APPLIED MEASUREMENT IN EDUCATION, *20*(2), 205–225

Identifying Possible Sources of Differential Functioning Using Differential Bundle Functioning With Polytomously Scored Data

F. A. McCarty
Department of Behavioral Sciences and Health Education
Emory University

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

Nambury S. Raju
Illinois Institute of Technology

Oshima, Raju, Flowers, and Slinde (1998) described procedures for identifying sources of differential functioning for dichotomous data using differential bundle functioning (DBF) derived from the differential functioning of items and test (DFIT) framework (Raju, van der Linden, & Fleer, 1995). The purpose of this study was to extend the procedures for dichotomous DBF to the polytomous case and to illustrate how DBF analysis can be conducted with polytomous scoring, common to psychological and educational rating scales. The data set used was parent and teacher ratings of child problem behaviors. Three group contrasts (teacher vs. parent, boy vs. girl, and random groups) and two bundle organizing principles (subscale designation and random selection) were used for the DBF analysis. Interpretations of bundle indexes in the context of child problem behaviors were presented.

Recently researchers have attempted to address differential functioning within a more substantive framework by shifting the focus from an examination of individual items to an examination of groups of items. Although much of the

Correspondence should be addressed to F. A. McCarty, 1518 Clifton Road, Atlanta, GA 30322. E-mail: fmccart@sph.emory.edu

research involving differential item functioning (DIF) has been focused at the item level, some researchers have considered differential functioning at the test level (Longford, Holland, & Thayer,1993, Raju, van der Linder, & Fleer, 1995; Shealy & Stout, 1993a), as well as at the bundle level, where items are considered in an aggregated fashion (Douglas, Roussos, & Stout, 1996; Gierl, Bisanz, Bisanz, & Boughton, 2003; Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001; Gierl & Khaliq, 2001; Oshima, Raju, Flowers, & Slinde, 1998; Wainer, Sireci, & Thissen, 1991).

Douglas et al. (1996) introduced the concept of differential bundle functioning (DBF), in which item bundles were defined as clusters of items that had been chosen according to some organizational principle. In this case, the items in a bundle are not necessarily adjacent, nor do they have to refer to a common passage or text, as is the case with differential testlet functioning as proposed by Wainer et al. (1991). Their approach focused on identifying potential DIF bundles through the identification of dimensionally homogeneous item bundles. This approach was based on Shealy and Stout's (1993a) multidimensional model for bias and used SIBTEST (Shealy & Stout, 1993b) to assess DIF in bundles. Douglas et al. described two methods for identifying suspect bundles: (a) use of expert opinion alone and (b) use of a statistical dimensionality analysis along with expert opinion.

More recently, Gierl and colleagues (2003) expounded on the Roussos–Stout DIF analysis paradigm. They describe a two-stage approach where the first stage involves the generation of DIF hypotheses, with the DIF hypothesis specifying whether each item or bundle intended to measure the primary dimension also measures a secondary dimension. In the second stage, the DIF hypothesis is statistically tested. The authors point out that focusing on clusters of items based on some organizing principle to provide a priori structure to the data greatly enhances the interpretability of the results. They further suggest that an item serves as a poor level of analysis because a single item represents a small, somewhat unreliable sample of a behavior. Because bundles are likely to provide a broader sample of secondary dimensions, they should be more conducive to interpretation and lead to better explanations regarding the nature of group differences. In addition, this approach should reduce the Type I error rate because a smaller number of DIF hypotheses are tested (Gierl et al., 2003). In another study, by Gierl and Khaliq (2001), the previously described framework is used to examine differential item and bundle functioning on translated achievement tests.

Using a different approach, Oshima et al. (1998) described and demonstrated how differential bundle functioning could be examined within the differential functioning of items and tests (DFIT) framework. The DFIT framework, as proposed by Raju et al. (1995), is an item response theory (IRT)-based parametric procedure that can be used with unidimensional and multidimensional data that result from either dichotomous or polytomous scoring. This framework provides the only parametric, IRT-based measure of differential functioning at both the

test and item levels. When using the DFIT framework to conduct a DBF analysis, the first step is to calibrate all items on a test with an appropriate IRT model for the two groups of interest. In the Oshima et al. (1998) study, a 3-parameter logistic model was used and groups were based on gender or socioeconomic status. After obtaining the item parameters for each group, item parameters were equated or put on a common scale. Then, items were classified into substantively meaningful bundles (e.g., cognitive dimensions, instructional objectives, reading passage) and the DBF analysis was conducted within the DFIT framework.

Oshima et al. (1998), however, presented DBF only for the data with dichotomous scoring. Therefore, the purpose of this study is to extend their study to the unidimensional polytomous case. The focus is on describing and illustrating how DBF analysis can be conducted within the framework of DFIT when polytomous scoring has been employed, as is the case with many scales used in psychological and educational settings. In the next section, the potential usefulness of the DBF analysis with polytomous data will be presented, followed by a detailed description of polytomous DBF in the DFIT framework.

DBF WITH POLYTOMOUS DATA

Although much of the research on IRT and DIF has focused on tests designed to assess performance within a cognitive domain, there has been an increase in the application of this methodology to examine instruments that focus on assessing constructs that could be best described as coming from a more behavioral or psychological domain. Within this domain, instruments are more likely to be constructed with polytomous scoring and developed so they contain items that may be clustered meaningfully with respect to assessing a similar aspect of the particular trait under consideration. The notion of item clusters becomes readily apparent if one considers the frequent use of subscales when there is an interest in assessing psychological and behavioral variables. Given the structure of many tests within the psychological and behavioral domains, it would seem appropriate to examine potential differential functioning within a framework that actually accounts for the intended structure of the test (i.e., it may be more meaningful to examine groups of items as opposed to individual items). With this in mind, polytomous DBF analysis within the DFIT framework may provide useful psychometric information for use in the development and evaluation of tests within the psychological and behavioral domains.

