2010

A Comparison of Methods for Detecting Differential Distractor Functioning

Sharon Koon

THE FLORIDA STATE UNIVERSITY

COLLEGE OF EDUCATION

A COMPARISON OF METHODS FOR DETECTING DIFFERENTIAL DISTRACTOR FUNCTIONING

By

SHARON KOON

A Dissertation submitted to the
Department of Educational Psychology and Learning Systems
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Spring Semester, 2010

The members of the committee approve the dissertation of Sharon Koon defended on March 16, 2010.

_____
Betsy Jane Becker
Professor Co-Directing Dissertation

_____
Akihito Kamata
Professor Co-Directing Dissertation

_____
Adrian Barbu
University Representative

_____
Jeannine Turner
Committee Member

_____
Yanyun Yang
Committee Member

Approved:

_____
Betsy Jane Becker, Chair, Department of Educational Psychology and Learning Systems

The Graduate School has verified and approved the above-named committee members.

I dedicate this dissertation to my dear husband, Rich,
and sons, Clay and Casey.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to all of my committee members, especially Dr. Aki Kamata and Dr. Betsy Becker. Both have been with me from the beginning of this journey to the end. Dr. Kamata guided me through my first statistics course to the completion of this dissertation, even after leaving Florida State University. Dr. Becker served as a constant advisor throughout this program and "adopted me" in the end to keep me on track. Without their support, encouragement, patience, and understanding, completing this dissertation would have been almost impossible. A special "thank you" goes to Dr. Jeannine Turner, who was willing to join my dissertation committee at the last moment.

I would like to thank my friends in and outside of work. For almost five years, they have listened to me talk about the ups and downs of graduate school. My "psycho" friends at work helped me through this dissertation by challenging my ideas, critiquing my work, and caring about my progress. The Lanes and the Daniels helped me manage my life, so I could manage my coursework.

I also would like to thank my parents for their encouragement throughout my undergraduate and graduate education. They always supported me with their prayers and best wishes.

I was inspired to earn my degree in measurement and statistics by Dr. Cornelia Orr. When I met her, I knew what I wanted to do "when I grew up." I was fortunate to have the opportunity to learn so much from her, and I am quite sure I would not be finished with this degree until the year 2020 without the flexible work schedule she provided me.

Most of all, it was through the support and encouragement of my husband and two sons that I was able to finish this goal. They fill me with gratitude and joy.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This study examined the effectiveness of the odds-ratio method (Penfield, 2008) and the multinomial logistic regression method (Kato, Moen, & Thurlow, 2009) for measuring differential distractor functioning (DDF) effects in comparison to the standardized distractor analysis approach (Schmitt & Bleistein, 1987). Students classified as participating in free and reduced-priced lunch programs served as the focal group and students not participating in these programs served as the reference group. The comparisons were conducted in such a way as to provide insight into two research questions: 1) whether the magnitude and pattern of the DDF effect is constant across all methods, and 2) whether the pattern of DDF effects support differential item functioning (DIF) findings. Measures of effect size are reported. In addition, the relationship between item characteristics and DIF and DDF effects were explored for patterns.

Comparisons of three methods for detecting DDF were conducted in this study. The standardized distractor analysis and odds-ratio methods for detecting DDF were found to have very highly related results, with regard to both the magnitude and pattern of DDF effects. The multinomial logistic regression DDF results also were highly related to the standardized distractor analysis approach, but yielded slightly different patterns across distractors. The odds ratio and multinomial logistic regression methods are easily implemented with available software, such as the SPSS software package used in this study, unlike the standardized distractor analysis method which must be programmed. Despite these and the other discussed differences, all three methods present a viable option for use improving test items included in statewide assessment programs.

CHAPTER 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 being added to an assessment, the validity of the use of the item must be demonstrated. This validity is established through both statistical reviews, based on field-testing, and expert committee reviews. All of these processes are very expensive and are aimed at ensuring that the item is a valid measure of a student's ability or achievement.

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 behavior 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. As Camilli (2006) stated, "DIF is synonymous with statistical bias, whereas unfairness can only be established if these measurement differences are factors irrelevant to the test construct; there is no direct route from statistical bias to unfairness. To maintain the distinction between statistical bias and unfairness, DIF is used as one kind of screening mechanism for quality control"(p. 234).

