Learning Disabilities: A Contemporary Journal 2(2), 49-69, 2004

An Introduction to Differential Item Functioning Analysis

Akihito Kamata[1]
Brandon K. Vaughn
Florida State University

This article provides a brief primer overview of Differential Item Functioning (DIF) analysis. DIF analysis investigates a differential characteristic of a test item between subpopulations of examinees and is useful in detecting possibly biased items toward a particular subpopulation. As demonstration, a dataset from a 40-item math test in a statewide testing program is analyzed by two widely used approaches, namely, Mantel-Haenszel and logistic regression methods, for the purpose of detecting DIF items against learning disability students who received testing accommodations. As a result, three items were found to present DIF against LD students who received testing accommodation, while one item was found to present DIF against typical students. Also, a discussion on future directions of DIF analysis study is provided.

Keywords: differential item functioning, learning disability, testing accommodations

Through the expansion of educational reforms, several initiatives have been proposed in an attempt to provide similar testing experiences and opportunities to individuals with and without disabilities. The incorporation and modifications of the Individuals with Disabilities Education Act (IDEA) and the Americans with Disabilities Act (ADA) provided a strong incentive and motivation to include individuals with disabilities in nationwide, statewide and districtwide assessment measures, thereby helping these individuals achieve new standards in academic situations (Pitoniak & Royer, 2001). In addition, the No Child Left Behind Act of 2001 mandates all states have annual tests in reading and mathematics in Grades 3 to 8 by 2005-2006 (U.S. Department of Education, 2002). Thus, there has been increasing attention in recent years to testing accommodations in order to include more students with disabilities in testing programs. For example, beginning in 1996, National Assessment of Educational Progress (NAEP) started the incorporation of these accommodations, such as extended time, large print, transcription, oral reading, and signing of directions, for students whose schools felt it necessary to provide such accommodations during testing situations. Also, statewide testing programs, such as ones in the state of Florida, have rigorously implemented testing accommodations for students with disabilities, including students with learning disabilities, for many years (Florida Department of Education, 2004).

1. Address correspondence to Akihito Kamata, 307 Stone Building, Department of Educational Psychology & Learning Systems, Tallahassee, FL 32306-4453, (850) 894-6097, kamata@coe.fsu.edu

Psychometric Definition of Bias

Because of increasing focus on the inclusion of more students, there has also been increasing attention to equity of test scores for various subpopulations, including students with learning disabilities. Test equity is primarily achieved by ensuring that a test measures only construct-relevant differences between subpopulations (Messick, 1989). If test equity is not achieved, a test or test item is biased toward a particular subpopulation of the test taking population. Statistically, a test or test item is said to be biased if the expected test or item scores are not the same for examinees from different sub-populations, given the same level of trait that the test intends to measure. (In academic achievement tests, a trait level is a level of achievement, or more simply, ability. In this paper these terminologies are used interchangeably.) This definition of bias clearly indicates that a bias is a "conditional" mean difference, and we have to distinguish a bias from a simple observed mean difference between subpopulations. For example, when the learning disabled (LD) sub-sample has a lower mean score on an achievement test than the typical student sample (i.e., existence of observed mean difference), it does not necessarily mean that the test score is biased toward LD students. It may be simply an indication that LD students have a lower achievement level in the population, rather than an indication of bias against LD students. In order to establish evidence for bias, one needs to demonstrate difference between LD and typical students on their test scores by conditioning on their trait levels (ability, in this case).

In many cases, researchers are interested in bias at the item level. This is particularly useful in a process of test development, in which biased test items are revised or removed. This is a legitimate and important process in an attempt to achieve test equity. If bias is only investigated at the test level, then the existence of bias at the item level may be overlooked and ignored. Also, it is still important to investigate bias at item level for test items in operational test forms. Even though most standardized tests have gone through the process of removing any biased items, it is still likely that some test items are biased against a particular subpopulation of interest. Subpopulations of concern are typically based on ethnicity or gender in the test construction process. Therefore, it is not always the case that the subpopulation of one's interest has been considered in the test construction process. Thus, investigation of bias at the item level allows one to investigate the impact of such items for the subpopulation of interest.

Differential Item Functioning (DIF)

One way to investigate bias at the item level is differential item functioning (DIF) analysis. DIF is said to be present in a test item when examinees from two subpopulations with the same trait level have different expected scores on the same item. This differential item characteristic is also known as item invariance across subpopulations. One possible consequence of having an item with DIF is differential total test scores or trait level estimates for examinees with the same trait levels from different subpopulations, consequently an unfair disadvantage to a particular subpopulation. Therefore, it is a psychometrician's concern to detect items with a large magnitude of DIF and remove them from the test before the test is operationally administered. When an item is biased, it is expected that the item would show DIF. However, the

existence of DIF does not necessarily mean bias against a particular subpopulation. DIF may be present for reasons other than bias. Therefore, we should interpret an item with DIF as a "possibly biased item" or simply refer to it as a "DIF item".

DIF analysis is particularly of interest to researchers on students with learning disabilities for at least two reasons. First, when test scores, such as an academic test or personality inventory, are utilized as a variable in a research study, one has to make sure that LD students' achievement levels are measured without any bias. If a variable is measured with bias, any subsequent statistical analysis will be biased too. Second, DIF analysis can be extended to an investigation of variables (both item characteristics and person/school characteristics) that are related to the magnitude of DIF. This type of analysis may help us understand why DIF is happening and may reduce the magnitude of DIF in future construction and administration of tests. See the last section of this paper for brief but more detailed discussion on this point.

There are several terminologies that the reader may need to be familiar with before we present specific DIF analysis procedures. First, there are common terminologies that are used to reference subpopulations in a DIF study. The group of interest in DIF analysis is referred as the *focal group,* while the group to be compared is referred as the *reference group.* (Typically, DIF between two groups is studied.) For example, in LD research, a group of learning disability students will probably be referred as the focal group, and a group of typical students will probably referred as the reference group. Second, DIF analysis requires comparing performance difference between reference and focal groups by controlling for trait level. Therefore, a measure of trait level that is comparable between focal and reference groups is required to match the examinees on their trait levels. Such a measure of trait level is referred as a *matching criterion.* A matching criterion could be the performance of a test, such as total test scores or trait level estimates as a latent trait, from which the studied items come. Alternatively, an external criterion could be used, such as performance on another test that is believed to be measuring the same construct as the studied item. In either case, a matching criterion should be perfectly construct-valid to the extent that it is measured without contamination from any unintended factor, namely, free from DIF items (Shealy & Stout, 1993). In order to achieve this, a process called *purification* may be employed, especially when the performance of a test, which is constructed from the items of study is used as a matching criterion. (See the subsequent data analysis section for example.) Third, when the magnitude of DIF is the same across all trait levels, it is referred as *uniform DIF*. On the other hand, it is referred as *non-uniform DIF,* when the magnitude of DIF is not consistent across trait levels.

To date, many DIF analysis models/techniques have been proposed. They include the delta plot (Angoff & Ford, 1973), analysis of variance method (Plake, 1981), contingency table approaches, such as the Mantel-Haenszel method (Holland & Thayer, 1988), logistic regression approach (Swaminathan & Rogers, 1990; Rogers & Swaminathan, 1993), confirmatory factor analysis approach (Muthen, 1989), and item response theory (IRT) approaches, such as loglinear item response models (Kelderman, 1989), area measures (Raju, 1988; Raju, van der Linden, & Fleer, 1995), likelihood-ratio method (Thissen, Steinberg, & Wainer, 1993), standardized index of bias (SIB) (Muraki & Engelhard, 1989), multidimensional IRT approach (e.g., Ackerman, 1992; Roussos &

Stout, 1996), and generalized linear mixed model approach (e.g., Luppescu, 2002; Meulders & Xie, 2004). See Holland and Wainer (1993) and Millsap and Everson (1993) for more detailed overview for some of these methods.

