 17, 21, 22, 26, 30, 33, 37, 39, 42, 44, 45, 46, 47, 48, 53 | .140 |
| Understanding | 1, 3, 4, 9, 11, 12, 16, 24, 25, 28, 31, 32, 35, 41 | .023 |
| Thinking skills | 6, 15, 18, 19, 20, 23, 27, 29, 34, 36, 38, 40, 43, 49, 50, 51, 52, 54, 55 | .086 |
| | Total (DTF) | .249 |
| (b) Objectives | | |
| Specific details | 2, 5, 7, 10, 21, 30, 39, 44, 46, 47, 53 | .089 |
| Action, reason, and sequences | 8, 13, 14, 17, 22, 26, 33, 37, 42, 45, 48 | .051 |
| Inference and drawing conclusions | 1, 3, 4, 6, 11, 12, 25, 28, 31, 36, 51, 52 | .029 |
| Extending meaning | 9, 16, 24, 32, 35, 41, 50 | .009 |
| Critical analysis | 15, 18, 34, 38, 43 | .023 |
| Metacognition | 19, 20, 23, 27, 29, 40, 49, 54, 55 | .048 |
| | Total (DTF) | .249 |
| (c) Passages | | |
| Passage 1 | 1–5 | .023 |
| Passage 2 | 6–9 | .007 |
| Passage 3 | 10–13 | .016 |
| Passage 4 | 14–19 | .010 |
| Passage 5 | 20–25 | .018 |
| Passage 6 | 26–32 | .043 |
| Passage 7 | 33–37 | .010 |
| Passage 8 | 38–44 | .043 |
| Passage 9 | 45–50 | .046 |
| Passage 10 | 51–55 | .033 |
| | Total (DTF) | .249 |

Note. SES = socioeconomic status; CDIF = compensatory differential item functioning; DTF = differential test functioning.