Items that are identified as having DIF and possible item bias are evaluated for the magnitude of DIF and the type of DIF. DIF is said to occur in two ways – uniform and non-uniform. 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.

Penfield (in-press) showed that DIF may be partially explained by studying examinee responses to item distractors, or the incorrect options in a multiple-choice item. In a recent study, he found that constant differential distractor functioning (DDF; Green, Crone, & Folk, 1989) effects across all distractors can lead to uniform DIF, while non-uniform DIF is an indication that the DDF effects may vary in sign across the distractors.

Penfield's work, and the work of others in the area of distractor analysis, is in part based on the work of Green et al. (1989) and Dorans and Kulick (1983). Green et al. (1989) proposed a method for studying examinee responses for group differences in distractor selection rates, which they named differential distractor functioning (DDF). Specifically, DDF is a method for studying DIF based on distractor-level responses. While Schmitt and Bleistein (1987) extended the standardization approach developed by Dorans and Kulick (1983) to assessing DIF to identify possible item factors that may contribute to DIF, Green et al. (1989) are most often credited for their focus on distractor analysis.

Improvements in the methods for studying DDF have been made since the 1980s. Wang (2000a) proposed a factorial model of differential distractor functioning to allow for the examination of the effect of multiple grouping factors, as well as interactions between them. A multi-step logistic regression procedure was used by Abedi, Leon, & Kao (2008) to explore differential trends in the selection of one of three distractors by students with disabilities. Penfield (2008) proposed an odds-ratio (OR) method for assessing DDF effects as modeled under both the nominal response model (Bock, 1972) and the multiple-choice model (Thissen & Steinberg, 1984). 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. All of these methods are proposed for the purpose of extending the analysis of invariance to all item responses and not just between correct and incorrect responses.

This study examined the effectiveness of the OR method (Penfield, 2008) and the MLR method (Kato et al., 2009) for measuring DDF effects in comparison to the standardized distractor analysis (STD) approach (Schmitt & Bleistein, 1987). 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 any item characteristics appear to explain the DDF effects.

Statement of the Problem

The OR approach to detecting DDF was proposed to address limitations of existing methods. Penfield (2008) proposed that an alternative method was needed that: "(a) could be implemented with moderate group sizes; (b) yielded DDF effect estimates that could be interpreted with respect to established models for multiple-choice items; and (c) was

conceptually simple, such that it could be effectively implemented using readily available software" (p. 249). In studying this approach, he found that the OR approach can address these limitations for data generated under the nominal response model. For data generated under the multiple-choice model, he found that the OR approach underestimated the DDF effect, much like the Mantel-Haenszel method had been found to do when using models that include guessing parameters. He recommended that a comparison of the OR approach to parametric approaches would provide useful information. To extend his research, the MLR method was selected for this study as the parametric comparison due to the limited information in the literature on the use of this approach. In contrast, the STD method was selected because of its accepted use in detecting DDF.

Significance of the Study

The purpose of this study was to compare the performance of methods for detecting DDF in an attempt to understand their similarities and differences. In addition, this study sheds light on the benefits of using the OR approach with real testing data to understand DIF through the evaluation of DDF effects. The goal of this study was to discuss and make recommendations about DDF methodology choice and use for practitioners.

CHAPTER 2

REVIEW OF LITERATURE

This chapter reviews the major research efforts relating to differential distractor functioning (DDF). To date, there are two major areas of focus in the research, in addition to research related to the application of DDF. The first area is concentrated on the methods for detecting DDF. The second area is focused on the relationship between DDF and differential item functioning (DIF). This chapter summarizes both of these efforts.

Methods for Detecting Differential Distractor Functioning

Many different methods exist for the detection of DDF. In most cases, the DDF method is an extension of an already existing method for detecting DIF. Mapuranga, Dorans, and Middleton (2008) summarized DIF methods and procedures that have appeared in the literature since the 1990s and classified these methods into four categories: expected item score methods, nonparametric odds-ratio methods, generalized linear model methods, and item response theory (IRT) methods. Their classification taxonomy is used in this chapter to summarize the primary methods that have been applied in the detection of DDF.

Expected Item Score Methods

As classified by Mapuranga et al. (2008), expected item score DIF detection methods analyze the extent to which there is no difference in the expected item score between the reference and the focal groups, after controlling for ability. To extend this classification to methods of DDF detection, the focus is placed on identifying differences in response rate proportions, where the null definition is that there is no difference in response rate proportions between the reference and focal groups at each level of ability.