Traditionally, DIF analyses have focused on sources of variation related to examinee characteristics such as gender, socioeconomic status, race/ethnicity, and culture. However, in a study by Maurer, Raju, and Collins (1998), the measurement equivalence between raters was examined using the DFIT framework. In this study, the rater type was used to form the subgroups subjected to the DIF

analysis. In the case of multiple raters, determining whether or not ratings differ—not because of differing skill, but rather because of how the scale was used in the different rater populations—has implications for interpretation and use of data obtained from multiple raters. Although the context of the Maurer et al. study was that of organizational settings and performance appraisal ratings, where multiple raters are often used, the idea of examining rater type as a source of DIF is equally relevant in educational and psychological settings. Instruments designed to assess psychological and behavioral constructs often are developed such that multiple raters are called on to provide information. For example, behavior rating scales are frequently used to assess social competence and both adaptive and maladaptive behavior in children. In this case, ratings may be obtained from a teacher, parent, or psychologist. Again, in order to interpret the scores appropriately it is important to know whether the scale has the same meaning across raters. Although this problem was addressed by the Maurer et al. study, it was not approached from a DBF perspective, but from the more basic DFIT perspective. Again, given the structure of many psychological and behavioral assessments, DBF analysis provides a framework that can encompass both the internal structure of an instrument, for example, subscales, as well as the source of variation that may be present when multiple informants are called on to provide information.

Another important issue related to psychological data with subscales is dimensionality. As in any unidimensional IRT research, the assumption of unidimensionality needs to be tenable. An instrument with (substantively defined) subscales may imply multidimensionality, to some degree. According to Tate (2002), two aspects of validity must be addressed with an instrument with subscales, the internal structure of the instrument and the discriminant validity of the subscores. The former relates to the unidimensionality assumption. Prior to the DBF analysis, a dimensionality analysis needs to be conducted to assure that the data are reasonably unidimensional. Then the DBF analysis can address the latter type of validity, as DIF (or DBF) is manifestation of multidimensionality of data.

It should be pointed out that the iterative linking procedure is an essential part of the IRT-based DBF analysis with a presence of prevailing multidimensionality. The purpose of the iterative linking is to "purify" the linking items from two separate IRT calibrations (reference and focal groups) so that the linking items are fairly unidimensional for the data with multidimensionality (i.e., DIF). This purifying method was first introduced in Lord (1980) and later simplified by Candell and Drasgow (1988). In the Candell and Drasgow method, the linking coefficients are obtained in multiple stages (or two stages) without reestimation of IRT parameters, each time eliminating the apparent DIF items for the purpose of calculating the linking coefficients. Studies have shown that iterative linking is especially useful when the number of DIF items is large (e.g., Miller & Oshima, 1992). Given the nature of the DBF analysis with a prevailing secondary dimension (i.e., subscales), the iterative linking should prove useful.

POLYTOMOUS DBF IN THE DFIT FRAMEWORK

The DFIT framework by Raju et al. (1995) provides three indexes of differential functioning: (a) differential test functioning (DTF), a measure reflecting the differential functioning of the entire test; (b) compensatory DIF (CDIF), an item-level index that is additive with respect to DTF; and (c) noncompensatory DIF (NCDIF), an item-level index that is a special case of CDIF in which the assumption is made that all items but the one under consideration are free of DIF.

DTF

According to Raju et al. (1995), DTF across examinees may be defined as

$$DTF = \varepsilon_F\{[T_{sF}(\theta_s) - T_{sR}(\theta_s)]^2\} = \varepsilon_F[D_s^2(\theta_s)], \tag{1}$$

where the expectation (ε) of the squared (test-level) true score [$T_s(\theta_s)$] difference [$D_s(\theta_s)$] is computed over the focal group. In the dichotomous case, the true score for each examinee *s* is the sum of the probability of answering an item correctly based on a dichotomous IRT model. In the polytomous case, the true score is expressed as:

$$T_s(\theta_s) = \sum_{i=1}^{n} ES_{si}(\theta_s), \tag{2}$$

where *n* is the number of items in the test. The expected score (*ES*) or true score for examinee *s* on item *i* [$ES_{si}(\theta_s)$] can be computed as

$$\mathrm{ES_{si}}(\theta_s) = \sum_{k=1}^{m} P_{ik}(\theta_s) X_{ik}, \tag{3}$$

where X_{ik} is the score for category *k*, *m* is the number of response categories, and P_{ik} is the probability of responding to category *k* based on a polytomous IRT model (Samejima, 1969).

The true score difference at the test level [$D_s(\theta_s)$] can be viewed as the sum of the expected score difference at the item level:

$$D_s(\theta_s) = \sum_{i=1}^{n} d_{si}, \tag{4}$$

where

$$d_{si}(\theta_s) = ES_{siF}(\theta_s) - ES_{siR}(\theta_s). \quad (5)$$

The d_{si} in the preceding equation is the difference in the item-level true scores for examinee *s* on item *i*.