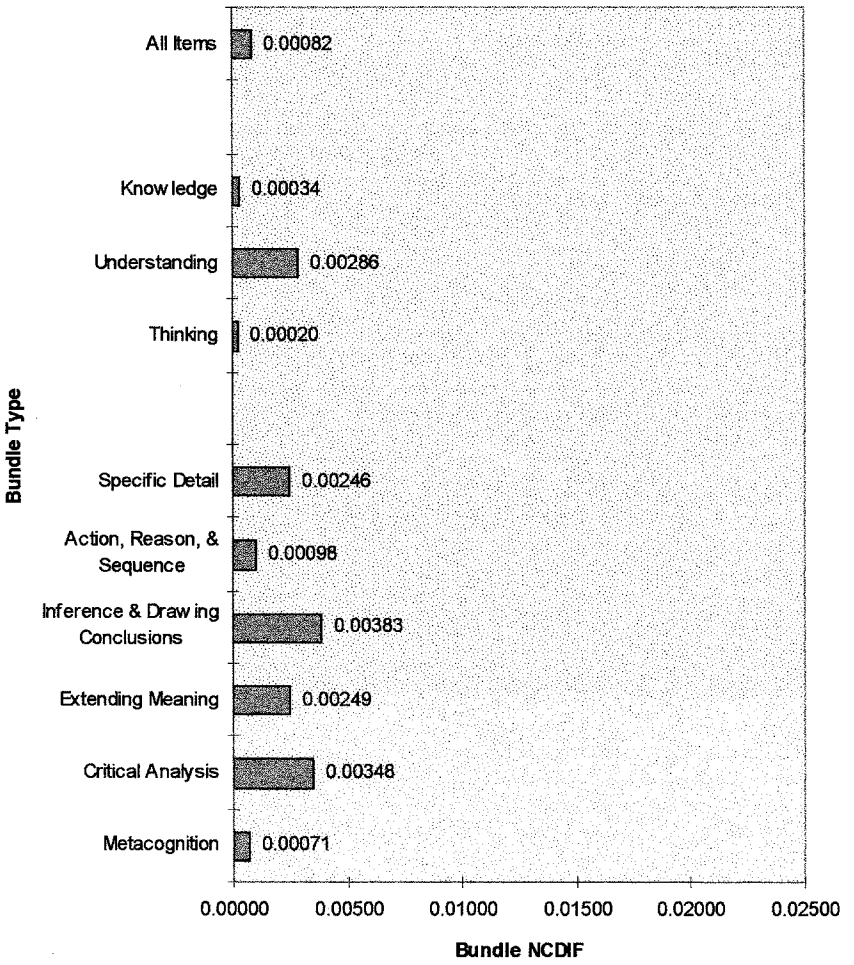


FIGURE 1 Differential bundle functioning (DBF; bundle noncompensatory differential item functioning; NCDIF) for boys and girls by cognitive classifications and objectives.

The pattern of bundle NCDIF for cognitive classifications and objectives for the SES comparison also differed from those for the gender comparison. This time, knowledge showed the highest bundle NCDIF of the cognitive classifications. Specific detail was the highest in the SES comparison among the six levels of objectives. Discussing the reasons for these results is beyond the scope of this article. What is of interest here is that DBF can provide a very different picture when contrasting groups are defined differently.

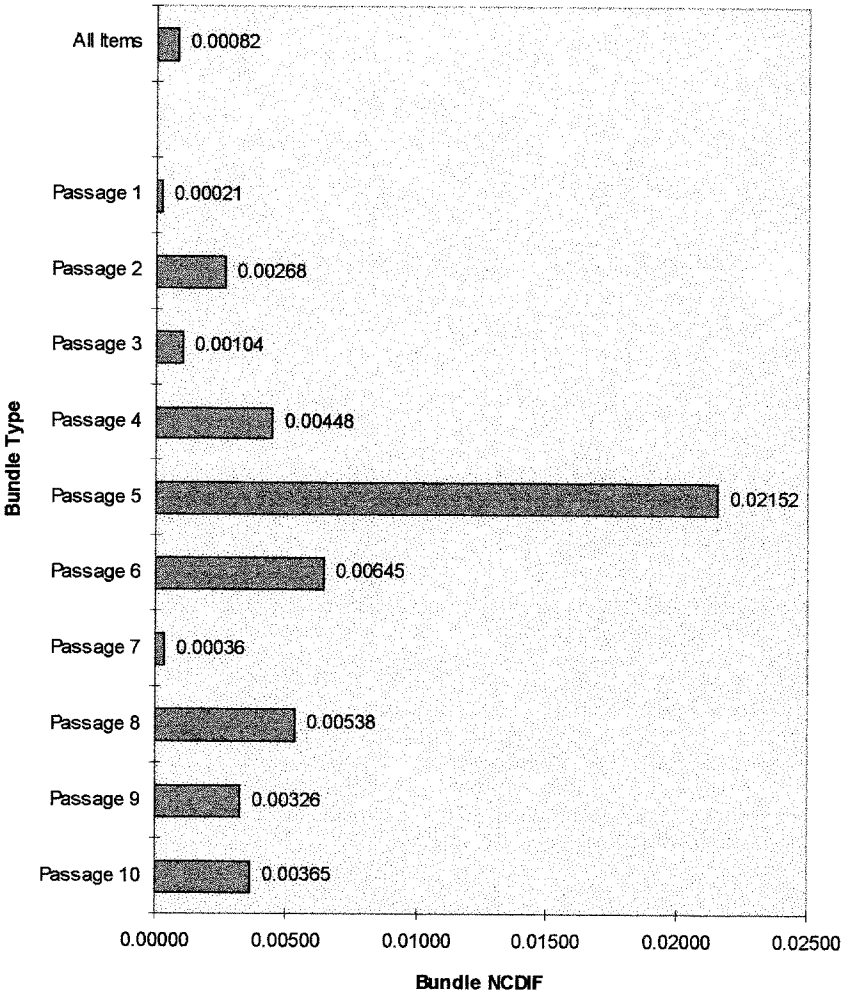


FIGURE 2 Differential bundle functioning (DBF; bundle noncompensatory differential item functioning; NCDIF) for boys versus girls by passages.