Standardized distractor analysis. Schmitt and Bleistein (1987) applied the standardization approach developed by Dorans and Kulick (1983) for assessing DIF to identify possible item factors that may contribute to DIF. This extension was ultimately referred to as the standardized distractor analysis (STD) method. STD extends the standardization approach by analyzing all distractors, not reached, and omits, instead of only analyzing differences between the correct and the total of all incorrect responses. This nonparametric approach allows for a test of DDF effects for each response option and is considered simple to implement.

As a basis for the extension to the distractor method, the standardization approach (Dorans & Kulick, 1983) to detecting DIF is explained first. This approach conditions on ability, such as the total test score, or *s*, and then makes use of a 2 x 2 contingency table to assess DIF. The format of this contingency table for each score level is provided in Table 2.1.

Table 2.1
Sample 2 x 2 Contingency Table

Group	Right (1)	Wrong (0)	Total Number (*n*)
Reference (*R*)	R_{1s}	R_{0s}	n_{rs}
Focal (*F*)	F_{1s}	F_{0s}	n_{fs}
Number (*n*)	n_{1s}	n_{0s}	n_s

Each cell in Table 2.1 represents the number of examinees in each group who answered the item right (1) or wrong (0) by score level. Using these values, a proportion correct is determined at each score level for both groups. It is the magnitude of the difference (*D*) in proportion right (*P*), or p-difference, between both groups at each score level that is of concern. The difference is calculated as $D_s = P_{fs} - P_{rs}$, where $P_{fs} = F_{1s}/n_{fs}$ and $P_{rs} = R_{1s}/n_{rs}$. At each score level, the definition of null DIF is $F_{1s}/n_{fs} - R_{1s}/n_{rs} = 0$, where $s = 1, \ldots, S$.

Both plots and item-discrepancy indices are used to identify potential problematic items and these approaches will be described in the context of the STD.

The STD makes use of the individual distractor in the analysis instead of collapsing all incorrect responses into one incorrect total. To conduct an analysis of responses to distractor *j*, for instance, the calculation of proportions and differences in proportions would be, $P_{fs}(j) = F_{js}/n_{fs}$ and $P_{rs}(j) = R_{js}/n_{rs}$, and $D_s(j) = P_{fs}(j) - P_{rs}(j)$, where F_{js} and R_{js} are the number of focal and reference group members who selected distractor *j*, respectively.

STD, and the standardization method in general, use two numerical indices for identifying items that need further investigation (Dorans, 1989; Dorans & Holland, 1993; Dorans & Kulick, 1986; Dorans, Schmitt, & Bleistein, 1988; Schmitt & Dorans, 1990). The first index is the standardized *p*-difference, *STD P-DIF*. Carrying forward the example related to the analysis of individual responses, the *STD P-DIF* for distractor *j* would be calculated as

$$\text{STD P-DIF (j)} = \sum_{s=1}^{S} w_s [P_{fs}(j) - P_{rs}(j)] \Big/ \sum_s w_s \qquad (1)$$

$$= \sum_s w_s P_{fs}(j) \Big/ \sum_s w_s - \sum_s w_s P_{rs}(j) \Big/ \sum_s w_s$$

$$= P_f - \hat{P}_f$$

where $w_s / \sum w_s$ is a standard weight applied to both the focal and reference group proportions. Dorans et al. (1988) provide several options for the value of w_s, as summarized in Table 2.2.

Table 2.2
STD Weighting Options

Weight	Description
$w_s = n_s$	Number of examinees at *s* in the total group
$w_s = n_{rs}$	Number of examinees at *s* in the reference group
$w_s = n_{fs}$	Number of examinees at *s* in the focal group
$w_s = n_{xs}$	Number of examinees at *s* in some reference group

While all of these weighting options are available and each may be appropriate depending on the circumstances of the application, the option used in the proposed method is to let w_s equal the number of examinees at *s* in the focal group, or n_{fs}. In the absence of a reason for selecting one of the other weighting options, n_{fs} is selected so that the greatest weight is applied to differences at score levels most obtained by the focal group.