Methods of DIF Analysis

Following sections demonstrate a series of DIF analysis procedures by the Mantel-Haenszel method, as well as by the logistic regression method. These two methods are chosen for this paper because they are widely implemented in both practical test construction and research settings and can be implemented by widely available statistical package software, such as SPSS. This paper only considers cases where all items are dichotomously scored, such as correct-incorrect and yes-no. However, DIF analysis is also possible for polytomously scored test items, such as Likert-type scale and rating scale.

Data

Data used in the following illustrative analyses were sampled from the 2003 administration of mathematics assessment for 3rd graders in a statewide testing program in a southeastern state in the United States. The test consisted of 40 items based on five subscales, including Number Sense, Measurement, Geometry, Algebraic Thinking, and Data Analysis. The items were all in multiple-choice format and scored dichotomously (correct or incorrect). All non-responded items were scored as incorrect answers. There were a total of 188,595 examinees, including 156,927 students in standard curricula and 12,317 students classified as specific learning disabled. For the illustrative data analyses, 1,842 students in standard curricula who did not receive any testing accommodation and 1,854 students who were classified as specific learning disabled and received at least one type of testing accommodation were randomly sampled. Descriptive statistics for the sample are summarized in Table 1.

Table 1 *Summary Statistics of Total Test Scores in the Sample Data*

	n	*M*	*SD*
LD Students	1,854	16.35	7.05
Students in Standard Curricula	1,842	23.57	7.45

Mantel-Haenszel Method

The Mantel-Haenszel (MH) procedure (Mantel & Haenszel, 1959) is a general statistical approach to test for the dependency of two variables in a three-way contingency table. Holland and Thayer (1988) proposed the use of the MH procedure to detect DIF between two sub-samples of examinees for a dichotomously scored test item, by summarizing test data in a form of a (2 scoring categories) $\times$ (2 subpopulations) $\times$ (k categories in matching criterion) contingency table.

The MH method of DIF analysis involves the construction of a contingency table, which gives the counts of correct (1) and incorrect (0) responses. These counts are broken up by the group indicator (focal and reference groups) and the matching criterion (k categories). The structure of a contingency table and associated notations are presented in Figure 1.

Figure 1. *Contingency table and associated notations.*

		Response to item j		
		Correct (1)	Incorrect (0)	Total
Group	Reference	a_j	b_j	n_{r_j}
	Focal	c_j	d_j	n_{f_j}
	Total	n_{1_j}	n_{0_j}	N_j

The MH method employs several quantities. First, it computes a quantity called the *common odds-ratio*. The estimate of the common odds-ratio for item i ($\hat{\alpha}_{MH_i}$) is obtained by

$$\hat{\alpha}_{MH_i} = \frac{\sum_j^S a_j d_j / N_j}{\sum_j^S b_j c_j / N_j},$$

where a_j, b_j, c_j, and d_j are defined in the 2×2 contingency table in Figure 1, j is the jth category in the matching criterion ($j = 0, \ldots, S$), and N_j is the number of examinees in the jth category. If there is no difference between the reference and focal groups by controlling for the level of matching criterion, then $\hat{\alpha}_{MH_i}$ will be equal to 1. If the reference group performs better on the item, then $\hat{\alpha}_{MH_i}$ will be smaller than 1, an indication of possible bias against the focal group. On the other hand, if the common odds-ratio is greater than 1, it is an indication of possible bias against the reference group.

Another common quantity in the MH method is a *signed index*. This signed index is simply the natural-log of the common odds ratio. The signed index for item i is denoted by $\hat{\beta}_{MH_i}$ and computed as follows:

$$\hat{\beta}_{MH_i} = \ln(\hat{\alpha}_{MHi}).$$

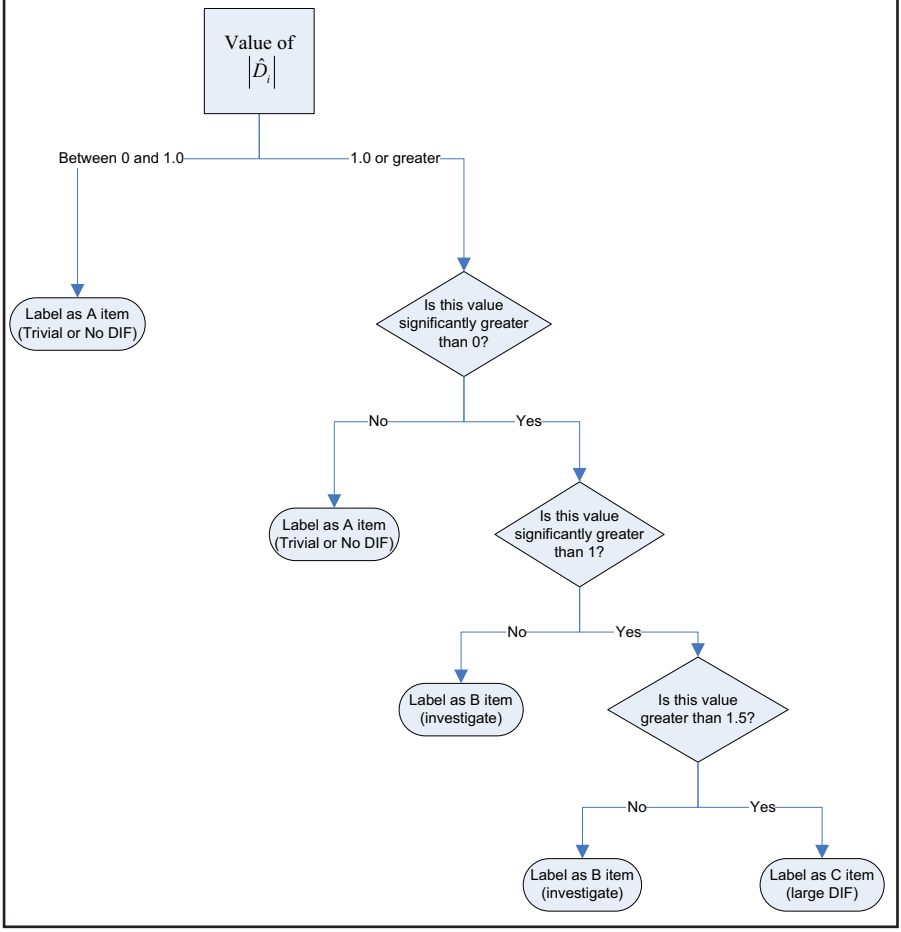
A signed index is in a scale of log-odds-ratio, which is equivalent to a coefficient in a logistic regression model. (We will make a connection between the MH and logistic regression methods on this quantity later in the paper.) When $\hat{\alpha}_{MH_i}$ is smaller than 1, $\hat{\beta}_{MH_i}$ is a negative value (possible bias against focal group). On the other hand, when $\hat{\alpha}_{MH_i}$ is larger than 1, $\hat{\beta}_{MH_i}$ is a positive value (possible bias against reference group). When $\hat{\alpha}_{MH_i} = 1$, $\hat{\beta}_{MH_i} = 0$, indicating that there is no difference on the proportion of correct answers between the reference and focal groups given the same level in the matching criteria.