An additional point of concern is the stability of DBF across different samples. When there are enough examinees, replicating the analysis (cross-validation) will give researchers more confidence in declaring DBF. To illustrate this process, a cross-validation of the data described previously was conducted for the boys versus girls comparison. In the first (i.e., original) set of samples, the most obvious DBF was exhibited by Passage 5. Therefore, test data for other independent random samples of 1,000 boys and 1,000 girls were examined for DBF using passages as

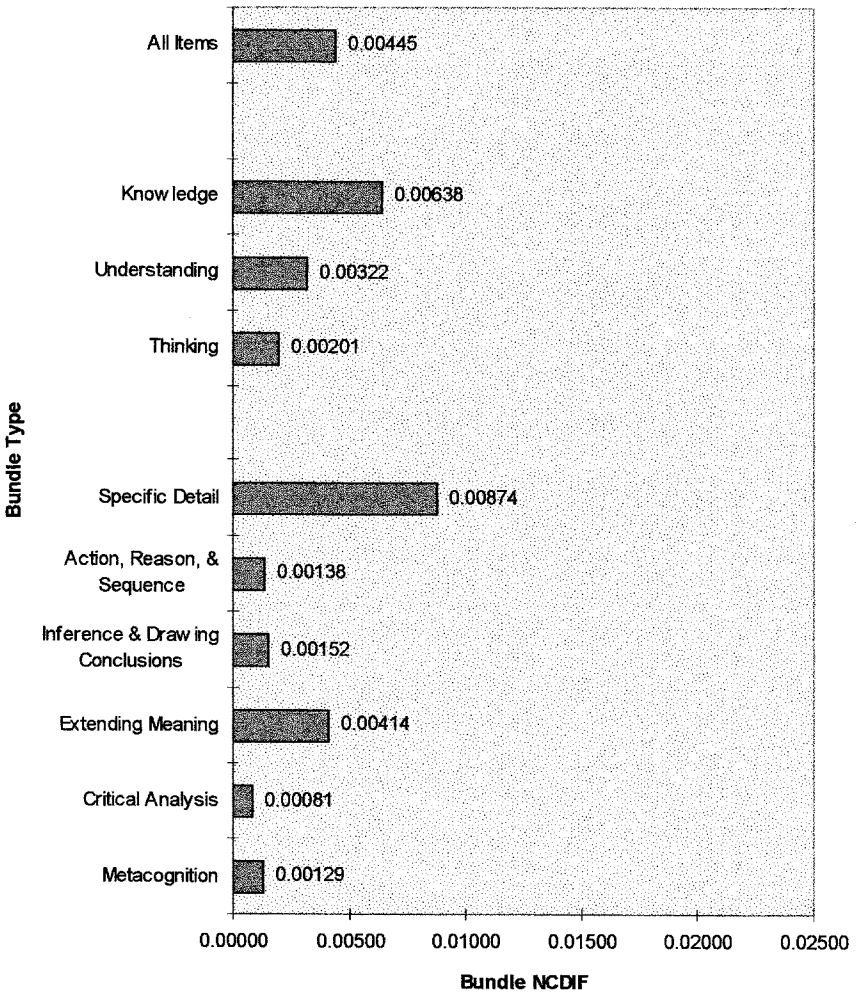


FIGURE 3 Differential bundle functioning (DBF; bundle noncompensatory differential item functioning; NCDIF) for high socioeconomic status (SES) versus low SES by cognitive classifications and objectives.