CDIF

CDIF is defined as:

$$CDIF_i = \varepsilon_F(d_i D) = Cov(d_i, D) + \mu_{d_i}\mu_D, \quad (6)$$

where $Cov(d_i, D)$ is the covariance between d_i and D, μ_{d_i}, is the mean of d_t, and μ_D is the mean of D in the focal group. CDIF has an additive property; that is,

$$DTF = \sum_{i=1}^{n} CDIF_i. \quad (7)$$

NCDIF

NCDIF is expressed as:

$$NCDIF_i = \varepsilon_F[ES_{siF}(\theta_s) - ES_{siR}(\theta_s)]^2 = \varepsilon_F d_{si}^2(\theta_s) = \sigma_{d_i}^2 + \mu_{d_i}^2. \quad (8)$$

The NCDIF index is simply the expectation (taken over the focal group) of the squared difference in true scores at the item level. A comparison of Equations 1 and 8 shows that the definitions of DTF and NCDIF are identical, except that DTF is concerned with the squared true score differences at the test level, whereas NCDIF is concerned with the squared true score differences at the item level.

Within the DFIT framework, statistical significance tests can be performed on the DTF and NCDIF indexes. However, because the χ^2 tests for the DTF and NCDIF indexes are known to be sensitive to sample size, Raju et al. (1995) suggested additional criteria for assessing the significance of the observed DTF and NCDIF indexes. The current recommendation for determining whether or not an NCDIF index is significant when an item is graded on a 3-point scale is an NCDIF value greater than .024 (Raju, 1999). This cutoff value is based on an

extension of the cutoff value of .006 for the dichotomous case, proposed by Fleer (1993). These cutoff values are considered to be rough guidelines for practitioners. Recently, a new significance test was proposed and tested for dichotomous DFIT (Oshima, Raju, & Nanda, 2006). The new method, based on the item parameter replication (IPR) algorithm, derives the cutoff values for each item by simulating a large number of DFIT indexes under the no-DIF condition. The new method was reported to work well for the dichotomous case. This new significance test has not yet been fully developed and evaluated for the polytmous case; however, the test as described in a recent paper by Raju, Oshima, Fortmann, Nering, and Kim (2006) was applied in this study. For the full descriptions of the DFIT procedures for the dichotomous and polytomous data, interested readers are referred to the original DFIT articles (Flowers, Oshima, & Raju, 1999; Raju, et al., 1995).

DBF

Oshima et al. (1998) described DBF analysis for dichotomously scored items. This description is equally valid for the polytomous case, as shown in the Appendix. DBF within the DFIT framework begins by calibrating all items with an appropriate IRT model for the two groups of interest and then placing the item parameters on a common scale. Using the predefined bundle classification scheme, items are placed into different bundles. Specifically, let there be v mutually exclusive bundles ($B_1, \ldots B_j, \ldots B_v$) with n_j items in bundle j and $n_1 + \ldots + n_j + \ldots + n_v = n$.

As shown in Equation A8 in the Appendix, CDBF (compensatory differential bundle functioning) is an extension of CDIF where CDIF values are added for items in each bundle to obtain CDBF. Therefore, $CBDF_j$ for bundle j may be expressed as:

$$CDBF_j = \sum_{i \varepsilon B_j} CDIF_i. \tag{9}$$

It should be noted that CDIF is the expectation (average) of the product of the true score difference at the item level and the true score difference at the test level (Equation 6). Because the true score difference at the bundle level is simply the sum of true score difference at the item level, CDBF becomes the sum of CDIF indexes of all items that make up the bundle. Because the sum of CDIFs is equal to DTF (Equation 7), the sum of CDBFs, in view of Equation 9, is also equal to DTF. CDBFs can be used to examine the impact (on DTF) of removing a certain bundle from the test (Oshima et al., 1998).

As shown in the Appendix, NCDBF (noncompensatory differential bundle functioning index) is not simply the sum of NCDIF values. Instead, the DFIT

analysis needs to be carried out separately for each bundle, resulting in a value of DTF for each bundle, which then becomes the NCDBF for that bundle. That is,

$$NCDBF_j = DTF_j, \tag{10}$$

where DTF_j refers to the DTF based on all items in bundle j. According to Equation 8, NCDIF is the expectation (average) of the squared difference in true score at the item level. Similarly, according to Equation A11 in the Appendix, NCDBF is also the expectation of the squared difference in true scores at the bundle level. Even though the sum of item-level true score differences is equal to the true score difference at the bundle level, the sum of squared item-level true score differences will not in general be equal to the squared true score difference at the bundle level. Therefore, as previously noted, NCDBF for a given bundle is not simply a sum of the NCDIF indexes of all items that make up the bundle. In computing the different NCDBF indexes, it should be noted that estimates of ability parameters (θ_s) are based on all items on the test and are obtained only once. This approach allows one to investigate whether a particular bundle of items is responded to differently by two groups matched by ability as measured by the entire test. When bundles contain different numbers of items, it is not recommended that the different NCDBF indexes across bundles be directly compared because the observed size of an NCDBF index typically varies as a function of the number items in the bundle (i.e., other things being equal, the more items, the larger the NCDBF index). In view of this, each $NCDBF_j$ may be divided by the number of items in the bundle, resulting in an average $NCDBF_j$ for the bundle. That is,

$$\overline{NCDBF}_j = \frac{NCDBF_j}{n_j}. \tag{11}$$

This average index may be used to compare bundles for differential functioning. Please refer to the Appendix for additional information about the CDBF and NCDBF indexes.