Under this notation and selection of weighting, P_f is the observed response of the focal group on the distractor and $\hat{P}_f$ is the expected performance of the focal group predicted from the reference group regression curve $P_{rs}(j)$.

The calculated *STD P-DIF* index ranges from -1 to +1, with a positive value indicating that the distractor is more attractive to the focal group and negative value indicating the distractor is less attractive to the reference group. *STD P-DIF* values outside the range of -.10 to +.10 indicate distractors that should be studied carefully, with values within that range indicating moderate to negligible effects.

The second index is the root-mean-weighted-squared difference (RMWSD),

$$\text{RMWSD}\ (j) = \quad [\sum_{s=1}^{S} w_s D_s(j)^2 / \sum_{s=1}^{S} w_s]^{.5}. \qquad (2)$$

Unlike the *STD P-DIF*, this index is not a signed index and does not allow cancellation to occur. While a RMWSD of 0.08 has been identified as an indication the distractor needs further review, practical use has found that this index is sample specific and is not as meaningful as the *STD P-DIF*. In light of this limitation, the *STD P-DIF* index is used most often.

A delta metric version, discussed later in this chapter, of the STD P-DIF for distractor *j* is

$$\text{STD D - DIF}(j) = -2.35 \ln\{[\hat{P}_f / 1 - \hat{P}_f)] / [P_f / (1 - P_f)]\}. \qquad (3)$$

A positive *STD D-DIF* (*j*) indicates that the focal group is more likely to select distractor *j*.

In addition, plots of conditional proportion correct and conditional differences are used to visually analyze differential performance between the focal and reference groups. The conditional proportion correct plot allows for an examination of the conditional mean item score at each score level. The plot of conditional differences allows for an examination of the variation in performance between both groups.

The STD was used by Schmitt and Bleistein (1987) to study differential speededness. Like the study of an individual distractor described earlier, this analysis used a non-response indicator to study differential response rates to items appearing at the end of a test. Similarly, Rivera and Schmitt (1988) used STD to study omitted responses of Hispanic examinees on the Scholastic Aptitude Test. Based on the insights learned from these and other studies of differential speededness and differential omit rates using STD, Dorans, Schmitt, and Bleistein (1992) expanded the notion of DIF detection to what they called comprehensive differential item functioning (Cdif). Cdif was identified as new terminology for the use of STD and was described as an important "adjunct to Mantel-Haenszel DIF detection" (p. 309) because it could be used to understand the factors contributing to items flagged as having a significant level of DIF, known as "C" 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. The authors applied the method as described in this section, but noted several limitations. The first limitation reported was that the use of a signed index may mask the true DDF effects because cancellation or reduction may occur. "For

example, students at the lower end of the distribution in the focal group may be differentially attracted to a particular distractor, but students at the higher end of the distribution in the reference group may be differentially attracted to the same distractor" (p. 14). Secondly, like in DIF analyses, sparse data at the extreme ends of the score scales resulting from different ability distributions may result in less reliable DDF estimates.

Dorans et al. (1992) contrasted the potential use of Cdif with the Green et al. (1989) log-linear method and the Thissen, Steinberg, and Wainer (1992) IRT method, discussed later in this chapter. The authors concluded that the Cdif approach to evaluating invariance was preferred when the sample size is sufficient because it is able to condition on all ability levels, by specific focal and reference group pairs, without having to make parametric assumptions. On the contrary, when fewer items are being studied, the data are sparse, and when the IRT model fits the data for all responses, the authors stated the IRT model may work best.

Nonparametric Odds-Ratio Methods

DIF detection methods that are based on nonparametric odds ratios are testing whether there is a difference in the odds of a correct response, at matching ability levels, between the reference and focal groups as determined by the significance of a log-odds ratio estimate. The same analysis may be made to detect DDF, where instead of the odds of a correct response, the analysis is concerned with the odds of a response to a certain response option.

Odds-ratio (OR) analysis of DDF. Penfield (2008) proposed the OR approach for assessing DDF effects as modeled under both the nominal response model (Bock, 1972) and the multiple-choice model (Thissen & Steinberg, 1984). The proposed method is an extension of the Mantel-Haenszel (MH) method originally proposed by Mantel and Haenszel (1959) and then extended by Holland and Thayer (1988) for analyzing dichotomously-scored items.