Further transformation of $\hat{\beta}_{MH_i}$ may be helpful for evaluating the magnitude of DIF. The criteria involved a transformation of the signed index into a quantity called the *MH-DIF statistic* $\hat{D}_i$ (Holland & Thayer, 1988):

$$\hat{D}_i = -2.35\,\hat{\beta}_{MH_i} = -2.35\ln(\hat{\alpha}_{MH_i}).$$

When $\hat{D}_i$ is a positive value, it is an indication of possible bias against the focal group, while a negative value is an indication of possible bias against the reference group. According to the Educational Testing Service (ETS) criteria (Dorans & Holland, 1993), an item is flagged for large DIF when the absolute value of $\hat{D}_i$ is greater than 1.5 and significantly greater than 1.0. Such items (commonly referred to as "C" items) are removed from the test or revised. While common statistical computer packages that calculate MH do not report the significance level for this hypothesis test, a researcher has several options. For example, 95% confidence intervals for $\hat{\alpha}_{MH_i}$ are typically given, and can be used to assess if $\hat{D}_i$ is significantly greater than 1.0 by use of the transformation of $\hat{D}_i$. Any items with values of $\hat{D}_i$ less than 1.0 in magnitude or not statistically different from zero (commonly referred to as "A" items) are considered negligible for DIF and not deleted or revised. All remaining items with $\hat{D}_i$ values greater than 1.0 but not meeting the first condition (commonly referred to as "B" items) are considered suspicious and recommended for further inspection. These classification rules are summarized in Figure 2. It should be

Figure 2 *DIF Classification rules.*



noted that the transformation to $\hat{D}_i$ is not an absolutely necessary process for the MH method. The motivation to transform $\hat{\alpha}_{MH_i}$ or $\hat{\beta}_{MH_i}$ into $\hat{D}_i$ is strictly based on the interest to interpret the magnitude of DIF in the scale of the item difficulty as measured by the existing scale (see Holland & Thayer, 1985). For example, $\hat{D}_i = 0.0$ corresponds to $\hat{\alpha}_{MH_i} = 1.00$, $\hat{D}_i = 1.0$ corresponds to $\hat{\alpha}_{MH_i} = 0.65$, and $\hat{D}_i = 1.5$ corresponds to $\hat{\alpha}_{MH_i} = 0.53$. These $\hat{\alpha}_{MH_i}$ values and associated confidence intervals can be used for the ETS criteria, in place of $\hat{D}_i$.

The MH Chi-square statistic is defined as

$$MH_{\chi^2} = \frac{\left[\left|\sum_{j=1}^{S}\left(a_j - E(a_j)\right)\right| - 0.5\right]^2}{\sum_{j=1}^{S} Var(a_j)},$$

where

$$Var(a_j) = \frac{n_{r_j} n_{f_j} n_{1_j} n_{0_j}}{N_j^2 (N_j - 1)} \quad \text{and} \quad E(a_j) = \frac{n_{r_j} n_{1_j}}{N_j}.$$

The MH_{X^2} tests the null hypothesis that the odds of getting an item correct is the same for both the focal group and reference group across all levels in the matching criterion. A rejection of the null hypothesis is an indication of the presence of DIF. Since the MH_{X^2} is chi-square distributed with $df = 1$, the null hypothesis will be rejected for any MH_{X^2} value greater than 3.84 (a critical value for χ^2 with $df = 1$ and 0.05 significance level).

We now illustrate the data analysis of DIF using the MH method. The statewide testing program data is analyzed using the SPSS[1] software. Prior to the MH procedure, a total test score is calculated as a matching criterion. Also, examinees in the focal group (LD students) are coded 1, and examinees in the reference group (students in standard curricula) are coded 0. To implement the MH method, the "Crosstabs" procedure in SPSS can be used. More details for running the procedure are described in the Appendix.

A sample output from SPSS for the first item is given in Figure 3. Here, the common odds ratio ($\hat{\alpha}_{MH_i}$) is estimated as .577, indicating the odds of a correct answer for LD students is about 58% of the odds for students in standard curricula, given the same level of matching criterion. It can be transformed into a signed index $\hat{\beta}_{MH_i}$ $\ln(\hat{\alpha}_{MH_i}) = -.551$ and an MH-DIF statistic $\hat{D}_i = -2.35\hat{\beta}_{MH_i} = 1.295$. The positive sign of the MH statistic indicates that the DIF is against LD students (focal group). Also, the 95% confidence intervals are estimated as [.456, .729] and [–.785, –.316] for $\hat{\alpha}_{MH_i}$ and $\hat{\beta}_{MH_i}$, respectively, which further provides the 95% confidence interval of [.743, 1.845] for $\hat{D}_i$. This confidence interval indicates that $\hat{D}_i$ is significantly different from 0, but not significantly different from 1. Therefore, according to Holland and Thayer's criteria, the item is classified as a "B item", which is flagged as a possible biased item toward LD students. A significance test based on MH_{X^2} can also be conducted. In this example, $MH_{X^2} = 20.593 (df = 1, p < .001)$, indicating the existence of DIF. However, one should note that the chi-square test is sensitive to sample size, and other measures, such as MH-DIF, should be considered as a primary criterion for detecting DIF.

1. Version 11.5 was used for illustrative data analyses in this paper.

Figure 3 *A sample output of the Mantel-Haenszel procedure by SPSS*

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
pes * m1 * mtot	3696	100.0%	0	.0%	3696	100.0%

Tests of Homogeneity of the Odds Ratio

	Chi-Squared	df	Asymp. Sig. (2-sided)
Breslow-Day	15.791	26	.941
Tarone's	15.789	26	.941

Tests of Conditional Independence

	Chi-Squared	df	Asymp. Sig. (2-sided)
Cochran's	21.288	1	.000
Mantel-Haenszel	20.593	1	.000

Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected is 0.

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			.577
ln(Estimate)			-.551
Std. Error of ln(Estimate)			.120
Asymp. Sig. (2-sided)			.000
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	.456
		Upper Bound	.729
	ln(Common Odds Ratio)	Lower Bound	-.785
		Upper Bound	-.316

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption. So is the natural log of the estimate.

For each of remaining items, the common odds ratio is obtained by SPSS and the MH-DIF is calculated. The results of the first phase of DIF analysis is shown in Table 2. When considering the values of $\hat{D}_i$, no items meet Holland and Thayer's criteria for large DIF, since $|\hat{D}_i| < 1.5$ for all items. Four items are between 1.0 and 1.5 in

absolute magnitude: item 1, item 2, item 25, and item 30. Items with positive $\hat{D}_i$ values (items 1, 2, and 25) indicate DIF against the focal group (LD students), while the item with negative $\hat{D}_i$ value (item 30) indicate DIF against the reference group (students in standard curricula).

At this point, we suspect these four items as showing DIF, and enter into a purification process. Following Holland and Thayer's (1988) recommendation, for each DIF item, a new total score is calculated by taking the previous total score, and subtracting the DIF items with the exception of the DIF item of interest. As an example, for item 2, the total score is adjusted by subtracting the scores of items 1, 25, and 30, but not item 2. For non-DIF items, the total score is adjusted by subtracting all DIF items from the original sum. The MH procedure reanalyzes all 40 items with these new totals. The results after purification are presented in Table 3. Since it resulted in no additional DIF items, the purification process stops here. Therefore, our conclusion is that items 1, 2, 25 are suspected being biased toward LD students, while item 30 is suspected as biased toward students in standard curricula. DIF against students in standard curricula can be interpreted that testing accommodations may have given unfair advantage. Note that in case additional items are identified as showing DIF, the purification process continues by adjusting the total test scores further. The process continues until no additional items are found to show DIF.