As for interpreting the size of the DBF indexes, the current polytomous DFIT/DBF framework has a recently developed significance test that can be applied. Because the test has not undergone an extensive evaluation, we recommend the use of the baseline method, where the baseline data are created by random bundles from randomly divided groups in addition to the significance tests. The NCDIF values produced by those baseline data can be descriptively compared to those produced by other meaningful comparisons. For this study, the DBF analysis can be viewed as a kind of profile analysis and descriptive presentations including graphs should serve as a valuable tool for visual inspection. This

approach should remain useful, even after an appropriate significance test becomes readily available, because its descriptive nature would be helpful in viewing the magnitude of DBF as well as determining the practical significance.

METHOD

Data Source

The rating scale data used in this study were obtained as part of a National Head Start/Public School Project. In this project, children were followed from the beginning of their kindergarten year to the end of their third-grade year. With the exception of the initial assessments, data were collected in the spring of the school year. The data used to demonstrate this methodology consist of parent and teacher ratings for approximately 1,626 second-grade children.

The rating scale used in this study was the Social Skills Rating System (SSRS; Gresham & Elliott, 1990). The SSRS consists of two distinct questionnaires, the social skills (SS) questionnaire and the problem behaviors (PB) questionnaire. There are two forms that have been developed for use with grades K–6, a teacher form and a parent form. Both forms ask the rater to rate the child's present behavior using descriptors that define how often a particular behavior occurs, using the following categories: never (0), sometimes (1), and very often (2). Although the DBF analysis was conducted on both scales (SS and PB), only the results from the problem behavior questionnaire were reported here as an illustration. The problem behaviors questionnaire consists of 18 items on the teacher form and 17 on the parent form. Three subscales, internalizing behaviors, externalizing behaviors, and hyperactive behaviors, have been identified for the problem behaviors questionnaire. For the purposes of this study, only the common items (16), which appeared on both the parent and teacher forms of the instrument were used.

Item Bundles

The item bundles were created using the items that were conceptually related and that have been defined by the author of the instrument as constituting subscales according to the teacher version of the instrument. In addition, three random bundles were created. The number of items in the random bundles was based on the number of items in the subscale bundles (two bundles of 5 items and one bundle of 6 items). This type of bundling was used for establishing a baseline for the potential magnitude of DBF when bundles have been created without using a conceptually or theoretically based organizing principle. Yen (1993) used a similar approach in creating testlets for comparison purposes in an investigation of the effects of local item dependence.

Group Contrasts

The first contrast was between randomly formed groups. These groups were formed by using the random selection after combining the parent and teacher ratings into a single data file. The groups were then randomly formed by selecting approximately 50% of the cases for Group 1 and the remaining cases for Group 2. For this comparison, Group 1 served as the reference group and Group 2 as the focal group. This contrast was set up to provide a comparison value for groups that theoretically have no distinguishing characteristics.

Then, two more sets of contrasting groups were used in this study for demonstration purposes; the teacher–parent and boy–girl contrast. These particular contrasts were chosen because of their appearance in the literature on child behavior ratings. Several researchers have addressed behavior ratings from the standpoint of differences in ratings for boys versus girls as well as for ratings provided by parents versus teachers (e.g., McGee & Feehan, 1991; Rowe & Kandel, 1997). In the teacher–parent contrast, the parent rating served as the reference or anchor in making the comparison. This comparison was similar to the comparison addressed in the study by Maurer et al. (1998), in which the measurement equivalence of a performance appraisal scale across peer and subordinate rater populations was examined using the DFIT framework. The boy–girl contrast was between boy students and girl students for each type of rater, parent or teacher, with girls serving as the reference or anchor for the comparison. This contrast represents a more traditional comparison within the context of DIF.

Assessment of Unidimensionality

Since a unidimensional model was applied to the data, unidimensionality was assessed with confirmatory factor analysis (CFA), using LISREL 8 (Joreskog & Sorbom, 1996). The CFA analysis was conducted separately by rater group, but with identical factor patterns. Prior to conducting CFA, polychoric correlations between items within each scale and rater group were obtained using PRELIS 2 (Joreskog & Sorbom, 1996). Results from the CFA analysis are summarized in Table 1. It shows four commonly used indexes of fit: Root mean square approximation (RMSEA), goodness-of-fit index (GFI), adjusted goodness-of-fit index (AGFI), and comparative fit index (CFI), separately by scale and rater group. For the RMSEA index, .05 or a lower value is a typical benchmark for good fit, and .08 is generally considered an upper bound for acceptable fit (Browne & Cudek, 1993). The two RMSEA indexes in Table 1 are .073 for teachers and .055 for parents, and both are below the upper bound. CFI, GFI, and AGFI values at or above .95 are generally considered indicative of good fit (Byrne, 1998; Hu & Bentler, 1999), whereas .90 is an appropriate lower bound of adequate fit. As shown in Table 1, these indexes are above .95, with the exception of the CFI

TABLE 1
Goodness-of-Fit Indexes by Rater Group From a Confirmatory Factor Analysis (One-Dimensional Solution)

Group	*RMSEA*	*GFI*	*AGFI*	*CFI*
Teachers	.073	.984	.980	.976
Parents	.055	.981	.974	.905

Note. RMSEA = root mean square error of approximation;
GFI = goodness-of-fit index;
AGFI = adjusted goodness of fit index;
CFI = comparative fit index.

index for parents, which is .905. This latter index meets the lower bound criterion of .90 for adequate fit. Overall, these indexes appear to suggest that the PB scale is unidimensional across the two rater groups. According to Byrne (1998, pp. 103–119), the indexes shown in Table 1 also appear to meet the generally recommended cutoffs for an acceptable fit.