Kamata and Vaughn (2004) provided an introduction to the MH method and its use of a three-way contingency table to test for the dependency of two variables by a matching criterion. The two variables in this test are the two scoring categories and the two groups of interest, the reference group and the focal group. The item responses are totaled for both groups, using a dichotomous indicator (0, 1), and tallied in the format of Table 2.1. As such, the same data are used for the standardization approach discussed earlier and the MH approach. Dorans and Holland (1993) discussed this common framework and demonstrated that both approaches share

the same definition of null DIF and only vary in the method for measuring departures from the null DIF.

Using the data in the contingency table, the MH approach uses an odds ratio at each score level. For item *i*, the MH common odds ratio is computed as follows

$$\hat{\alpha}_{MH_i} = \frac{\sum_{s=1}^{S} R_{1s} F_{0s} / n_s}{\sum_{s=1}^{S} R_{0s} F_{1s} / n_s}. \quad (4)$$

Because the MH estimate is an odds ratio, with the odds of the reference group compared to the focal group, an odds ratio of one indicates no difference, an odds ratio of greater than one indicates that the item benefits the reference group and an odds ratio less than one indicates the item benefits the focal group. When the odds ratio is converted to a log-odds ratio by taking the natural log, the value becomes a signed index. This signed index is referred to as $\hat{\beta}_{MH_i}$ and is calculated by $\hat{\beta}_{MH_i} = \ln(\hat{\alpha}_{MH_i})$, where a positive number indicates that the item benefits the reference group and a negative number indicates that the item benefits the focal group. While a value of zero indicates no statistical bias, each nonzero value indicates a possible bias toward one of the groups.

To determine the magnitude of this potential bias on an existing item difficulty scale, the Educational Testing Service (ETS) converts $\hat{\alpha}_{MH_i}$ to the delta (Δ) metric. Large delta values indicate difficult items, while small delta values indicate easy items. The conversion may be accomplished by $MH\ D\text{–}DIF = -2.35 \ln(\alpha_{MH_i})$ or $MH\ D\text{–}DIF = -2.35 \hat{\beta}_{MH_i}$.

A positive *MH D-DIF* statistic indicates that the item favors the focal group, whereas a negative value indicates that the item favors the reference group.

Once in the delta metric, absolute values of the *MH D-DIF* statistic are used to classify items into three categories, as indicated in Table 2.3.

Table 2.3
Definition of the DIF Categories A, B, and C in the Currently Used Procedures Based on the MH-D-DIF Statistic, and the Action to Be Taken by Test Development

Category	Absolute Value and Significance of MH D-DIF	Action	
		During Test Assembly	Before Score Reporting
A	MH D-DIF not significantly different from 0 (.05 level) OR Absolute value less than 1	Select freely	No action required
B	MH D-DIF significantly different from 0 (.05 level) AND EITHER (1) Absolute value at least 1 but less than 1.5 OR (2) Absolute value at least 1 but not significantly greater than 1 (.05 level)	If there is a choice among otherwise equivalent items, select the item with the smallest absolute value of MH D-DIF	No action required
C	Absolute value of MH D-DIF at least 1.5 and significantly greater than 1 (.05 level)	Select only if essential to meet specifications. Documentation and corroboration by reviewer required	Documentation and corroboration by independent review panel required

Source: Peterson (1987) as reproduced in Longford, Holland, and Thayer (1993).

As presented in Dorans (1989), the MH chi-square test statistic for each item i, $\alpha^2_{MH_i}$, is chi-square distributed with a $df = 1$ and is computed by

$$\alpha^2_{MH_i} = \left[\sum_{s=1}^{S} R_{1s} - \sum_{s=1}^{S} u_s - 0.05\right]^2 \Big/ \sum_{s=1}^{S} \sigma_s^2 \qquad (5)$$

where

$$\mu_s = E(R_{1s} \mid \alpha = 1) = n_{rs} n_{1s} / n_s \qquad (6)$$

$$\sigma_s^2 = VAR(R_{1s} \mid \alpha = 1) = \frac{n_{rs} n_{fs} n_{1s} n_{0s}}{n_s^2 (n_s - 1)}. \qquad (7)$$

The null hypothesis being tested by $\alpha^2_{MH_i}$ is that $\hat{\alpha}_{MH_i}$=1, or that the reference and focal groups have the same odds of getting an item correct at all score levels. The null hypothesis will be rejected at the .05 significance level if $\alpha^2_{MH_i}$ is greater than 3.84.

