2013

An Applied Examination of Methods for Detecting Differential Distractor Functioning

Sharon Koon and Akihito Kamata

An applied examination of methods for detecting differential distractor functioning

Sharon Koon*

Florida State University,
2010 Levy Avenue, Suite 100,
Tallahassee, Florida 32310, USA
E-mail: skoon@fcrr.org
*Corresponding author

Akihito Kamata

University of Oregon,
102S Lokey Education Building,
Eugene, Oregon 97403-5267, USA
E-mail: kamata@uoregon.edu

Abstract: This study applied the odds-ratio (Penfield, 2008), the multinomial logistic regression (Kato et al., 2009), and the standardised distractor analysis (Schmitt and Bleistein, 1987) methods in the examination of differential distractor functioning (DDF) effects. Using data from the administration of one statewide mathematics assessment, these methods were applied to provide insight into two research questions; 1) whether the magnitude and pattern of the DDF effect is constant across all methods; 2) whether the pattern of DDF effects supports differential item functioning (DIF) findings. While some differences in results were found, all three methods present a viable option for use in improving test items included in statewide assessment programmes.

Keywords: differential item functioning; DIF; differential distractor functioning; DDF.

Reference to this paper should be made as follows: Koon, S. and Kamata, A. (xxxx) 'An applied examination of methods for detecting differential distractor functioning', *Int. J. Quantitative Research in Education*, Vol. X, No. Y, pp.000–000.

Biographical notes: Sharon Koon is an Associate in Research at the Florida Center for Reading Research at Florida State University. Her primary research interests relate to the development and analysis of statewide assessment programmes through the use of psychometrics.

Akihito Kamata is Professor and Department Head of the Department of Educational Methodology, Policy, and Leadership at the University of Oregon. His primary research interest is psychometrics, focusing on development and implementation of item-level test data analysis methodology through item response theory modelling, multilevel modelling and structural equation modelling.

1 Introduction

It is estimated that the development costs for one item on a high-stakes, statewide assessment average $1,800 to $2,000, and this development process takes at least two years (Florida Department of Education, 2009). Prior to an item contributing to a student's test score, the validity of the item must be demonstrated. This validity is established through both statistical reviews, based on field-testing, and expert committee reviews.

Despite these safeguards, not all items perform in expected ways. In some cases, items are found to 'behave differently' among groups, after controlling for ability. This behaviour is known as differential item functioning (DIF). If an item is found to have DIF, there is a statistical indication that the item may be biased, warranting further review of the item.

Items that are identified as having DIF are evaluated for the magnitude of DIF and the type of DIF. DIF in dichotomously scored items is said to occur in two ways – uniform and non-uniform. Uniform DIF occurs when the magnitude of the DIF is constant across ability levels, unlike non-uniform DIF, which occurs when the magnitude is not constant across ability levels (e.g., Kamata and Vaughn, 2004). The magnitude of DIF is typically considered in categories of small, medium, and large, using different effect-size criteria depending on the method used. If the magnitude is small, little or no action is taken. If the magnitude is large, it is recommended that the item be removed from future test construction until it can be reviewed and judged as being an appropriate item. In some cases, the item is revised and field tested again. Central to this revision process is the identification of a possible reason for the DIF.

To identify possible item factors that may contribute to DIF, Schmitt and Bleistein (1987) extended the standardisation approach to assessing DIF developed by Dorans and Kulick (1983). This extension was ultimately referred to as the standardised distractor analysis (STD) method. STD extends the standardisation approach by analysing all distractors, not reached, and omits, instead of only analysing differences between the correct and the total of all incorrect responses. Similarly, Green et al. (1989) advocated the use of information that can be garnered by examining an examinee's distractor choice. They proposed a log-linear method for studying examinee responses for group differences in distractor selection rates, which they named differential distractor functioning (DDF). Penfield (2010) showed, using an odds-ratio method (OR; Penfield, 2008) that DIF may be partially explained by studying examinee responses to item distractors.

The evaluation of DDF effects continues to be relevant to the study of DIF. Middleton and Laitusis (2007) used STD to determine if there were differences in distractor functioning between students with disabilities, with and without accommodations, in comparison to students without disabilities. A multi-step logistic regression procedure was used by Abedi et al. (2008) to explore differential trends in the selection of one of three distractors by students with disabilities. Kato et al. (2009) proposed the use of a multi-step multinomial logistic regression (MLR) approach as an extension of the logistic regression approach used by Abedi et al. (2008) to simultaneously evaluate DIF and DDF in reading assessments for students with disabilities.

Building upon this research, this study applied the OR (Penfield, 2008), the MLR (Kato et al., 2009), and the STD (Schmitt and Bleistein, 1987) methods for measuring

DDF effects to data collected from the administration of a statewide mathematics assessment. Measures of DIF statistical significance, when applicable, and effect size are reported. In addition, the relationship between DIF and DDF effects are examined for each of the three methods, as well as an exploration into whether item characteristics explain the DDF effects.

2 Methods

Data collected from the 2006 administration of the Grade 3 Florida Comprehensive Assessment Test® (FCAT) Mathematics were used in this study. The FCAT is Florida's statewide criterion-referenced assessment test in the subjects of reading, mathematics, science, and writing. In addition to state and federal accountability stakes, the FCAT is a high stakes test for students because student promotion decisions also are tied to FCAT results.

Of the 206,678 cases in the data file, 202,140 cases were used in this study after an analysis of missing data, with a distribution of 46% White, 25% Hispanic, 23% African-American, 2% Asian, 4% Multiracial, and less than 1% American Indian. Approximately 54% of these students were served by free or reduced-priced lunch programmes.

Data in the file included student responses to the same 40 test items included in the analyses (known as core items). While there are multiple test forms used in Florida, each test form contained the same core items. All test items consisted of four options. The option chosen for each test item was captured in the file, as well as an indicator of whether the response was correct (coded as 1) or incorrect (coded as 0).

This specific assessment was chosen because it was released to the public as an example. When an assessment is released to the public, all of the test items that contribute to a student's score (the core items) from the test booklet are released and those items that do not contribute to a student's score (field test and anchor items used in equating) are not released. Because the test items included in this study have been released, they are used freely in the appropriate sections to inform the results.