IRT Calibrations and Equating

Each scale was calibrated using PARSCALE2 (Muraki & Bock, 1993). This program is especially appropriate for calibrating graded response, polytomous items. The maximum marginal likelihood and expectation–maximization (EM) algorithm were used to estimate the item parameters. The default values were used for all estimations. In addition, estimation of underlying abilities was made using Bayesian EAP procedures using normal priors. The item-level (chi-square) goodness-of-fit indexes from PARSCALE were used to assess the fit of Samejima's graded response model to the current data set. All of the item-level chi-squares were nonsignificant, except for an occasional item in some groups with significant chi-square.

Prior to running the DFIT and DBF analyses, the item parameter estimates from PARSCALE2 were put on a common metric using EQUATE 2.0 (Baker, 1993). The estimation of equating coefficients was made by using Baker's modified test characteristic curve method.

A two-stage linking procedure was used. At the first stage, after item parameters were placed on a common scale using all items on the test, the DFIT program (Raju, 1997) was used to identify items with large DIF (NCDIF > .024). Identified items were then deleted from the linking in the second stage. Finally, all items were transformed using the linking coefficients obtained by the second stage linking procedure. The item parameters from the reference group were equated to the metric of the focal group.

DFIT and DBF Analyses

The DFIT program was used to calculate the DFIT indexes as well as the DBF indexes. The standard DIF/DTF program was first run with all *n* items to obtain the item-level CDIF values. The CDIFs for all items in a bundle were then summed to obtain the CDBF indexes (Equation 9). For NCDBF, the standard DIF/DTF program was run for each bundle as if the bundle was the whole test (Equation 10). This process was repeated as many times as the number of bundles under consideration. Once the CDBF and NCDBF indexes were obtained, these values were interpreted with respect to differences across bundle types and group contrasts within the context of the literature related to behavior ratings.

RESULTS

Random-Group Comparison

Results from the random-group comparison are reported in Table 2. This group comparison can serve as a baseline for the differential functioning that might be expected when groups are formed without respect to any particular group characteristic.

Teacher–Parent Comparison

The teacher–parent comparison is presented in Table 3. Although not reported in Table 3, the problem behavior bundles displayed considerably less evidence of

TABLE 2
CDBF and NCDBF for Each Bundle of the Problem Behavior Scale for the Random Groups, Group1 (Reference), and Group 2 (Focal) Comparison

Bundle Name	*Number of Items*	*CDBF*	$\overline{NCDBF}$
Subscale			
Externalizing	6	.000	.36
Internalizing	5	.000	.85
Hyperactivity	5	.000	.18
Random			
Bundle 1	6	.000	.07
Bundle 2	5	.000	.02
Bundle 3	5	.000	.14

Note. CDBF = Compensatory differential bundle functioning; $\overline{NCDBF}$ = noncompensatory differential bundle functioning. Actual $\overline{NCDBF}$ values have been multiplied by 1,000.

TABLE 3
CDBF and NCDBF for Each Bundle of the Problem Behavior Scale for the Parent (Reference) vs. Teacher (Focal) Comparison

Bundle Name	*Number of Items*	*CDBF*	$\overline{NCDBF}$
Subscale			
Externalizing	6	–.039	51.04[a]
Internalizing	5	.024	30.11[a]
Hyperactivity	5	.034	16.94[a]
Random			
Bundle 1	6	–.018	8.70
Bundle 2	5	.038	28.14
Bundle 3	5	–.001	.82

Note. CDBF = compensatory differential bundle functioning; NCDBF = noncompensatory differential bundle functioning. Actual $\overline{NCDBF}$ values have been multiplied by 1,000.
[a]Indicates significant $\overline{NCDBF}$ for subscales at $\alpha = .001$.

differential functioning than the social skills bundles. This outcome may be because problem behaviors in general are easier to recognize and are less context specific. That is to say, problem behaviors generalize more than social skills to multiple contexts. The set of items with the highest $\overline{NCDBF}$ value is the externalizing bundle. This result may occur because the type of behavior being assessed by this bundle would be more readily recognized in a school setting where acting-out behaviors come to the attention of the teacher more quickly than other behaviors.

In terms of the random bundles, one can see that even when bundles are formed without any conceptual framework, there appears to be some differential functioning with respect to the teacher–parent comparison. Furthermore, one can observe some variation among Bundles 1–3, which is expected given a random assignment with small sample sizes (6, 5, and 5 items). These values can serve as guidelines for evaluating other meaningful bundles. For example, CDBF values larger than .04 and $\overline{NCDBF}$ values larger than .03 (30/1,000) may suggest nonrandomness for these data. By these estimates, one particular bundle that is beyond the range associated with randomness is the $\overline{NCDBF}$ value for the externalizing subscale. Based on the newly developed significance test, all subscale bundles display statistically significant $\overline{NCDBF}$.