Like the simple extension of the standardization approach, Penfield's (2008) extension of the MH approach decomposes the "wrong" designation into separate distractor-level analyses. He proposed this method to provide an alternative to the differential alternative model (DAF; Thissen, Steinberg, & Fitzpatrick, 1989) discussed later in this chapter. While the DAF provides a parametric method for detecting DDF, the method is much more complex to implement than the MH method. Penfield's goal was to propose a method that did not depend on large group sizes and could provide DDF effect-size estimates that are parallel to those from multiple-choice models.

As proposed, the DDF effect was modeled under the nominal response model (NRM; Bock, 1972), such that invariance is modeled through DDF effects and not a DIF effect. The NRM parameterization for each *j*th contrast was provided as

$$P(Y = y_1 | Y = y_1 or\ y_j) = \frac{\exp(z_j)}{1 + \exp(z_j)} \tag{8}$$

where $z_j = c_j + a_j\theta$.

The c_j and a_j parameters represent the location and slope parameters, respectively, in the model, $Y = y_1$ and $Y = y_j$ denote the selection of the correct option or the *j*th distractor, respectively. The probability of selecting the correct option given that the examinee selected either the correct option or *j*th distractor is conditioned on the ability (θ) of the examinee. The DDF effect parameter for the *j*th contrast function (ω_j) and grouping variable (G) were added to the logit (z_j) by $z_j = c_j + a_j\theta + G\omega_j$.

Penfield (2008) demonstrated that ω_j could be estimated through the natural logarithm of the conditional odds ratio of the *j*th contrast, called λ_j, as follows

$$\begin{aligned} \lambda_j &= \ln\left\{ \frac{\dfrac{P(Y = y_1 | Y = y_1 or\ y_j, \theta, G = 1)}{1 - P(Y = y_1 | Y = y_1 or\ y_j, \theta, G = 1)}}{\dfrac{P(Y = y_1 | Y = y_1 or\ y_j, \theta, G = 0)}{1 - P(Y = y_1 | Y = y_1 or\ y_j, \theta, G = 0)}} \right\} \\ &= \ln\left\{ (\exp[c_j + a_j\theta + \omega_j]) \times (\exp[c_j + a_j\theta]^{-1} \right\} \\ &= \omega_j . \end{aligned} \tag{9}$$

Penfield (2008) proposed that for the *j*th contrast function, the conditional odds ratio across all score levels can be estimated by

$$\hat{\alpha}_j = \frac{\sum_{s=1}^{S} R_{1s} F_{js} / n_s}{\sum_{s=1}^{S} R_{js} F_{1s} / n_s}. \quad (10)$$

Taking the natural logarithm of $\hat{\alpha}_j$, $\hat{\lambda}_j = \ln(\hat{\alpha}_j)$, 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 favors the reference group, and a negative value indicates DDF favoring the focal group.

Table 2.4 illustrates that the data required for the analysis are consistent with the STD approach, where *j*=1, 2,…, *J* for each of the *J* distractors (i.e., options that are not the right answer).

Table 2.4
Sample Odds Ratio DDF Contingency Table

Group	Right (1)	Wrong (*j*)	Total Number (*n*)
Reference (*R*)	R_{1s}	R_{js}	n_{rs}
Focal (*F*)	F_{1s}	F_{js}	n_{fs}
Number (*n*)	n_{1s}	n_{js}	n_s

If the data do not fit the NRM, the estimates of the DDF effect could be moderately biased. However, Penfield (2008) noted, "one of the advantageous properties of $\hat{\lambda}_j$ is its nonparametric nature, which frees it from the restrictions of model fit associated with parametric approaches to DDF effect evaluation" (p. 254).

The standard error of $\hat{\lambda}_j$, as provided by Penfield (2008), is

$$SE(\hat{\lambda}_j) = \sqrt{\frac{\sum_{s=1}^{S} n_s^{-2}(R_{1s}F_{js} + \hat{\alpha}_j R_{js}F_{1s})(R_{1s} + F_{js} + \hat{\alpha}_j R_{js} + \hat{\alpha}_j F_{1s})}{2\left(\sum_{s=1}^{S} \frac{R_{1s}F_{js}}{n_s}\right)^2}}, \tag{11}$$

leading to a test statistic that is distributed approximately standard normal

$$z(\hat{\lambda}_j) = \frac{\hat{\lambda}_j}{SE(\hat{\lambda}_j)}. \tag{12}$$

To avoid inflation of the Type I error rate across all tests of DDF, the distractor-level Type I error rate can be adjusted by dividing the intended rate by the total number of distractors.