The 2006 Grade 3 FCAT Mathematics test design included the following specifications (Florida Department of Education, 2008):

- content categories – number sense, concepts, and operations (30%); measurement (20%); geometry and spatial sense (17%); algebraic thinking (15%); and data analysis and probability (18%)
- cognitive complexity – low (25–35%), moderate (50–70%), high (5–15%)
- 45–50 multiple-choice items, including field test or anchor items, depending on the test form.

FCAT Mathematics test item specifications (Florida Department of Education, 2005) require that all items "should not provide an advantage or disadvantage to a particular group of students" (p.2). The item development process includes several reviews for the purpose of identifying items that may not meet this requirement, including reviews for bias and community sensitivity. FCAT items that are accepted for placement on an assessment are field-tested and studied for statistical acceptability, including DIF. Items

that do not function appropriately may either be deleted or revised. Items that are revised must be reviewed and field-tested again. DDF analyses are not conducted.

The 3-parameter logistic (3-PL) IRT model is used for scaling FCAT multiple-choice items. The resulting item parameters are used in the determination of individual ability estimates, which are then placed on the FCAT score scale (Florida Department of Education, 2007). FCAT scale scores are reported on a scale of 100–500. FCAT achievement also is interpreted using achievement levels, with Level 1 the lowest level and Level 5 the highest level.

The standard error of measurement (SEM) of the 2006 Grade 3 FCAT Mathematics ranged across the five FCAT achievement levels, with larger SEM values at the lowest and highest achievement level cut scores. The IRT marginal reliability was .927, and the reliability as measured by Cronbach's alpha was .900 (Florida Department of Education, 2007).

All 40 items on the assessment were classified as having a small, or negligible, DIF rating, as measured by the standardised mean difference (SMD; Zwick et al., 1993). This finding indicates that, for the comparison groups studied (females versus males, African-American students versus White students, and Hispanic students versus White students), the DIF effect was found to have no practical importance, with a finding of statistical significance most likely due to the large sample size. It is expected that the DIF effect would be negligible for these comparison groups, since most items with larger DIF effects would have been removed during the test construction process.

2.1 DIF analyses

It has been proposed that the value of DDF analyses is primarily found as a supplement to a DIF analysis by item to investigate where in the test item the DIF may be occurring (Dorans et al., 1992; Penfield, 2010). Given this, a DIF analysis was conducted first. Logistic regression was used for this preliminary analysis because the method provides indices of both uniform and non-uniform DIF (Swaminathan and Rogers, 1990).

The logistic regression model for each item is expressed by

$$\ln\left[\frac{p_e}{(1-p_e)}\right] = \beta_0 + \beta_1 X_e + \beta_2 G_e + \beta_3 (XG)_e, \tag{1}$$

where p_e is the probability of examinee e to get the item correct, X_e is the value of the matching criterion for examinee e, G_e is the group indicator for examinee e, and $(XG)_e$ represents the interaction between X_e and G_e for examinee e. The FCAT scale score served as the matching criterion. To evaluate the magnitude of the DIF, the exponents of $\hat{\beta}_2$ and $\hat{\beta}_3$ were taken to place the values on the scale equivalent to the Mantel-Haenszel (MH) common odds ratio, referred to as $\hat{\alpha}_{MH_i}$.

2.2 DDF analyses

As applied by Schmitt and Bleistein (1987), STD was used to estimate DDF effects. The standardised p-difference, *STD P-DIF*, served as the index for assessing the direction and magnitude of the DDF for each distractor. *STD P-DIF* for distractor *j* was estimated by

$$STD\ P\text{-}DIF(j) = \sum_{s=1}^{S} w_s \left[P_{fs}(j) - P_{rs}(j) \right] \Big/ \sum_{s} w_s , \tag{2}$$

where

$$P_{fs}(j) = F_{js} \big/ n_{fs}$$

and

$$P_{rs}(j) = R_{js} \big/ n_{rs} .$$

$P_{fs}(j)$ and $P_{rs}(j)$ are the proportions of focal group, F, and reference group, R, members, respectively, selecting distractor j at each score level s. It is the magnitude of the difference in proportions, or p-difference, between both groups at each score level that is of concern. The standard weight, w_s, applied to both the focal and reference group proportions, was set equal to n_{fs}, the number of examinees at s in the focal group. The FCAT scale score was used as the matching criterion (s) and score levels were not grouped so that the information was maximised and consistent across all approaches.

The STD P-DIF values were transformed to the odds-ratio scale, an index referred to by Wright (1986) as α_{STD}. The magnitude of the odds ratio was aligned to the Educational Testing Service (ETS) categories of A, B, and C, by Monahan et al. (2007). Odds ratios greater than 1.89 are categorised as C+, while those less than 0.53 are C– effects. Odds ratios greater than 1.53 are categorised as B+, while those less than 0.65 are B- effects. Odds ratios greater than 1, but less than or equal to 1.53 are categorised as A+ effects, and those less than 1 but greater than or equal to 0.65 are A- effects. This classification was used in summarising the magnitude of the DDF effects.

The OR method proposed by Penfield (2008) served as the second method for estimating DDF effects. For the j^{th} contrast function (i.e., the contrast between examinees choosing the correct answer, coded as 1, and the j^{th} distractor), the conditional odds ratio across all FCAT scale score levels was estimated by

$$\hat{\alpha}_j = \frac{\sum_{s=1}^{S} R_{1s} F_{js} / n_s}{\sum_{s=1}^{S} R_{js} F_{1s} / n_s} . \tag{3}$$

Taking the natural logarithm of $\hat{\alpha}_j$, $\hat{\lambda}_j = \ln\left(\hat{\alpha}_j\right)$, provides an estimate of λ_j, the natural logarithm of the conditional odds ratio of the j^{th} contrast function. This transformation allows the interpretation of the DDF effect to be through a signed index where a value equal to 0 indicates no DDF for the j^{th} distractor, a positive value indicates that the DDF favours the reference group, and a negative value indicates DDF favouring the focal group.