Boy–Girl Comparison

The subscale bundle analysis for the boy–girl comparison for parents and teachers is presented in Table 4. For the parent as rater, it is interesting to note that for

TABLE 4
CDBF and NCDBF for Each Bundle of the Problem Behavior Scale for the Girls (Reference) vs. Boys (focal) Comparison for Parents and Teachers

Bundle Name	*Number of Items*	*CDBF*	$\overline{NCDBF}$
Parents			
Subscale			
Externalizing	6	.000	.02
Internalizing	5	.000	8.73[a]
Hyperactivity	5	.000	8.59[a]
Random			
Bundle 1	6	.000	.70
Bundle 2	5	.000	.79
Bundle 3	5	.000	.14
Teachers			
Subscale			
Externalizing	6	.006	1.45
Internalizing	5	–.021	61.00[a]
Hyperactivity	5	.020	51.44[a]
Random			
Bundle 1	6	–.001	2.39
Bundle 2	5	–.003	1.57
Bundle 3	5	.003	.98

Note. CDBF = compensatory differential bundle functioning; NCDBF = noncompensatory differential bundle functioning. Actual $\overline{NCDBF}$ values have been multiplied by 1000. The asterisks
[a]Indicates significant $\overline{NCDBF}$ for subscales at $\alpha = .001$.

problem behaviors the bundle indexes are quite small, suggesting that parents are equitable raters across all three types of behaviors assessed with these bundles of items.

For the bundle indexes for teachers' boy–girl comparison, the hyperactivity bundle and the internalizing bundle have the highest values of $\overline{NCDBF}$. However, looking at the CDBF values, one can see that boys are favored in one bundle and girls are favored in the other, essentially serving to cancel each other. With this cancellation, there seems to be some evidence of the measurement equivalence across gender for teacher ratings of problem behaviors when a total score is obtained. However, there is evidence of differential functioning across gender when the focus is specifically on internalizing or hyperactive behaviors. The $\overline{NCDBF}$ and CDBF values for the externalizing bundle are very small compared to the other two bundles, suggesting that teachers provide equivalent ratings for boys and girls with respect to externalizing behaviors. This finding could be due to the fact that externalizing behaviors are more observable and in fact easier to

see, so that gender has little influence on the rating of these behaviors. The greater degree of differential functioning seen with the internalizing and hyperactivity bundles may be because these behaviors are less visible and behavioral expectations related to gender may play a role in teacher ratings.

A random bundle analysis was also conducted for the boy–girl comparison for parents and teachers. The CDBF and $\overline{NCDBF}$ values are quite small for all of the comparisons. Again, using the estimate of randomness, two particular bundles, internalizing and hyperactivity subscales (especially for teachers), appeared to show nonrandomness. The significance test for the boy–girl comparisons for both parents and teachers indicated statistically significant values for the internalizing and hyperactivity bundles.

Graphic Display

To facilitate a comparison of the $\overline{NCDBF}$ values across different bundle types and different group comparisons, the $\overline{NCDBF}$ values by group comparison are displayed in Figure 1. In Figure 1, several observations discussed earlier can be highlighted. First, as expected, the most striking feature is that the $\overline{NCDBF}$ values for the random group comparison are considerably smaller than the values seen for other meaningful group comparisons. Second, externalizing behaviors seems to show the highest DBF for the teacher–parent comparison. Third, parents appeared to exhibit less differential functioning than teachers in terms of the boy–girl comparison. Fourth, internalizing and hyperactivity behaviors are more susceptible to DBF than externalizing behaviors for the boy–girl comparison.

DISCUSSION

A procedure to conduct a DBF analysis was introduced within the DFIT framework for the polytomous data. Using polytomous questionnaire data, a teacher–parent contrast as well as the gender contrast within each group (teacher or parent) was used to demonstrate the information obtained when a DBF analysis was undertaken to address differential functioning for bundles created by subscales. The random-group comparison and the randomly created bundles were included to provide a type of baseline comparison value.

Researchers have suggested that analyzing groups of items, as opposed to single items, may provide a more useful way to examine differential functioning. While an item-level analysis is a crucial part of examining differential functioning, the nature of a DBF analysis encourages or even requires the researcher to approach the analysis from a more substantive standpoint by a priori identifying item characteristics that could theoretically lead to differential functioning.

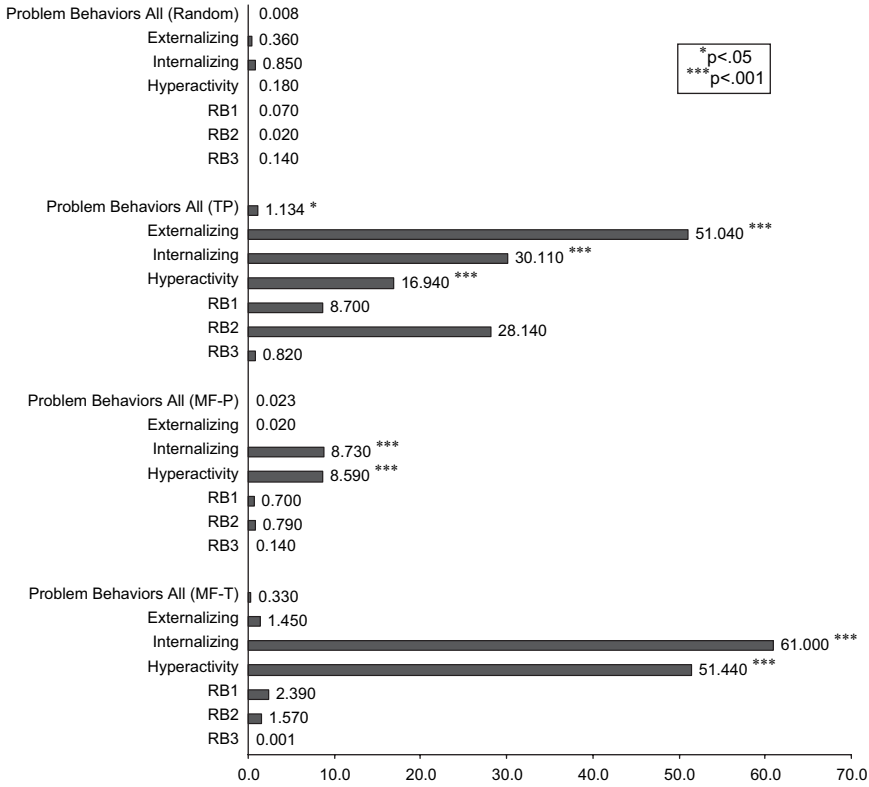

FIGURE 1 A comparison of $\overline{NCDBF}$ values for all group comparisons for the problem behavior bundles. All values have been multiplied by 1,000. NCDBF = noncompensatory differential bundle functioning.