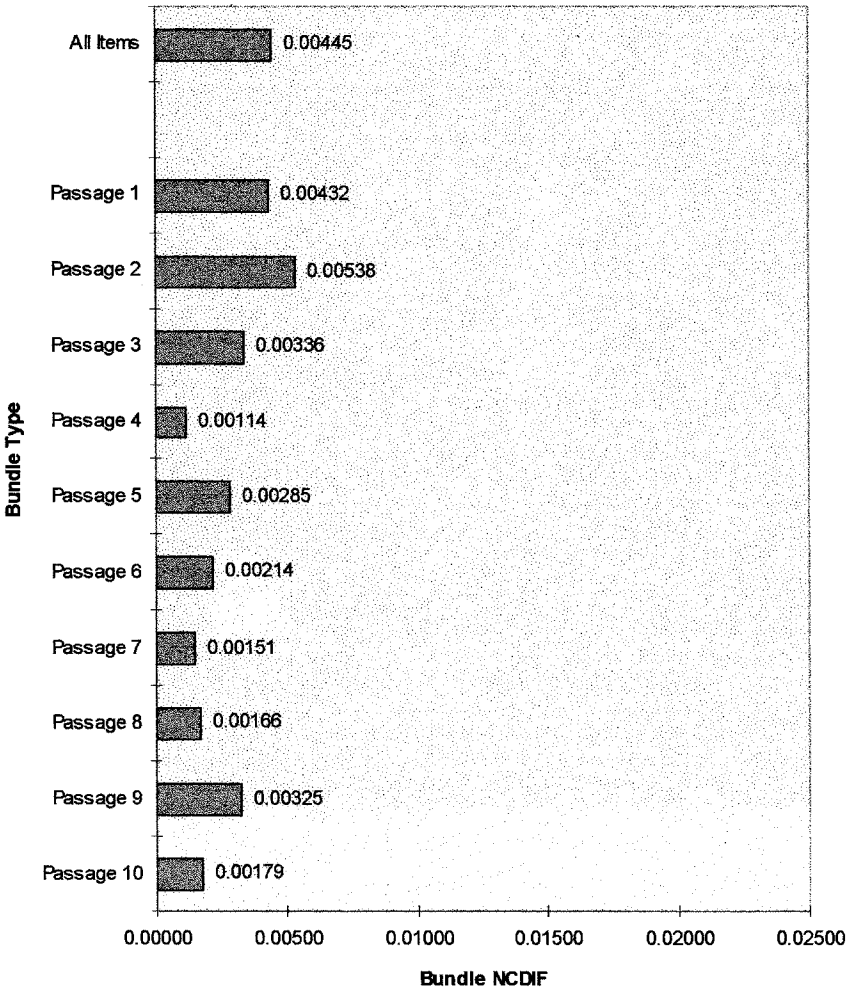


FIGURE 4 Differential bundle functioning (DBF; bundle noncompensatory differential item functioning; NCDIF) for high socioeconomic status (SES) and low SES by passages.

the criterion for the bundles. Passage 5 again had a substantially larger value of DBF (bundle NCDIF = .027408) than did the remaining passages (bundle NCDIF ranged from .00018 to .01046). The order of the magnitudes of DBF for the remaining nine passages changed somewhat, suggesting that small differences in the magnitudes of DBF in a given set of samples should be interpreted with caution.

## DISCUSSION

As stated earlier, the purpose of this article was not to evaluate the particular test, but to show the technique for DBF analysis. Therefore, we refrain from any speculation as to potential causes of differential functioning on this particular test for this fourth-grade sample. Instead, this article demonstrated a possible tool for examining causes of differential functioning and the effect of removing bundles that exhibit DBF from the test.

In the DFIT framework, DBF can be investigated using bundle CDIF and bundle NCDIF. It is useful to have both types of DBF. Consider the DBF identified in the data examined in this study. Passage 5 showed the highest bundle NCDIF favoring boys over girls, suggesting that the type of context used in Passage 5 is a possible source of differential functioning. Should test developers avoid using the context of Passage 5? Not necessarily. Before making such a decision, they also need to look at bundle CDIF to examine the effect of the particular bundle on the total test. In our example, it happened that bundle CDIF for Passage 5 was also fairly large. However, if there had been another bundle that happened to favor girls over boys, the exact same bundle (Passage 5) could have had a small or even near-zero value for bundle CDIF. If test developers choose only passages that are least likely to cause differential functioning, such as Passage 7, the validity of the test may be in question because those passages can be artificial and may not particularly reflect the examinees' daily lives. The use of bundle CDIF along with bundle NCDIF can help test developers balance the test so that no particular group has an unfair advantage although the test as a whole reflects our diverse culture.