The final method considered was the MLR method, proposed by Kato et al. (2009). DDF was analysed by fitting a MLR model with the correct choice as the base category

$$\ln\left[\frac{p_{je}}{\left(1 - p_{je}\right)} \right] = \beta_{j0} + \beta_{j1} X_e + \beta_{j2} G_e \tag{4}$$

where $j = 1, 2, \ldots, J$ and the log odds for each distractor was interpreted as the log odds to choose the distractor as compared to the correct answer.

Two models were compared. The first model included the ability measure (X_e) under the assumption that there were no group differences. The second model included both the ability measure and group indicator (G_e). As in the other methods, the FCAT scale score (s) was used as the matching criterion and score levels were not grouped.

The coefficient for the group indicator was used to indicate the magnitude of the DDF effects in the model. To evaluate the magnitude of the DDF effects, the exponent of $\hat{\beta}_{j2}$ was taken to place the values on the odds-ratio scale.

To judge the consistency of the DDF effect-size results for each distractor between the OR and STD approaches, and the MLR and STD approaches, correlation indices were used. These comparisons were made to the STD approach because it is a recognised procedure for DDF detection. To determine the consistency at the item level between STD and the other two approaches, DDF effects for each item were summarised after transforming the odds-ratio effect sizes to the log-odds ratio index. The summary data included the effect-size range, whether the effects were divergent in direction, and the mean effect size. Items were classified as having divergent distractor-level effects if the combination of effects included both a negative and positive effect among the three DDF effects within each item.

The statistical package IBM® SPSS® Statistics software (SPSS) 12 was used for the logistic regression, OR, and MLR analyses. SAS software (version 9.1) was used to generate item response frequencies which were then used in Microsoft Excel (version 97–2003) to calculate STD effects.

2.3 Relationship between DIF and DDF

Penfield (2010) found that a condition leading to uniform DIF is when there is a constant DDF effect across all distractors. In addition, he found that crossing DIF effects, which he distinguishes as non-uniform DIF effects that cancel due to differences in sign, can only occur in the presence of DDF effects that vary in sign. For items with significant uniform DIF, the item stem or correct option may be the source of the DIF effect. For items with significant non-uniform DIF, the distractors are likely the source of the DIF effect.

Using the logistic regression DIF results, items that were found to only have significant uniform DIF effects were investigated for a constant DDF effect across all distractors based on the magnitude of the range. All items identified as having significant non-uniform DIF in the logistic regression analysis were plotted to determine if the non-uniform DIF was crossing DIF. In addition to plotting, crossing DIF was investigated through determining the direction of the DIF effect at each FCAT achievement level cut point.

FCAT test items are classified by the several item characteristics, including the item content category, cognitive complexity, item difficulty, and item discrimination. While these classifications primarily are applied to the item and the correct response, they also apply to the item distractors. Using these item characteristics and the DDF item summary statistics (i.e., DDF effect-size range, mean effect size, divergence), an analysis was conducted to determine if relationships existed.

For items with large DIF effects identified through logistic regression, the utility of the information provided by the DDF analyses was explored with the assistance of a Florida Department of Education content expert. The exploration focused on the practical use of DDF results in the item revision process.

3 Results

Of the 206,678 cases in the 2006 Grade 3 FCAT Mathematics data file, a total of 3,871 cases were deleted from the data file because the demographic data collected during testing did not match the demographic data on the Florida Department of Education student database. These cases were deleted to ensure that the free or reduced-price lunch grouping indicator was measured with minimal error. In addition, 667 cases were deleted because an FCAT scale score, MATHSS, was not reported. For all analyses, students classified as participating in free or reduced-price lunch programmes (coded lunch = 1) served as the focal group, while students not participating in these programmes served as the reference group. The final sample included 202,140 cases.

An analysis of items with no responses indicated that the mean rate of no response was .324%, with a minimum of .06% and a maximum of 1.369%. These rates were considered minimal and not sufficient to conduct separate DDF analysis of 'no responses'. For the purpose of the DIF analysis, unanswered items were considered incorrect and included in the analysis.

3.1 DIF and DDF analyses

3.1.1 DIF results

Statistical evidence for DIF existed for all 40 items, which may be expected due to the large sample size. Thirty-four items had statistical evidence of small, uniform DIF (classified as A DIF). Thirty of these items also had statistical significance for non-uniform DIF, and 17 of these items displayed medium to large non-uniform DIF effects across the ability scale. The results for six items indicated only significant non-uniform DIF. For those items with significant non-uniform DIF, the magnitude of the non-uniform DIF at the four FCAT achievement level cut scores (253, 294, 346, and 398, respectively) was determined in the odds-ratio scale. The results of the analyses are provided in Table 1.

The predicted probability of achieving a correct response was saved for each case and graphed for both the focal and reference groups. Evidence of crossing DIF can be seen when the odds ratio shifts from lower odds to greater odds, or vice versa, across the scale-score range. Figure 1 shows a graph of item #38, which supports the evidence of crossing DIF found in Table 1. Figure 1, as well as Figures 3 and 4, are reprinted Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

Table 1 Significant DIF results

Item	*DIF type*	*DIF*	*FCAT achievement level cut scores*			
			Level 1/2: 253	*Level 2/3: 294*	*Level 3/4: 346*	*Level 4/5: 398*
1	N	0.974	1.278	1.041	0.802	0.618
2	N	0.919	1.030	0.874	0.709	0.576
3	U	0.940	-	-	-	-
4	N	0.852	0.923	0.886	0.841	0.798
5	N	-	1.123	0.993	0.849	0.727
6	N	0.957	0.966	1.006	1.060	1.116
7	N	1.070	1.200	1.061	0.908	0.776
8	N	0.818	0.976	0.828	0.672	0.546
9	N	0.865	1.134	1.003	0.858	0.734
10	N	1.045	1.536	1.250	0.963	0.742
11	N	1.195	1.137	1.092	1.036	0.984
12	N	-	1.302	1.199	1.080	0.974
13	N	-	1.310	1.111	0.902	0.732
14	N	0.944	1.356	1.151	0.934	0.759
15	N	0.922	1.394	1.183	0.960	0.780
16	N	0.888	1.639	1.179	0.777	0.511
17	N	1.163	1.458	1.343	1.210	1.090
18	N	0.956	1.274	1.037	0.799	0.616
19	N	0.938	1.483	1.159	0.847	0.620
20	U	1.038	-	-	-	-
21	N	0.892	1.058	0.935	0.800	0.684
22	N	1.056	1.751	1.368	1.001	0.732
23	N	0.949	1.708	1.334	0.976	0.714
24	N	0.973	1.750	1.312	0.911	0.632
25	N	-	2.251	1.619	1.066	0.702
26	N	0.928	1.891	1.360	0.896	0.590
27	N	-	1.338	1.233	1.111	1.001
28	U	0.827	-	-	-	-
29	N	0.739	0.845	0.747	0.639	0.547
30	N	0.907	1.106	0.938	0.762	0.618