The examples presented in this study are based on this conceptual approach, calling on research in the area of child behavior ratings to identify relevant comparison groups as well as item organizing principles. Following this example, the computation and interpretation of CDBF and $\overline{NCDBF}$ indexes were discussed. The CDBF index was shown to be useful in examining the cancellation effect across bundles as well as the overall direction of the differential functioning for a particular bundle. This type of information could be useful in a test development phase, where the information could provide direction for the writing or editing of items for the purposes of reducing differential functioning. The $\overline{NCDBF}$ index, on the other hand, shed light on the possible sources of differential functioning. For example, the greatest source of differential functioning came from the externalizing subscale for the teacher–parent comparison, perhaps due to the fact that acting-out behaviors are less tolerated in school than at home.

In considering the utility of this methodology, it is useful to look at the results of the analyses in terms of theoretical expectations. Generally speaking, the results of the analyses were consistent with the behavior rating literature. In addition, this study showed that the use of both random groups and random bundles can be useful as a means of identifying meaningful values of DBF in the DIFT framework.

The strengths of the DBF method introduced here include the versatility of the DFIT framework, which allows the examination of differential functioning at three different levels (item, bundle, and test) using the parametric IRT. Until recently, the major weakness of the DFIT framework was the lack of a significance test that was not overly sensitive to large sample sizes. However, a new significance test based on the IPR method was recently introduced and evaluated for dichotomous data. The work on the significance test for polytomous DFIT is currently under way. Future research is needed to evaluate the performance of the DFIT–DBF analysis with the new significance test.

ACKNOWLEDGMENT

We would like to dedicate this article to Dr. Nambury Raju, who passed away on October 27, 2005. This article could not have been completed without the contributions made by Dr. Raju. In addition to sharing his ideas and his exceptional intellectual insight, Dr. Raju was quick to provide positive feedback and to encourage persistence. We are grateful to have had the opportunity to work with a scholar who has contributed so widely to the field of psychometric theory. Dr. Raju's intellectual insight was overshadowed only by his positive spirit and good nature.

REFERENCES

Baker, F. B. (1993). *EQUATE2: Computer program for equating two metrics in item response theory* [Computer software]. Madison: University of Wisconsin, Laboratory of Experimental Design.

Browne, M. W., & Cudek, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking matrices and assessing item bias in item response theory. *Applied Psychological Measurement, 12,* 253–260.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential 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 (Doctoral dissertation, Illinois Institute of Technology, 1993). *Dissertation Abstracts International, 54–04B,* 2266.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23*, 309–326.

Gierl, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. *Journal of Educational Measurement, 40*, 281–306.

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

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

Gresham, F. M., & Elliott, S. N. (1990). *The social skills rating system.* Circle Pines, MD: American Guidance Services.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Joreskog, K. J., & Sorbom, D. (1996). *LISREL 8: User's reference guide*. Chicago: Scientific Software.

Keogh, B. K., & Bernheimer, L. P. (1998). Concordance between mothers' and teachers' perceptions of behavior problems of children with developmental delays. *Journal of Emotional and Behavioral Disorders, 6*, 33–41.

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.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Maurer, T., Raju, N. S., & Collins, W. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology, 83*, 693–702.

McGee, R. & Feehan, M. (1991). Are girls with problems of attention underrecognized? *Journal of Psychopathology and Behavioral Assessment, 13*, 187–198.

Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement, 16*, 381–388.

Muraki, E., & Bock, R. D. (1993). *PARSCALE2: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks* [Computer software]. Chicago: Scientific Software International.

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.

Oshima, T. C., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance in the Differential Functioning of Items and Tests (DFIT) Framework. *Journal of Educational Measurement, 43*, 1–17.

Raju, N. S. (1997). *DFITPU: A Fortran program for calculating DIF/DTF* [Computer software]. Atlanta: Georgia Institute of Technology.

Raju, N. S. (1999). *Some notes on the DFIT framework*. Unpublished manuscript, Illinois Instititute of Technology, Chicago.

Raju, N. S., Oshima, T. C., Fortmann, K., Nering, M., & Kim, W. (2006, February). *The new significance test for Raju's polytomous DFIT*. Paper presented at New Directions in Psychological Measurement With Model-Based Approaches Conference. Atlanta, GA.

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

Rowe, D. C., & Kandel, D. (1997). In the eye of the beholder? Parental ratings of externalizing and internalizing symptoms. *Journal of Abnormal Child Psychology, 25*(4), 265–275.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph,* No. 17.