Table 2 *Initial Results of Mantel-Haenszel Analysis on 40 Math Test Items*

Item	χ^2	$\hat{\alpha}_{MH_i}$	*95% CI for* $\hat{\alpha}_{MH_i}$	$\hat{D}_i$	*95% CI for* $\hat{D}_i$
1	20.59**	0.58	0.46, 0.73	1.28	0.74, 1.82
2	27.74**	0.61	0.51, 0.73	1.16	0.74, 1.58
3	0.45	0.95	0.81, 1.11	0.12	-0.25, 0.50
4	1.05	0.92	0.79, 1.07	0.20	-0.16, 0.55
5	2.07	1.14	0.96, 1.36	-0.31	-0.72, 0.10
6	0.26	1.05	0.88, 1.25	-0.11	-0.52, 0.30
7	3.33	0.86	0.73, 1.01	0.35	-0.02, 0.74
8	20.68**	1.45	1.24, 1.71	-0.87	-1.26, -0.51
9	3.13	0.87	0.75, 1.01	0.33	-0.02, 0.68
10	15.24**	1.52	1.23, 1.86	-0.98	-1.46, -0.49
11	0.88	1.08	0.93, 1.26	-0.18	-0.54, 0.17
12	4.49*	0.84	0.72, 0.99	0.41	0.02, 0.77
13	5.05*	0.83	0.70, 0.97	0.44	0.07, 0.84
14	0.00	1.00	0.84, 1.18	0.00	-0.39, 0.41
15	0.33	1.05	0.90, 1.23	-0.11	-0.49, 0.25
16	11.55**	1.35	1.14, 1.60	-0.71	-1.10, -0.31
17	0.47	0.95	0.81, 1.10	0.12	-0.22, 0.50
18	6.44*	1.25	1.06, 1.48	-0.52	-0.92, -0.14
19	5.70*	0.82	0.69, 0.96	0.47	0.10, 0.87
20	4.51*	1.18	1.02, 1.38	-0.39	-0.76, -0.05
21	3.15	1.16	0.99, 1.37	-0.35	-0.74, 0.02
22	20.37**	0.70	0.59, 0.81	0.84	0.50, 1.24
23	9.89**	1.29	1.10, 1.51	-0.60	-0.97, -0.22

Table 2 continued *Initial Results of Mantel-Haenszel Analysis on 40 Math Test Items*

24	0.69	0.92	0.77, 1.10	0.20	-0.22, 0.61
25	30.44**	0.63	0.54, 0.74	1.09	0.71, 1.45
26	1.28	1.10	0.94, 1.28	-0.22	-0.58, 0.15
27	0.39	0.95	0.82, 1.10	0.12	-0.22, 0.47
28	9.28**	1.29	1.10, 1.52	-0.60	-0.98, -0.22
29	1.13	0.91	0.77, 1.08	0.22	-0.18, 0.61
30	43.34**	1.86	1.54, 2.24	-1.46	-1.90, -1.01
31	0.48	0.94	0.79, 1.11	0.15	-0.25, 0.55
32	2.49	0.86	0.72, 1.03	0.35	-0.07, 0.77
33	4.39*	0.83	0.71, 0.98	0.44	0.05, 0.80
34	16.62**	1.42	1.20, 1.68	-0.82	-1.22, -0.43
35	1.82	0.90	0.77, 1.05	0.25	-0.11, 0.61
36	5.33*	0.81	0.68, 0.96	0.50	0.10, 0.91
37	0.40	1.06	0.90, 1.25	-0.14	-0.52, 0.25
38	10.91**	1.35	1.13, 1.60	-0.71	-1.10, -0.29
39	0.20	0.96	0.82, 1.13	0.10	-0.29, 0.47
40	0.71	0.93	0.79, 1.09	0.17	-0.20, 0.55

* $p < .05$; ** $p < .01$

Table 3
Results of Mantel-Haenszel Analysis on 40 Math Test Items After Purification

Item	χ^2	$\hat{\alpha}_{MH_i}$	*95% CI for* $\hat{\alpha}_{MH_i}$	$\hat{D}_i$	*95% CI for* $\hat{D}_i$
1	24.24**	0.56	0.44, 0.70	1.36	0.84, 1.93
2	33.98**	0.58	0.49, 0.70	1.28	0.84, 1.68
3	0.42	0.95	0.81, 1.11	0.12	-0.25, 0.50
4	1.73	0.90	0.78, 1.05	0.25	-0.11, 0.58
5	1.02	1.10	0.92, 1.30	-0.22	-0.62, 0.20
6	0.00	1.01	0.85, 1.19	-0.02	-0.41, 0.38
7	3.61	0.85	0.73, 1.00	0.38	0.00, 0.74
8	19.30**	1.43	1.22, 1.67	-0.84	-1.21, -0.47
9	3.24	0.87	0.75, 1.01	0.33	-0.02, 0.68
10	13.55**	1.48	1.21, 1.81	-0.92	-1.39, -0.45
11	0.58	1.07	0.91, 1.24	-0.16	-0.51, 0.22
12	5.84*	0.82	0.71, 0.96	0.47	0.10, 0.80
13	8.15*	0.79	0.67, 0.93	0.55	0.17, 0.94
14	0.00	0.99	0.84, 1.18	0.02	-0.39, 0.41
15	0.07	1.02	0.88, 1.20	-0.05	-0.43, 0.30
16	10.38**	1.33	1.12, 1.57	-0.67	-1.06, -0.27
17	0.80	0.93	0.80, 1.08	0.17	-0.18, 0.52
18	5.54*	1.22	1.04, 1.45	-0.47	-0.87, -0.09
19	8.83*	0.78	0.66, 0.92	0.58	0.20, 0.98
20	4.16*	1.17	1.01, 1.37	-0.37	-0.74, -0.02
21	2.55	1.15	0.97, 1.35	-0.33	-0.71, 0.07

Table 3 continued

Results of Mantel-Haenszel Analysis on 40 Math Test Items After Purification

22	22.34**	0.69	0.59, 0.80	0.87	0.52, 1.24
23	10.31**	1.29	1.11, 1.51	-0.60	-0.97, -0.25
24	2.20	0.87	0.73, 1.04	0.33	-0.09, 0.74
25	31.73**	0.63	0.53, 0.74	1.09	0.71, 1.49
26	1.23	1.10	0.94, 1.28	-0.22	-0.58, 0.15
27	0.92	0.93	0.80, 1.08	0.17	-0.18, 0.52
28	7.56**	1.26	1.07, 1.48	-0.54	-0.92, -0.16
29	2.02	0.88	0.75, 1.04	0.30	-0.09, 0.68
30	42.55**	1.84	1.53, 2.21	-1.43	-1.86, -1.00
31	0.34	0.95	0.80, 1.12	0.12	-0.27, 0.52
32	2.45	0.86	0.72, 1.03	0.35	-0.07, 0.77
33	4.07*	0.84	0.71, 0.99	0.41	0.02, 0.80
34	15.01**	1.39	1.18, 1.64	-0.77	-1.16, -0.39
35	1.88	0.90	0.77, 1.04	0.25	-0.09, 0.61
36	5.30*	0.81	0.68, 0.96	0.50	0.10, 0.91
37	0.11	1.03	0.88, 1.22	-0.07	-0.47, 0.30
38	10.39**	1.33	1.12, 1.58	-0.67	-1.07, -0.27
39	0.95	0.92	0.79, 1.08	0.20	-0.18, 0.55
40	1.31	0.91	0.78, 1.07	0.22	-0.16, 0.58