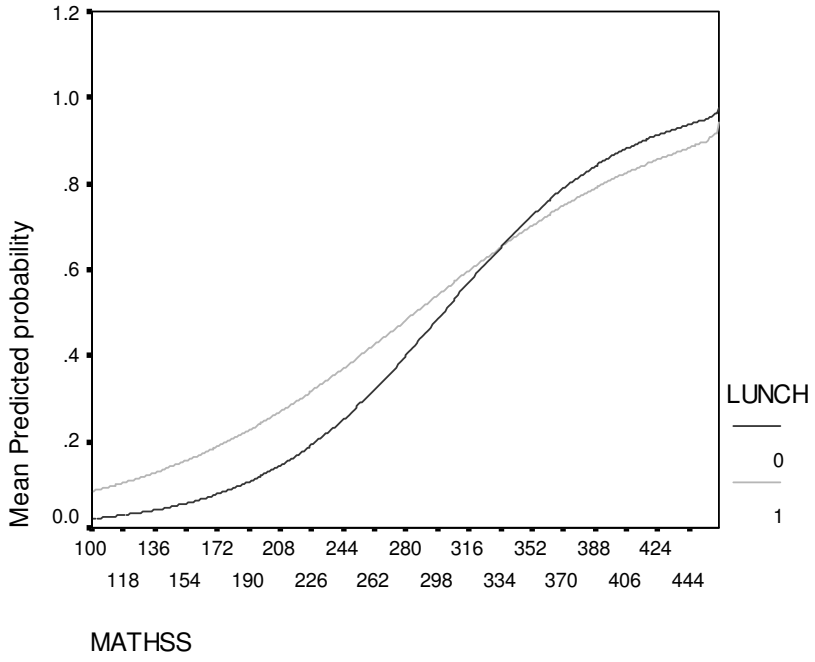
Notes: N = non-uniform DIF and U = uniform DIF. DIF values are the coefficient for the group indicator variable in the second logistic regression model (under the uniform DIF assumption) in the odds-ratio scale. All values in the table were found to be significant at the $p < 0.05$ level. Values not found to be significant at the $p < 0.05$ level were not included in the table and are represented by '-'.

Table 1 Significant DIF results (continued)

Item	*DIF type*	*DIF*	*FCAT achievement level cut scores*			
			Level 1/2: 253	*Level 2/3: 294*	*Level 3/4: 346*	*Level 4/5: 398*
31	N	1.086	1.369	1.210	1.035	0.885
32	N	1.056	0.977	1.018	1.073	1.130
33	N	-	1.047	0.964	0.869	0.783
34	N	1.062	1.100	0.972	0.832	0.711
35	N	0.897	1.214	1.030	0.836	0.679
36	U	1.074	-	-	-	-
37	N	0.856	1.159	0.906	0.662	0.484
38	N	1.071	1.747	1.365	0.998	0.730
39	N	0.962	1.014	0.897	0.767	0.656
40	N	0.819	0.996	0.778	0.569	0.416

Notes: N = non-uniform DIF and U = uniform DIF. DIF values are the coefficient for the group indicator variable in the second logistic regression model (under the uniform DIF assumption) in the odds-ratio scale. All values in the table were found to be significant at the $p < 0.05$ level. Values not found to be significant at the $p < 0.05$ level were not included in the table and are represented by '-'.

Figure 1 DIF results for item #38



3.1.2 STD results

While the effect-size estimates did range in size, none of the distractors were classified as having B or C DIF effects.

The effect-size estimates are based on a comparison of response rates of the reference group to those of the focal group, and values larger than one indicate that the reference group has greater odds of selecting the studied distractor. Values less than one indicate that the reference group has lower odds than the focal group of selecting the studied distractor, controlling for ability. In ordering effect-size estimates, distractor 2 of item 29 was the least attractive to the reference group ($\alpha_{STD_{29-2}}$ = .7721) while distractor 1 of item 31 was the most attractive, with the highest effect-size estimate ($\alpha_{STD_{31-1}}$ = 1.1964).

3.1.3 OR results

Consistent with the STD estimates, no item options were classified as having B or C DIF effects. The effect-size estimates are based on a comparison of the reference group to the focal group, such that values greater than one indicate that the reference group has higher odds of choosing the correct response over the studied distractor. This interpretation is different than that of the standardised distractor analysis, where the effect-size estimate is based on response rates to a particular distractor. Under this approach, values less than one indicate that the reference group has lower odds than the focal group of selecting the correct response over the studied distractor. In ordering effect-size estimates, distractor 2 of item 29 was the least attractive ($\lambda_{29\text{-}2}$ = 1.388) to the reference group, while distractor 1 of item 38 was the most attractive ($\lambda_{38\text{-}1}$ = .797).

3.1.4 MLR results

As discussed earlier, the first model analysed included only the FCAT scale score in the model. The second model included both the FCAT scale score and Lunch status. While the model fit improved with the addition of the variable Lunch, as indicated by the likelihood ratio tests, there was little to no improvement in the Nagelkerke *R*-square estimate. None of the analyses resulted in an improvement of at least .003; improvement no less than .003 was suggested by Kato et al. (2009) as a criterion for meaningful results. No distractors were classified as having B or C DIF effects.