Shealy, R., & Stout, W. (1993a). An item response theory model for test bias. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 213–316). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Tate, R. (2002). Test dimensionality. In J. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 181–211). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

APPENDIX

Let there be v bundles ($B_1, \ldots, B_j, \ldots, B_v$) with n_j items in bundle j, and let $n = n_1 + \ldots + n_j + \ldots + n_v$ represent the total number of items in the test or scale. Using the expected raw score or true score definitions in Equations 2 and 3, the expected raw score or true score for bundle j for examinee s may be expressed as:

$$T_{sB_j}(\theta_s) = ES_{sB_j}(\theta_s) = \sum_{i \varepsilon B_j} ES_{si}(\theta_j) = \sum_{i \varepsilon \boldsymbol{B}_j} \left[\sum_{k=1}^{m} P_{ik}(\theta_j) X_{ik} \right]. \tag{A1}$$

Furthermore,

$$T_s(\theta_s) = \sum_{j=1}^{v} ES_{sB_j}(\theta_s) = \sum_{j=1}^{v} \left[\sum_{i \in B_j} ES_{si}(\theta_s) \right] = \sum_{i=1}^{n} ES_{si}(\theta_s), \tag{A2}$$

where $T_s(\theta_s)$ is the true score at the test level for examinee s. Given Equation 5, differential functioning at bundle j may be defined as:

$$\begin{aligned} d_{sB_j}(\theta_s) &= ES_{sB_jF}(\theta_s) - ES_{sB_jR}(\theta_s) \\ &= \sum_{i \in B_j} [ES_{siF}(\theta_s) - ES_{siR}(\theta_s)] = \sum_{i \in B_j} d_{si}(\theta_s), \end{aligned} \tag{A3}$$

where F and R refer to the focal group and reference group, respectively. Within the framework of bundles, differential functioning at the test/scale level for examinee *s* may be expressed as:

$$D(\theta_s) = \sum_{j=1}^{v} d_{sB_j}(\theta_s) = \sum_{j=1}^{v}\left[\sum_{i\in B_j} d_{si}(\theta_s)\right] = \sum_{i=1}^{n} d_{si}(\theta_s). \quad \text{(A4)}$$

Given Equation A4, DTF, within the context of bundles, may be rewritten as:

$$DTF = \varepsilon\left\{[\sum_{i=1}^{n} d_{si}(\theta_j)]^2\right\} = \varepsilon\left\{[\sum_{j=1}^{v} d_{sB_j}(\theta_j)]^2\right\} = \sum_{j=1}^{v}\left[Cov(d_{B_j}, D) + \mu_{d_{B_i}}\mu_D\right], \quad \text{(A5)}$$

where the expectation (ε) is taken over the focal group. Now, within Raju et al.'s (1995) DFIT framework, compensatory differential bundle functioning for bundle *j* ($CDBF_j$) may be written as:

$$CDBF_j = Cov(d_{B_j}, D) + \mu_{d_{B_i}}\mu_D. \quad \text{(A6)}$$

Because

$$d_{sB_j}(\theta_s) = \sum_{i\in B_j} d_{si}(\theta_s), \quad \text{(A7)}$$

$CDBF_j$ may be written as:

$$CDBF_j = \sum_{i\in B_j}\left[Cov(d_i, D) + \mu_{d_i}\mu_D\right] = \sum_{i\in B_j} CDIF_i, \quad \text{(A8)}$$

which is exactly the definition of $CDBF_j$ given in Equation 9. That is, $CDBF_j$ is simply the sum of CDIF indexes of items in bundle *j*. In addition, because

$$D_s(\theta_s) = \sum_{j=1}^{v} d_{sB_j}(\theta_s), \quad \text{(A9)}$$

$CDBF_j$ may also be written as:

$$CDBF_j = Cov(d_{B_i}, \sum_{j=1}^{v} d_{B_j}) + \mu_{d_{B_j}}\mu_{\sum_{j=1}^{v} d_{B_j}} = \sum_{l=1}^{v}\left[\mathrm{cov}(d_{B_j}, d_{B_l}) + \mu_{d_{B_j}}\mu_{d_{B_j}}\right]. \quad \text{(A10)}$$

If we now assume that $d_{B_j}(\theta_s) = 0$ for all $i \neq j$ (that is, all bundles except bundle B_j have zero differential functioning), according to the definition of noncompensatory DIF in the DFIT framework, an index of noncompensatory differential bundle functioning for bundle j ($NCDBF_j$) may be expressed as:

$$NCDBF_j = Cov(d_{B_j}, d_{B_j}) + \mu_{d_{B_j}} \mu_{d_{B_j}} = \sigma^2_{d_{B_j}} + \mu^2_{d_{B_j}}. \qquad \text{(A11)}$$

In view of Equation A7, the above expression may be rewritten as:

$$NCDBF_j = \sum_{i \in B_j} (\sigma^2_{d_i} + \mu^2_{d_i}) + 2 \sum_{(i \prec k) \in B_j} \left[\text{cov}(d_i, d_k) + \mu_{d_i} \mu_{d_k} \right]. \qquad \text{(A12)}$$

That is,

$$NCDBF_j = \sum_{i \in B_j} NCDBF_i + 2 \sum_{(i \prec k) \in B_j} \left[\text{cov}(d_i, d_k) + \mu_{d_i} \mu_{d_k} \right]. \qquad \text{(A13)}$$

As previously noted, unlike the CDBF index, the NCDBF index is not simply the sum of NCDIF indexes of items in a given bundle (Equation A13). Based on Equation A11, however, NCDBF is DTF at the bundle level. Given this relationship between NCDBF and the bundle-level DTF, it is possible to define appropriate tests for assessing the statistical significance of NCDBF indexes. The chi-square tests previously proposed by Raju et al. (1995) for the DTF index would be equally appropriate for the NCDBF index when treated as a bundle-level DTF.