* $p < .05$; ** $p < .01$

Logistic Regression Method

Rogers and Swaminathan (1990) proposed the use of a logistic regression model to detect DIF between two sub-samples of examinees for a dichotomously scored item. Group membership, matching criterion (e.g., total test score), and an interaction effect between group and matching criterion are considered as independent variables in the model, while item response (0 = incorrect, 1 = correct) is considered as the dependent variable. The basis of this model focuses on the main effect of group and the interaction effect between group membership and matching criterion, conditioned on matching criterion. The main effect of group represents uniform DIF, while the interaction effect represents non-uniform DIF. The difference between uniform and non-uniform DIF can be explained graphically. Figure 4a depicts uniform DIF, where two curves have the same slope but are different in their locations. As a result, they are parallel and are visually separated from each other. In this example, the focal group falls below the reference group. This shows possible DIF against the focal group which is consistent across all levels of the matching criterion. Figure 4b depicts non-uniform DIF. In this case, the curves have different slopes and cross (showing evidence of non-uniform DIF), while the location of the curve is the same (showing no uniform DIF). In this example, there is DIF against the reference group for low levels of the matching criterion, and DIF against the focal group for higher levels of the matching criterion. Finally, Figure 4c shows an example of both uniform and non-uniform DIF. The curves are different in their slopes and location. As a result, they are visually separated at certain levels of the matching criterion, and also cross.

Graphical representations of uniform and non-uniform DIF

Figure 4. *(a) Only uniform Dif*

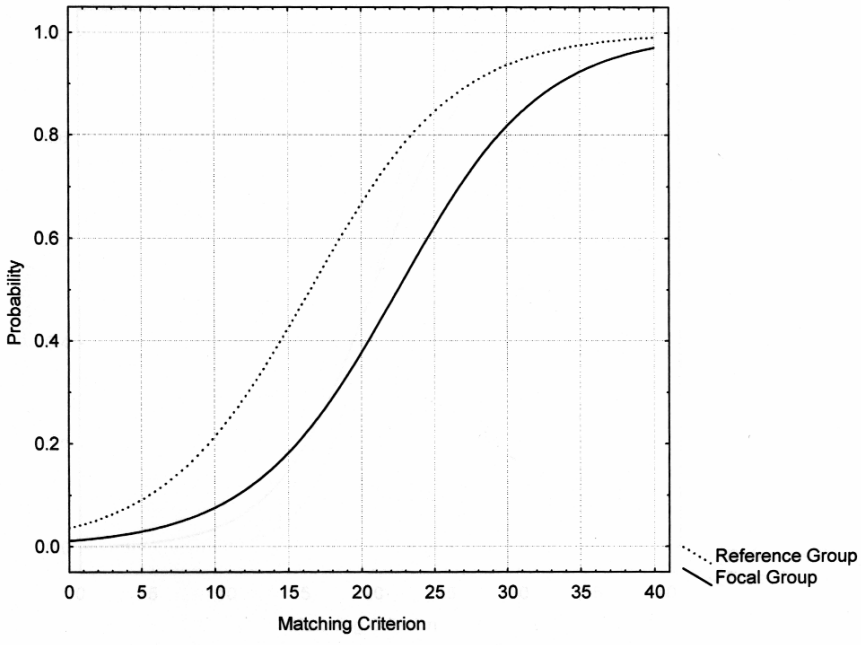


Figure 4. (b) *Only non-uniform DIF*

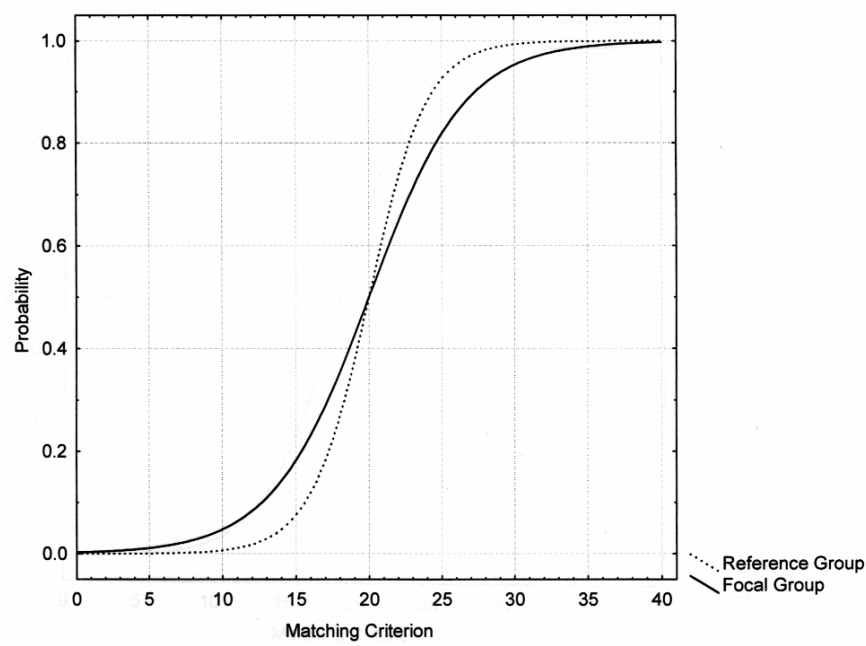
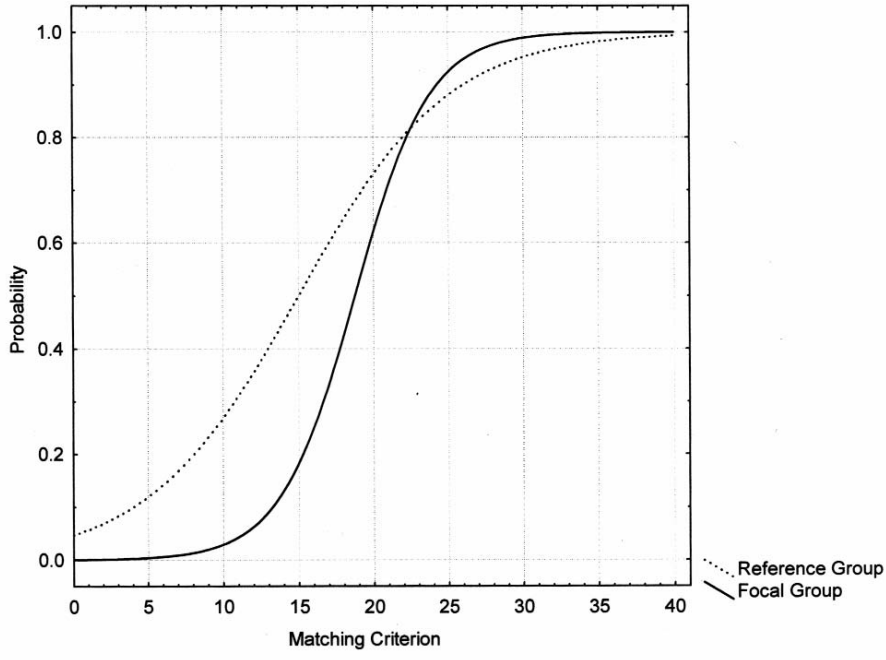


Figure 4. (c) *Both uniform and non-uniform DIF*



The logistic regression model for each item is expressed by

$$\ln\left[\frac{p_j}{(1-p_j)}\right] = b_0 + b_1 X_j + b_2 G_j + b_3 (XG)_j ,$$

where p_j is the probability of individual j to get the item correct, χ_j is the value of matching criterion for individual j, G_j represents the group membership for individual j, and $(XG)_j$ is the interaction between X_j and G_j for individual j. The left hand-side of the equation, $\ln[p_j/(1-p_j)]$ is a quantity called the log-odds-ratio. An unbiased item is indicated if b_2 and b_3 are not different from 0. Note that a subscript for item is dropped from the equation for notational simplicity.