The estimates are based on a comparison of the reference group to the focal group, with the correct response serving as the base category, such that the estimates are the odds for the students who do not receive free or reduced-price lunch services to choose the distractor over the correct response. Odds ratios greater than one indicate that the reference group has greater odds of choosing the studied distractor over the correct response. Odds ratios less than one indicate the reference group has lower odds of selecting the studied distractor over the correct response. Distractor 2 of item 29 was again the least attractive [$\exp(\hat{\beta}_{29-22})$ = .654] to the reference group, while distractor 2 of item 17 was the most attractive [$\exp(\hat{\beta}_{17-22})$ = 1.230].

3.1.5 Comparison of results

Spearman rank and Pearson correlations were calculated as measures of consistency. All correlations were significant and indicated a strong linear relationship. The patterns of both measures of correlation were similar, with the highest correlation between the OR and MLR approaches, which are both based on contrasts between the distractor and the correct answer. The correlations are summarised in Table 2.

Table 2 Spearman rank correlations (lower triangle) and Pearson correlations (upper triangle) between DDF effect-size estimates

Index	*STD*	*OR*	*MLR*
STD	1	–.922**	.907**
OR	–.898**	1	–.958**
MLR	.883**	–.953**	1

Note: **Correlation is significant at the 0.01 level (two-tailed).

The DDF effect-size estimates for each item were compared by their mean effect size, the effect-size range, and whether the DDF effects were divergent. The mean DDF effect size by item was calculated by taking the mean of the three DDF effects estimated for each item. The effect-size range was determined by taking the difference between the highest and lowest DDF effects estimated for each item. Items were classified as having divergent distractor-level effects if the combination of the three DDF effects included both a negative and positive effect among the three DDF effects within each item.

Table 3 summarises the correlation between the effect-size means. All of the methods had very high correlations, indicating a strong linear relationship in the effect-size means. The range of each item's DDF effect-size estimates also is consistent among the three methods, with the OR and STD methods having the strongest linear relationship as shown in Table 4.

Table 3 Spearman rank correlations (lower triangle) and Pearson correlations (upper triangle) between DDF effect-size means

Index	*STD*	*OR*	*MLR*
STD	1	–.961**	.951**
OR	–.958**	1	–.967**
MLR	.949**	–.957**	1

Note: **Correlation is significant at the 0.01 level (two-tailed).

Table 4 Spearman rank correlations (lower triangle) and Pearson correlations (upper triangle) between DDF effect-size ranges

Index	*STD*	*OR*	*MLR*
STD	1	.929**	.882**
OR	.905**	1	.888**
MLR	.849**	.848**	1

Note: **Correlation is significant at the 0.01 level (two-tailed).

As mentioned previously, items were coded as having divergent distractor-level effects if at least two of the three DDF effects within each item included a positive and negative effect (with divergent effects coded as '1' and non-divergent effects coded as '0'). The significance of a linear relationship between the methods in the identification of divergent DDF effects was tested with a phi coefficient. The phi coefficient is a Pearson correlation coefficient and is used when the data are dichotomous (Falk and Well, 1997). Both the OR and MLR methods were found to have a statistically significant phi coefficient with the STD approach. The strongest relationship was between the OR and STD methods (φ = .616). The OR and MLR methods were found to have a statistically significant positive correlation of medium magnitude (φ = .451), as well as the STD and MLR methods (φ = .423).

Patterns in the direction of DDF effect-size estimates by distractor were further examined by coding the effect-size estimates as positive or negative, and when available, noting the significance of the estimates. As evidenced by the large negative correlation between the estimates of the OR method and those of the STD and MLR methods (–.922 and –.958, respectively), the interpretation of the OR estimates is the opposite of the other two estimates, consistent with its conceptualisation.

Of the 120 DDF effects estimated, 87.5% of the OR and STD estimates shared the same pattern. Fourteen of the 15 estimates that differed did not have significant odds-ratio estimates. Fewer estimates were in agreement between the STD and MLR methods; 32 estimates did not indicate the same direction resulting in a consistency rate of 73.3%. Differences included those items found to be significant by the MLR model. The pattern of the OR and MLR estimates were slightly more consistent at 75.8%.

3.2 Relationship between DIF and DDF

3.2.1 Uniform DIF and DDF

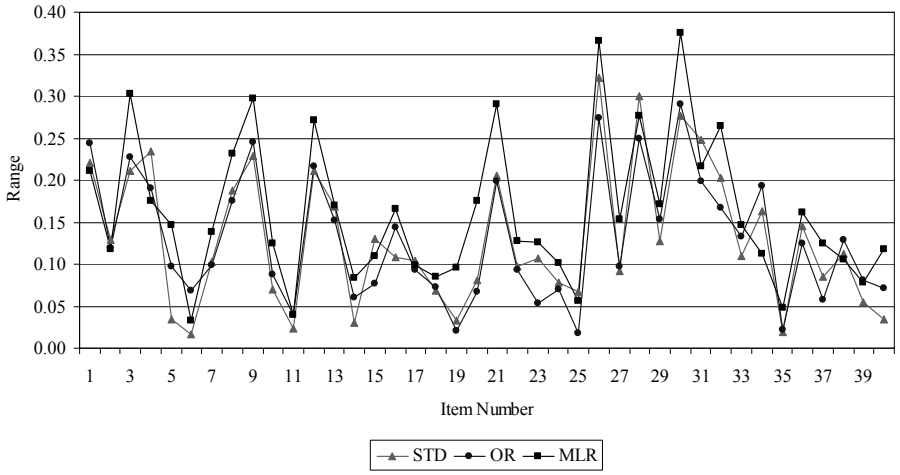
Four items were found to have only significant uniform DIF effects. These items were investigated for a constant DIF effect across all distractors based on the magnitude of the effect-size range. Table 5 provides the DIF effect-size estimates and the DDF effect-size range for these four items. None of the items exhibited constant DDF effects across the item distractors, as evidenced by the non-zero range estimates. The non-existence of constant DDF effects across each item's distractors was consistent across methods.

Table 5 Summary of uniform DIF effect-size estimates and DDF effect-size ranges by item

Item	*DIF*	*STD range*	*OR range*	*MLR range*
3	–0.062	0.211	0.227	0.303
20	0.037	0.080	0.068	0.175
28	–0.190	0.299	0.249	0.277
36	0.071	0.146	0.125	0.162

Because the DDF effects are estimated with error, the DDF effect-size ranges were graphed across all items to determine if the four studied items had ranges that were lower than other items, indicating a relatively constant effect across all distractors. For these four items, no patterns were found to indicate that smaller ranges were indicative of only uniform DIF (see Figure 2).