Bundle NCDIF can be used to examine certain characteristics of examinees. A cognitive psychologist may want to apply this technique to study how cognitive skills change over time. For example, research has shown that although boys and girls perform similarly in math during their grade-school years, certain gender differences tend to appear during adolescence. Such differences include boys scoring consistently higher than girls on math tests, with these differences being especially marked among gifted students (Benbow & Stanley, 1982). DBF examines differential performance when examinees are matched by ability. It would be interesting to examine whether adolescent boys and girls equally matched by ability would solve the problem differently in terms of cognitive skills.

DBF analysis is not limited to the traditional groupings often used for DIF analysis. In this article, the SES comparison was shown as an example. Comparisons of public schools versus private schools and rural schools versus suburban schools would be other possibilities. DBF makes it possible to examine whether there is differential functioning in certain cognitive skill areas between certain school systems. Finding potential sources of the differential functioning may help develop better curricula.

This article demonstrated the application of DBF only to the unidimensional dichotomous case. As noted earlier, in the DFIT framework, any type of test data, either multidimensional or unidimensional, whether scored polytomously or dichotomously, can be submitted to DBF analysis in a very similar way. The only difference would be the way in which the true scores or expected scores are calculated. The remaining process would be identical. Computer programs necessary for unidimensional dichotomous, unidimensional polytomous, and multidimensional dichotomous data are currently available. DBF analysis for questionnaires based on a Likert-type scale may be useful in the field of psychology.

There are still various technical areas to be investigated with regard to DBF analysis in the DFIT framework. First, the issue of linking is closely related to the performance of DFIT. As described earlier, a new linking program, IPLINK, is currently being tested to enhance DFIT performance in both unidimensional and multidimensional test data. Second, the significance test associated with DTF needs to be investigated further. Simulation studies examining the Type I error rate are also currently underway to find the appropriate level of the cutoff score for declaring significance.

In summary, this study introduced a possible tool to be used for examining causes of differential functioning and its impact on the total test. The emphasis is not only on the test (i.e., whether the test is valid or fair) but also on the examinees (i.e., how and why certain groups matched by ability answer items differently). Some possible applications of this technique were discussed. Further research studies were suggested both in the application area and in the technical area.

## REFERENCES

Balow, I. H., Farr, R. C., & Hogan, T. P. (1993). *Compendium of instructional objectives: Metropolitan Achievement Tests* (7th ed.). San Antonio, TX: Psychological Corporation.

Benbow, C., & Stanley, J. (1982). Intellectually talented boys and girls: Educational profiles. *Gifted Child Quarterly, 26,* 82–88.

Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook 1: The cognitive domain.* New York: McKay.

Chang, 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.

Douglas, J., Roussos, L., & Stout, W. (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential item functioning. *Journal of Educational Measurement, 33,* 465–484.

Fleer, P. F. (1993). *A Monte Carlo assessment of a new measure of item and test bias.* Unpublished doctoral dissertation, Illinois Institute of Technology, Chicago.

Fleer, P. F., Raju, N. S., & van der Linden, W. J. (1995, April). *A Monte Carlo assessment of DFIT with dichotomously scored unidimensional tests.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Lee, K., & Oshima, T. C. (1996). *IPLINK: Multidimensional and unidimensional item parameter linking in item response theory* [Computer program]. Atlanta: Georgia State University.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of MH D–DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171–196). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific Software.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy–Stout's test for DIF. *Journal of Educational Measurement, 30,* 293–311.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31,* 200–219.

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. *Journal of Educational Measurement, 34,* 253–272.

Psychological Corporation. (1993). *Metropolitan Achievement Tests* (7th ed.). San Antonio, TX: Author.

Raju, N. S. (1995). *DFITDU: A Fortran program for calculating DIF/DTF* [Computer program]. Atlanta: Georgia Institute of Technology.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1992, April). *An IRT-based internal measure of test bias with applications for differential item functioning.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning items and tests. *Applied Psychological Measurement, 19,* 353–368.

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

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning. *Journal of Educational Measurement, 28,* 197–220.