A DIF analysis first considers only the matching criterion (X_j) in the model (1st model). Then, the main effects of the matching criteria (X_j) and the group variable (G_j) are considered and the model is reanalyzed (2nd model). Finally, the interaction effect is added to the model, and the model is reanalyzed (3rd model). A Chi-square statistic is derived for each model based on the quantity called log-likelihood ratios.

In order to test for DIF, a quantity χ^2_{DIF} is computed by

$$\chi^2_{DIF} = \chi^2_{3rd\ model} - \chi^2_{1st\ model},$$

which is the difference between two Chi-square statistics from the 3rd and 1st models.

χ^2_{DIF} is distributed as a Chi-square distribution with 2 degrees of freedom. It is a simultaneous test of both uniform and non-uniform DIF. Therefore, if the test is not significant, we conclude that the item does not show either uniform or non-uniform DIF. On the other hand, if the test is significant, we conclude that we have enough evidence for either uniform or non-uniform DIF (or both). The magnitude of DIF

Figure 5 *A sample output of the logistic regression procedure by SPSS*

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	944.017	1	.000
	Block	944.017	1	.000
	Model	944.017	1	.000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	mtot	.229	.010	541.378	1	.000	1.257
	Constant	-2.145	.142	227.052	1	.000	.117

a. Variable(s) entered on step 1: mtot.

Block 2: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	22.760	1	.000
	Block	22.760	1	.000
	Model	966.777	2	.000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	mtot	.214	.010	440.620	1	.000	1.238
	pes	-.556	.118	22.144	1	.000	.574
	Constant	-1.552	.188	68.143	1	.000	.212

a. Variable(s) entered on step 1: pes.

Block 3: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	.792	1	.373
	Block	.792	1	.373
	Model	967.569	3	.000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	mtot	.202	.017	146.612	1	.000	1.224
	pes	-.833	.333	6.262	1	.012	.435
	mtot by pes	.019	.021	.800	1	.371	1.019
	Constant	-1.361	.284	22.948	1	.000	.257

a. Variable(s) entered on step 1: mtot * pes .

can be evaluated by the estimates of b_2 and b_3 in the same way as in the MH method. Since b_2 and b_3 are in the same scale as $\hat{\beta}_{MH_i}$ (in the scale of log-odds-ratio), multiplying b_2 and b_3 by –2.35 will transform them into the same scale as $\hat{D}_i$. Similarly, b_2 and b_3 can be transformed into the same scale as $\hat{\alpha}_{MH_i}$ by exponentiating them.

We apply a logistic regression approach to the first item of the statewide testing program data. The matching criterion is obtained by the total score in the same way as in the previous example with MH procedure. Also, the group indicator variable is coded in the same way (0 = students in standard curricula, 1 = LD students). Parts of the logistic regression procedure output in SPSS are displayed in Figure 5, and the results are summarized in Table 4. Running the 1st model against the 3rd model gives $\chi^2_{DIF} = 967.57 - 944.02 = 23.55$ with 2 degrees of freedom. This value is significant at the 0.01 level, indicating the presence of DIF.

Table 4 *Logistic Regression Results for Math Item 1 in the Statewide Testing Program Data*

	Model 1	Model 2	Model 3
$\chi^2_{(1)}$	944.02**	966.78**	967.57**
Model Coefficients			
Intercept	–2.145	–1.552	–1.361
Test Score X_j	0.23**	0.21**	0.20**
Group G_j		–0.56**	–0.83*
		(1.32)	(1.95)
Interaction $(XG)_j$			0.02
			(–0.05)

* p < .05; ** p < .01

Note: Values in parentheses indicate transformed valued in the same scale as $\hat{D}_i$ in the MH method by multiplying the estimated coefficient by –2.35.

To assess the type of DIF, we note that for uniform DIF,

$$\chi^2_{\textit{uniform DIF}} = \chi^2_{\textit{2nd model}} - \chi^2_{\textit{1st model}} = 966.78 - 944.02 = 22.76$$

with 1 degree of freedom, and is significant at the 0.01 level. This indicates that the item has characteristics of uniform DIF. The Group coefficient (b_2) is significantly different from 0 and estimated as –.83 in the final model. It can be translated into $-.83 \times (-2.35) = 1.95$, in the same scale as $\hat{D}_i$, indicating a possible bias against the focal group (LD students). For assessment of non-uniform DIF, we note that

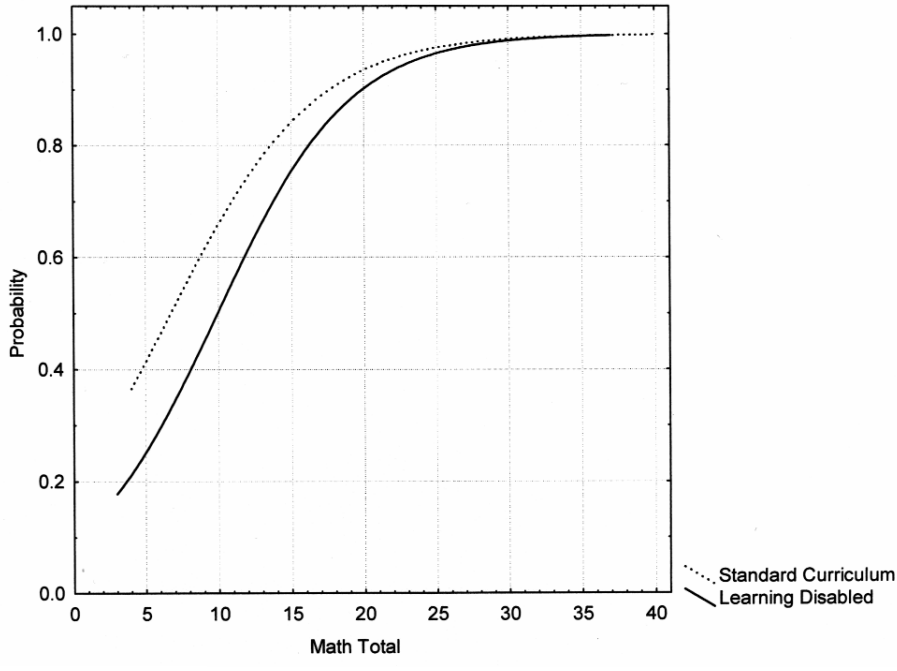
$$\chi^2_{\textit{non-uniform DIF}} = \chi^2_{\textit{3rd model}} - \chi^2_{\textit{2nd model}} = 967.57 - 966.78 = 0.79$$

with 1 degree of freedom, which is not significant at $\alpha = .05$. Thus, there is no indication of non-uniform DIF for this item. The coefficient for the interaction effect (b_3) is estimated as .02, which can be translated into $.02 \times (-2.35) = -.05$ in the same scale as $\hat{D}_i$. One may also observe that the interaction effect is not significantly different from 0 in the model. We see that in the case of item 1, DIF exists yet is consistent across varying ability levels. The results are comparable to the chi-square test in the MH procedure, yet with more detail on the two types of DIF. Graphically, a logistic curve that represents the relationship between the matching criterion and the probability of correct answer is drawn for each group separately in Figure 6.