Figure 2 DDF effect-size ranges by item and method



Patterns in the direction of DDF effect-size estimates across distractors for these items were examined. For the purpose of this analysis, OR and MLR estimates that were not found to be significant are not included. Given this, Table 6 shows that the significant estimates of the OR and MLR methods are consistent with STD method. The only item found to have a constant pattern (i.e., the same sign) across distractors was item 20 under the STD method; however, there is no test of statistical significance of these results.

Table 6 Distractor effect-size patterns for items with uniform DIF only

Item	*STD*			*MLR*			*OR*		
	1	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*
3	–	–	+	– –	– –	++	++	++	– –
20	+	+	+		++			– –	
28	+	–	–		– –	– –		++	++
36	–	+	–	– –	++			– –	

Notes: Patterns in the direction of DDF effect-size estimates were coded as positive ('+') or negative ('-') using the log-odds scale indices. MLR and OR estimates that were significant at .05 were coded as '++' or '—' and estimates not found to be significant were left blank. STD estimates are coded as '+' or '-' in the absence of a significance test.

3.2.2 Non-uniform and crossing DIF and DDF

All items identified as having significant non-uniform DIF in the logistic regression analysis were plotted to determine if the non-uniform DIF also was crossing DIF. In addition to plotting, crossing DIF was investigated through determining the direction of the DIF effect at each FCAT Achievement Level cut point. Two items were found to have significant non-uniform DIF with no crossing DIF. Figure 3 displays a graph of item 17. As shown in Table 7, the DDF pattern was consistent across all three methods for these items.

Figure 3 Item #17 non-uniform DIF

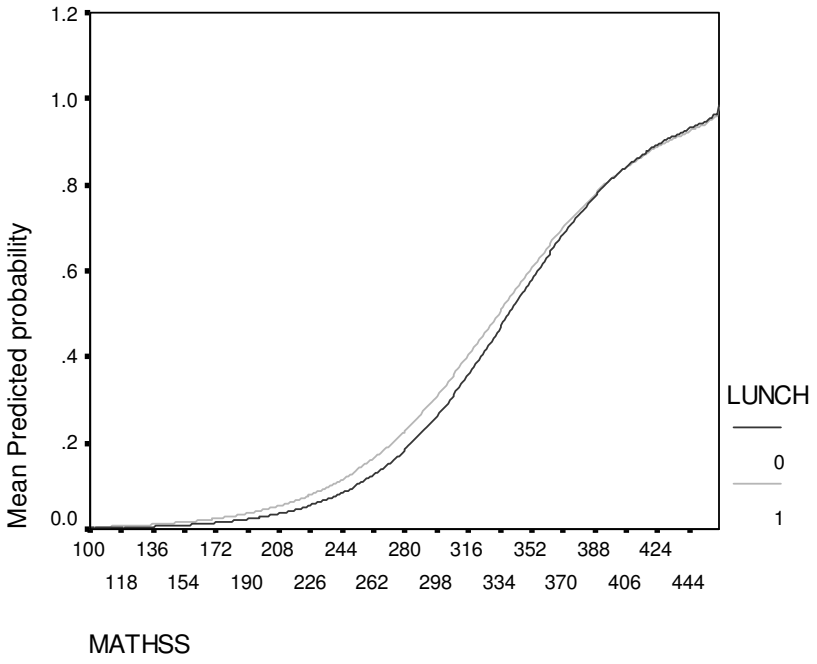


Table 7 DDF patterns: non-uniform DIF effect-size estimates

Item	*FCAT cut score*				*STD*			*MLR*			*OR*		
	253	*294*	*346*	*398*	*1*	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*
4	A-	A-	A-	A-	+	–	–		– –	– –		++	++
17	A+	A+	A+	A+	+	+	+	++	++	++	– –	– –	– –

Notes: Patterns in the direction of DDF effect-size estimates were coded as positive ('+') or negative ('-') using the log-odds scale indices. MLR and OR estimates that were significant at .05 were coded as '++' or '--' and estimates not found to be significant were left blank. STD estimates are coded as '+' or '-' in the absence of a significance test.

Thirty-four items were found to have significant non-uniform DIF that crossed at some point across the ability scale. For all patterns of DIF under this condition, the STD and OR results were consistent, with the exception of one DDF estimate. The STD and MLR results were much less consistent. After grouping items by their DIF pattern, the consistency percent was determined and is summarised in Table 8. The consistency percent was calculated as the percent agreement across all three methods by DIF pattern. The lowest level of consistency was found when the divergence in non-uniform DIF effects occurred between the FCAT Achievement Level 3 and Level 4 cut scores (294 and 346, respectively). The mean FCAT score for this assessment was 324, with a standard deviation of 64.73.

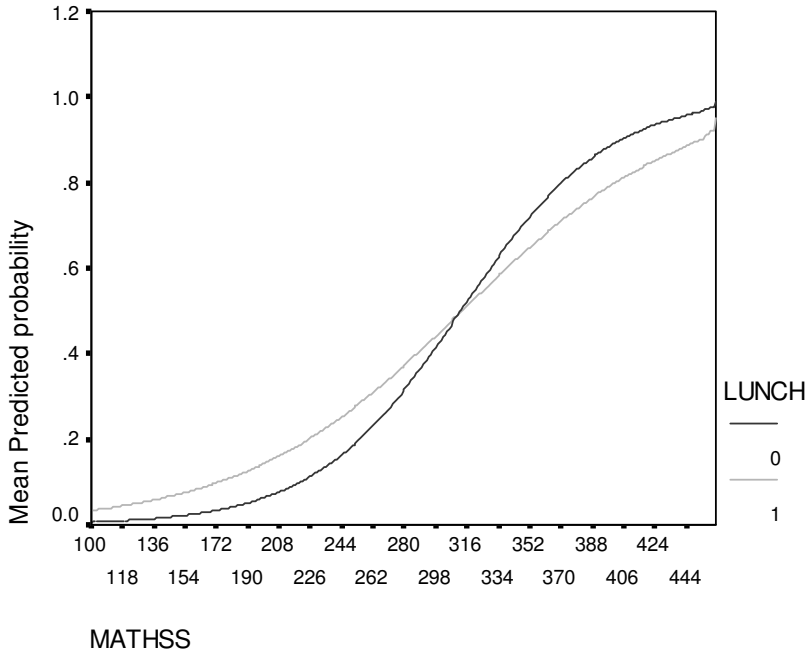
Table 8 Percent consistent by non-uniform DIF pattern

FCAT cut score				*Percent consistent*		
253	*294*	*346*	*398*	*Consistent*	*Total*	*Percent*
+	+	+	+	1	1	100
+	+	+	–	3	5	60
+	+	–	–	6	15	40
+	–	–	–	4	8	50
–	–	–	–	2	3	67
–	+	+	+	2	2	100

Notes: The six observed non-uniform DIF effect patterns are summarised in this table using '+' to indicate a positive effect and '-' to indicate a negative effect, using the log-odds scale indices.

Items 16, 24, and 26 exhibited large non-uniform, crossing DIF. Figure 4 graphs the mean probability of a correct response by FCAT scale score for item 16.

Figure 4 Item #16 crossing DIF



As can be seen in Table 9, the STD and OR methods were consistent in identifying divergent DDF effects when considering both significant and non-significant departures from 0. The results of the comparison between the MLR and OR were not as consistent and do not indicate divergent DDF in all cases (see the MLR results for item 16).

Table 9 Substantial non-uniform crossing DIF: DDF pattern by method

Item	*MLR*			*STD*			*OR*		
	1	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*
16	–	– –	– –	+	–	–	– –	++	++
24	–	– –	++	+	–	+	–	+	– –
26	– –	– –	+	–	+	+	++	–	– –

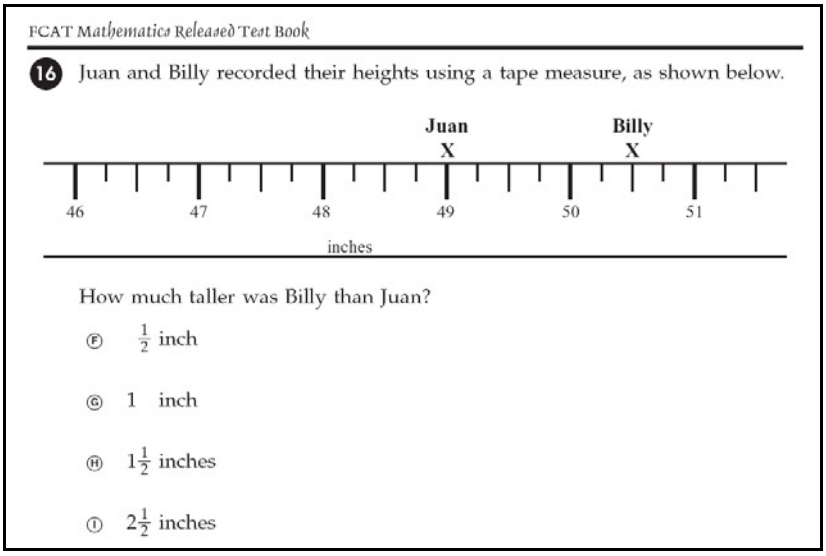
Notes: Patterns in the direction of DDF effect-size estimates were coded as positive ('+') or negative ('-') using the log-odds scale indices. MLR and OR estimates that were significant at .05 were coded as '++' or '--' and estimates not found to be significant were coded as '+' or '-'. STD estimates are coded as '+' or '-' in the absence of a significance test.

4 Item characteristics

FCAT test items are classified by several item characteristics, including the item content category, cognitive complexity, item difficulty, item discrimination, and the item guessing parameter. The relationship between item characteristics and the DDF item summary statistics (i.e., DDF effect-size range, mean effect size) was explored through the use of scatter plots and by calculating Pearson correlations. No significant correlations were found.

The three items identified as having large crossing DIF were studied to explore the utility of the information provided by the DDF analyses. The result of this exploration for item 16 is summarised here, and is consistent with the findings for the other items.

Figure 5 Item #16 content



Source: Florida Department of Education (2006)

Item 16 (see Figure 5) was classified as a moderate complexity item with a focus on measuring a student's understanding of content related to length, with 59% of the students selecting H as the correct response. Small percentages of students selected the first two options, with response rates for each of 6%.

Both the STD and OR DDF effect estimates indicated that distractor 1 (option F) was more attractive to the reference group in comparison to the focal group. In addition, all three methods found distractor 2 (option G) and distractor 3 (option I) less attractive to reference group members. The patterns can be found in Table 9. As identified by the content expert, options F and G presented answer options that were less than the correct amount presented by option H, and option I presented an answer option which was greater than option H. Other than these obvious distractor characteristics, there were no plausible reasons for the identified behaviour. This item had relatively high discrimination (a=1.134), moderate difficulty (b = 0.078), and moderate guessing (c = .208).

Similarly, no clear patterns emerged with the other two identified items to demonstrate the contributions that DDF analyses could make to the item revision process. It appears that the contributions would be greatest when items are being field tested. This type of information could assist content expert specialists in determining which distractors are problematic after DIF analyses on field test results. The items reviewed in this section were operational items that had already been subjected to many qualitative reviews prior to use.

5 Discussion

Three methods for detecting DDF were applied in this study. The STD and OR methods for detecting DDF were found to have very highly related results, with regard to both the magnitude and pattern of DDF effects. The MLR DDF results also were highly related to the STD approach, but yielded slightly different patterns across distractors. The OR and MLR methods are easily implemented with available software, such as SPSS 12 or higher, unlike the STD method, which is not a standard statistical package option. Despite these and the other discussed differences, all three methods present a viable option for use in improving test items included in assessment programmes.

As with any empirical study, the results of this study are limited by the focus on one statewide, operational assessment. The findings of this study may be different for assessment programmes that differ in the scoring model used and the steps used to screen items for placement on the assessment. Florida's statewide assessment did not include items with large uniform DIF that could be used in the investigation. Analysis of items with large uniform DIF may have resulted in different conclusions across methods, and interpretations of the usefulness of the resulting DDF information.