They are different in locations (indication of uniform DIF) but very close to parallel because of small but non-zero b_3.

Unlike the MH procedure, the logistic regression approach can investigate both uniform and non-uniform DIF. Also, the logistic regression approach can handle continuous matching criterion variable, such as ability estimates from item response theory models. Furthermore, it can be extended to polytomously scored test items quite easily (see e.g., Zumbo, 1999), although an attempt to extend MH method to polytomous items has been made (e.g., Zwick, Donoghue, & Grima, 1993).

Figure 6 *Graphical representation of DIF for the first item in the statewide testing data*



Analysis Beyond Detecting DIF

This paper focused on detecting DIF items. However, recent studies demonstrated possibilities to go beyond detecting DIF items and obtain additional information about DIF items. Although these types of analyses are not extensively discussed or demonstrated in this paper, a brief discussion is provided.

When the magnitudes of DIF are compared between test items, some items may show larger DIF magnitude, while some others show relatively small DIF magnitudes. In such a situation, it may be of interest to researchers to investigate sources of such a variation. Bolt (2000), for example, found that multiple-choice items had more DIF characteristic than constructive-response items between males and females on SAT math pretest items. Another example, Walker and Beretvas (2001) found that DIF between proficient writers and non-proficient writers were significant only for constructed-response items that required writing about their solutions.

These results can possibly provide suggestions that may be informative to minimize DIF items in the future by many different means, including instruction, policy, and test construction. These types of analyses are possible by logistic regression approach (e.g., Swanson, Clauser, Case, Nungester, & Featherman, 2002) and multidimensional IRT approach (e.g., Roussos & Stout, 1996).

Similarly, there is a possibility that the magnitude of DIF varies across group units, such as schools. This is of particularly strong interest in DIF studies for the test accommodated subpopulation, because each school typically makes its own decision on which students shall be provided with testing accommodations. Also, it is very likely that the degree by which accommodated students benefit from accommodations depends on how much they use such accommodations, which in turn may reflect how serious schools are concerning testing accommodations. This type of information will be useful not only for detecting DIF, but also providing useful suggestive information for many audiences, including school administrators and policy makers. As one attempt, Yanling (2002) proposed a two-way factorial DIF analysis taking into account the interaction of gender DIF and ethnicity. Yanling's two-way factorial DIF analysis can easily incorporate one group characteristic variable in order to detect group-varied DIF as a three-way interaction between item difficulty, person characteristic variable, and a group characteristic variable. However, this approach is limited to the addition of only one categorical person- and group-characteristic variable. Some authors demonstrated that DIF can be parameterized in the framework of generalized linear mixed model (also known as hierarchical generalized linear model) (e.g., Kamata & Binici, 2003; Luppescu, 2002; Meulders & Xie, 2004). This approach can be easily extended to a "multilevel" DIF analysis model that allows one to incorporate one or more continuous or categorical person- and group-characteristic variables, which also takes into account both within-group and between-group variations.

Summary

This paper presented detailed procedures for DIF detections by two widely used approaches. Also, a brief discussion on more recent direction of DIF study that goes beyond detections of DIF items was provided. The motivation for studying DIF might be global (as in the case of studying differences in trait manifestation) or more detailed (as in the case of test construction). Whatever the reason, the use of DIF analysis gives the researcher richer insights into the issue of bias for continuing improvement in inclusion and testing of LD students in educational practice. Therefore, it is highly suggested that researchers on learning disability students consider utilizing DIF analysis when a test is constructed or a test score is used as a variable in their research.

Akihito Kamata *is an Associate Professor of Educational Measurement and Statistics at Florida State University. His research interests include multilevel modeling and measurement theory.* ***Brandon K. Vaughn*** *is a doctoral student in Measurement and Statistics at Florida State University. His research interests are applied statistics and measurement theory.*

References

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67-91.

Angoff, W.H., & Ford, S.F. (1973). Item race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10,* 95-105.

Bolt, D. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement, 37,* 307-327.

Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.

Florida Department of Education (2004). FCAT 2004 reading, mathematics, and science test administration manual. Tallahassee, FL.

Holland, P.W. & Wainer, H. (Eds.) (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum.

Holland, P.W., & Thayer, S.T. (1985). *An alternative definition of the ETS delta scale of item difficulty.* Princeton, NJ: Educational Testing Service, Research Report RR-85-43.

Holland, P.W., & Thayer, S.T. (1988). Differential item performace and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.) *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Kamata, A. & Binici, S. (2003). *Random effect DIF analysis via hierarchical generalized linear modeling.* Paper presented at the annual International Meeting of the Psychometric Society, Sardinia, Italy.

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika, 54,* 681-697.

Luppescu, S. (2002). *DIF detection in HLM item analysis.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), E*ducational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Meulders, M., & Xie, Y. (2004). Person-by-item predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 213-240). New York: Springer.

Millsap, R.E., & Everson, H.T. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 17,* 297-334.

Muraki, E. & Engelhard, G. (1989) *Examining differential item functioning with BIMAIN.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Muthen, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557-585.

Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research, 71,* 53-104.

Plake, B.S. (1981). An ANOVA methodology to identify biased test items that takes instructional level into account. *Educational and Psychological Measurement, 41,* 365-368.

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

Raju, N.S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 495-502.

Rogers, H.J. & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17,* 105-116.

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

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

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Swanson, D.B., Clauser, B.E., Case, S.M., Nungester, R.J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational & Behavioral Statistics, 27,* 53-75.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

U.S. Department of Education. (2002). *Draft regulations to implement Part A of Title I of the Elementary and Secondary Education Act of 1965 as amended by the No Child Left Behind Act of 2001.* Washington, DC.

Walker, C.M., & Beretvas, S.N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement, 38,* 147-163.

Yanling, Z. (2002). *DIF in a large scale mathematics assessment: The interaction of gender and ethnicity.* Paper presented at the annual meeting of the American Education Research Association, New Orleans, LA.

Zwick, R., Donoghue, J.R., & Grima, A. (1993). Assessing differential item functioning in performance tasks. *Journal of Educational Measurement, 30,* 233-251.

Zumbo, B.D. (1999). *A handbook on the theory and methods for differential item functioning: Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Received July 17, 2004
Revision received July 24, 2004
Accepted July 30, 2004

APPENDIX
PROCEDURE FOR DIF ANALYSES ON SPSS

Data File

The data file should contain a) a group indicator variable, b) item responses for each item, and c) matching criterion variable. In the illustrative data analyses, the group-indicator variable (0 = standard curricula, 1 = specific learning disabled) is labeled as pes. The variables that represent item responses (0 = incorrect, 1 = correct) are labeled as m1 through m40. The matching criterion is labeled as mtot1 as the sum of m1 through m40 for the first step in the purification process. In addition, an interaction variable for each item is created by multiplying the group-indicator variable and the matching criterion variable, and they are labeled as int1. For the second step in the purification process, for each DIF item, a new total score is calculated by taking the previous total score, and subtracting the DIF items with the exception of the DIF item of interest. For non-DIF items, the total score is adjusted by subtracting all DIF items from the original sum. As a result, 40 separate total scores are generated. Accordingly, interaction variables are also recomputed for all 40 items. Although all observed scoring categories between 0 and 40 are used as matching criterion, they can be collapsed into a smaller number of categories. This may be reasonable when the number of scoring categories is large relative to the sample size.