All methods potentially suffered from the loss of precision which results from sparse data. To maximise the information across methods, all FCAT scale scores were treated as distinct score levels. Because score levels were not grouped, many scale score levels may not have had adequate representation of one or both groups for each distractor. While this imprecision affected all models, it may have impacted the multinomial regression model the most as evidenced by the inconsistencies in the DDF patterns between the STD and MLR results.

The MLR model that was implemented in this study did not include an interaction term between FCAT scale score and Lunch status. Given the discrepant DDF patterns across methods when non-uniform, crossing DIF occurs in the middle of the ability scale, it is likely that DDF effects also are divergent within a distractor. To investigate this, MLR models were estimated with FCAT scale score, Lunch status, and the interaction term between these two variables. The change in the Nagelkerke *R*-square was compared to the model estimated with only FCAT scale score. Improvement greater than .003, as recommended by Abedi et al. (2008), in the Nagelkerke *R*-square was found for 18 of the 40 items. Of interest is that items 16, 24, and 26, with evidence of substantial crossing DIF, were three of the five items with the most improvement in the Nagelkerke *R*-square. The addition of the interaction term could improve the interpretation of the DDF effects.

The utility of the DDF effect estimates should be studied within the context of field test items. This context appears to have the most potential for using the information that can be provided by DDF results. After an item is field tested, a decision must be made regarding the validity of using an item in the calculation of a student's score. When an item is found to exhibit medium to large DIF effects, it often becomes ineligible for use on a statewide assessment. Losing items after field testing is not desirable, due to the costs associated with item development and field testing. A more desirable approach would be to revise such items. DDF information could be used to provide insights into this revision process. Specifically, items with problematic distractors could be rewritten and field tested again in lieu of being discarded. In addition, the insights gained through these analyses could be used to inform the item writing and review processes to continuously make each process more efficient and effective.

References

Abedi, J., Leon, S. and Kao, J.C. (2008) *Examining Differential Distractor Functioning in Reading Assessments for Students with Disabilities*, CRESST Tech. Rep. No. 743, National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.

Dorans, N.J. and Kulick, E. (1983) *Assessing Unexpected Differential Item Performance of Female Candidates on SAT and TSWE Forms Administered in December 1977: An Application of the Standardization Approach*, Research Rep. No. 83-9, Educational Testing Service, Princeton, NJ.

Dorans, N.J., Schmitt, A.P. and Bleistein, C.A. (1992) 'The standardization approach to assessing comprehensive differential item functioning', *Journal of Educational Measurement*, Vol. 29, No. 4, pp.309–319.

Falk, R. and Well, A.D. (1997) 'Many faces of the correlation coefficient', *Journal of Statistics Education*, Vol. 5, No. 3, pp.1–18.

Florida Department of Education (2005) *FCAT Mathematics Test Item Specifications, Grades 3–5*, Author, Tallahassee, FL.

Florida Department of Education (2006) *FCAT 2006 Mathematics Released Test, Grade 3*, Author, Tallahassee, FL.

Florida Department of Education (2007) *FCAT Reading and Mathematics: Technical Report for the 2006 FCAT Test Administrations*, Author, Tallahassee, FL [online] http://fcat.fldoe.org/pdf/releasepdf/06/FL06_Rel_G3M_AK_Cwf001.pdf (accessed 11 November 2009).

Florida Department of Education (2008) *FCAT Test Design Summary*, Author, Tallahassee, FL.

Florida Department of Education (2009) *Frequently Asked Questions* [online] http://www.fldoe.org/faq/default.asp?ALL=Y&Dept=202 (accessed 11 November 2009).

Green, B.F., Crone, C.R. and Folk, V.G. (1989) 'A method for studying differential distractor functioning', *Journal of Educational Measurement*, Vol. 26, No. 2, pp.147–160.

Kamata, A. and Vaughn, B.K. (2004) 'An introduction to differential item functioning analysis', *Learning Disabilities: A Contemporary Journal*, Vol. 2, No. 2, pp.49–69.

Kato, K., Moen, R.E. and Thurlow, M.L. (2009) 'Differentials of a state reading assessment: item functioning, distractor functioning, and omission frequency for disability categories', *Educational Measurement: Issues and Practice*, Vol. 28, No. 2, pp.28–40.

Middleton, K. and Laitusis, C.C. (2007) *Examining Test Items for Differential Distractor Functioning among Students with Learning Disabilities*, ETS Research Rep. No. RR-07-43, Educational Testing Service, Princeton, NJ.

Monahan, P.O., McHorney, C.A., Stump, T.E. and Perkins, A.J. (2007) 'Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression', *Journal of Educational and Behavioral Statistics*, Vol. 32, No. 1, pp.92–109.

Penfield, R.D. (2008) 'An odds ratio approach for assessing differential distractor functioning effects under the nominal response model', *Journal of Educational Measurement*, Vol. 45, No. 3, pp.247–269.

Penfield, R.D. (2010) 'Modeling DIF effects using distractor-level invariance effects: implications for understanding the causes of DIF', *Applied Psychological Measurement*, Vol. 34, No. 3, pp.151–165.

Schmitt, A.P. and Bleistein, C.A. (1987) *Factors Affecting Differential Item Functioning for Black Examinees on Scholastic Aptitude Test Analogy Items*, ETS Research Rep. No. RR-87-23, Educational Testing Service, Princeton, NJ.

Swaminathan, H. and Rogers, H.J. (1990) 'Detecting differential item functioning using logistic regression procedures', *Journal of Educational Measurement*, Vol. 27, No. 4, pp.361–370.

Wright, D.J. (1986) *An Empirical Comparison of the Mantel-Haenszel and Standardization Methods of Detecting Differential Item Performance*, Statistical Report No. SR-86-99, Educational Testing Service, Princeton, NJ.

Zwick, R., Donoghue, J.R. and Grima, A. (1993) 'Assessment of differential item functioning for performance tasks', *Journal of Educational Measurement*, Vol. 30, No. 3, pp.233–251.