Mantel-Haenszel Method

The Crosstabs command under Descriptive submenu is used. The dialog box for Crosstabs procedure is presented in Figure A1. Notice that two dichotomous variables (pes and m1) are selected as the row and column variables, and the matching criterion (mtot1) is selected as the layer variable. Then, by clicking the Statistics button, another dialog box like Figure A2 will appear. Check mark the Cochran's and Mantel-Haenszel statistics option, and specify the value for Test common odds ratio equals 1. (Please see page 68, Figures A1 and A2) Alternatively, the following SPSS syntax will do the same analysis.

```
CROSSTABS
    /TABLES = m1 BY pes BY mtot1
    /STATISTIC = CMH(1).
```

Logistic-Regression Method

The Logistic Regression procedure is used. This procedure is selected by the Logistic Binary command under Regression submenu. In Figure A3, m1 is selected as a Dependent variable, and

mtot1 is selected as a covariate. This is a specification for the 1st model. Then, click on the next button to specify the 2nd model (see Figure A4). Here, m1 is still the dependent variable, and mtot1 and pes are specified as covariates. Lastly, specify the 3rd model by clicking the next button and specifying mtot1, pes, and int1 as covariates (see Figure A5). (Please see page 69, Figures A3, A4, and A5.) Alternatively, the following SPSS syntax will do the same analysis.

```
LOGISTIC REGRESSION VAR = m1
    /METHOD = ENTER mtot1
    /METHOD = ENTER mtot1 pes
    /METHOD = ENTER mtot1 pes int1.
```

Figure A1. *A dialog box for the Mantel-Haenszel procedure by SPSS – 1*

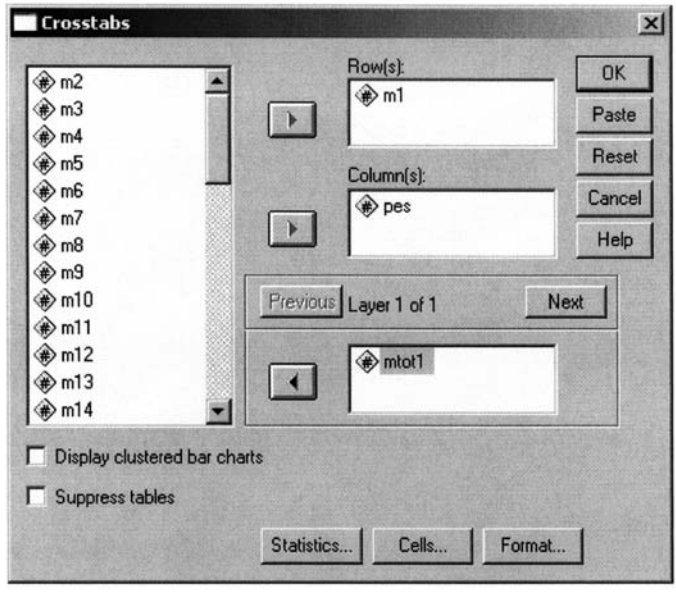


Figure A2. *A dialog box for the Mantel-Haenszel procedure by SPSS – 2*

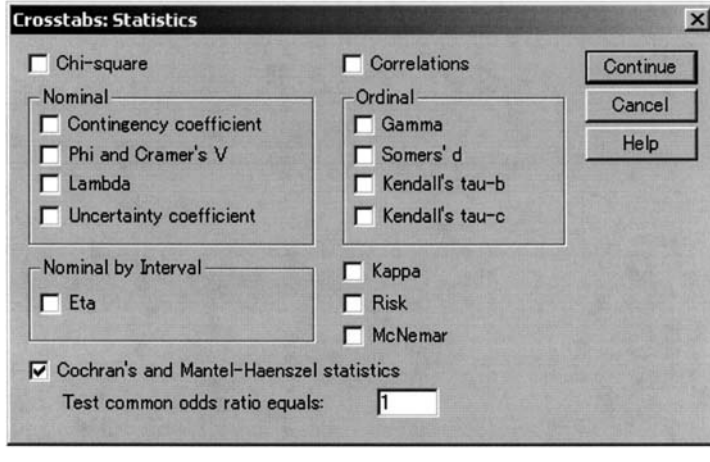


Figure A3. *A dialog box for the logistic regression procedure by SPSS – 1*

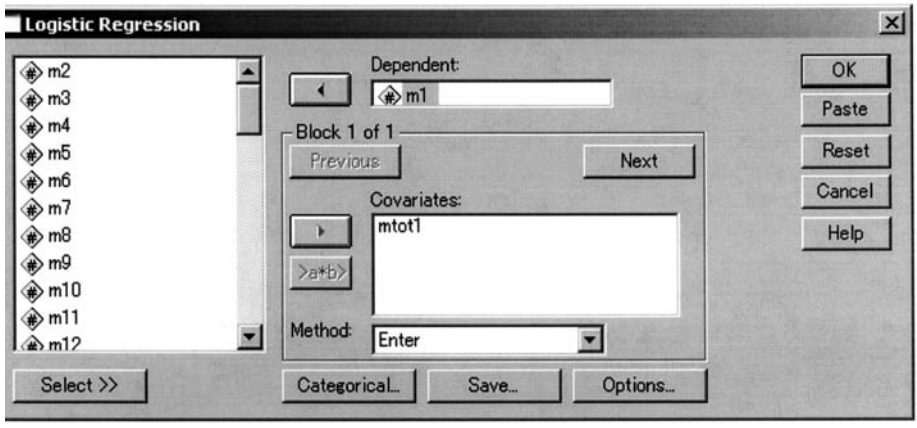


Figure A4. *A dialog box for the logistic regression procedure by SPSS – 2*

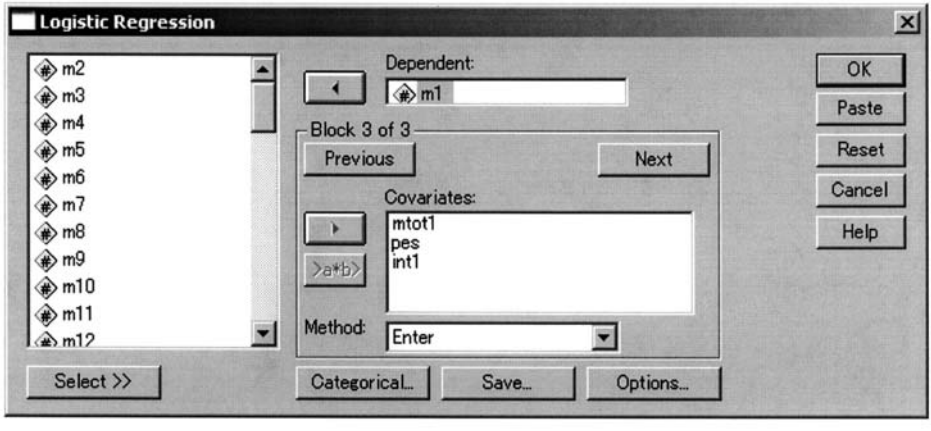


Figure A5. *A dialog box for the logistic regression procedure by SPSS – 3*