Penfield (2008) conducted simulation studies to evaluate the performance of $\hat{\lambda}_j$ and $z(\hat{\lambda}_j)$ when data are generated using the NRM and the multiple-choice model (MCM; Thissen and Steinberg, 1984). Results of the studies indicated that the DDF effect estimators performed well under the NRM, with only minimal bias. Under the MCM, the DDF effect estimators did not perform as well. Under the non-null conditions simulated, $\hat{\lambda}_j$ resulted in biased estimates, with the estimates biased toward zero at varying levels. It was noted that this same type of bias is found when the MH odds ratio is applied to data generated by the three-parameter logistic model and may result from the presence of guessing. Taking into account this bias, Penfield (2008) stated that $\hat{\lambda}_j$ can be viewed as a "conservative estimator of the true DDF effect" (p. 265).

Generalized Linear Model Methods

Mapuranga et al. (2008) classify methods in this category because they model data that are linearly based on assumed probability distributions. Included in this category are logistic regression, mixture models, and hierarchical models. The log-linear method is a specialized case of the generalized linear model, in which only the association between variables is being tested. All of the methods are parametric and, therefore, require testing the fit of the model to the data.

Log-linear analysis. Green et al. (1989) proposed a log-linear method to extend the DIF analysis of test items to include all item responses, in an analysis of what they named differential distractor functioning (DDF). This method makes use of a three-way contingency table where item choice is analyzed by subgroup by ability level. While the method is flexible to

enough to include all answer choices, only incorrect choices were analyzed in the proposed method.

The log-linear analysis allows for an examination of the contribution of both main effects and interaction effects in accounting for the data. The main effect of ability group provides information on any differences in ability groupings, just as the main effect of subgroup accounts for differences in the numbers of examinees in each subgroup and the main effect of distractor choice explains differences in distractor choice. The interaction effects include the ability by subgroup, ability by distractor choice, and the subgroup by distractor choice interactions. The first two interactions help to account for differences in proportions unique to the data set, where the ability by subgroup interaction helps to explain differences in the abilities of subgroups and the ability by distractor choice interaction helps to explain differences in the distractors chosen by examinees of different abilities. The third interaction, the subgroup by distractor choice interaction, accounts for differences in the numbers of examinees within a subgroup choosing a particular distractor. This third interaction serves as an indication of DDF.

The hypothesis being tested under this method is that the subgroup by distractor choice interaction does not improve the model fit, as measured by the chi-square statistic, degrees of freedom, and p-value. To test this hypothesis, a model is fit with all of the main effects and only the ability by subgroup and ability by distractor choice interactions. This model is compared to the same model with the addition of the subgroup by distractor choice interaction. When the model fit improves with the addition of this interaction, DDF is inferred.

To provide a measure of effect size by distractor choice, the collected frequencies can be used to produce plots of distractor choice by subgroup and ability. These plots allow for visual inspection of the differences in distractor choice for items that are found to function differently by subgroup. For example, a plot of the conditional percent choosing a distractor at each score level would allow for a visual inspection of differences in response rates for that particular distractor. Similarly, one can plot differences in conditional percents choosing a distractor at each score level.

Green et al. (1989) demonstrated the use of the log-linear method through a study of the verbal section of the Scholastic Aptitude Test (SAT-V). All 85 items on the SAT-V were analyzed for DDF using the total verbal raw score as the measure of ability and ethnicity as the grouping variable. Of the 85 items, 15 items showed evidence of DDF. Further qualitative

analyses explored the reason for the DDF, including the item type and item content. While there were no clear findings related to item content, the authors did conclude that item type did appear to explain 6 of the 15 DDF items. These 6 items were classified as sentence completion items and appeared to contribute to the differential selection by the Hispanic subgroup.

While the log-linear method does provide a test of DDF effects overall, it does not allow for a test of DDF by distractor. Therefore, fully understanding the DDF effects is limited to the use of the plots described earlier.

Logistic regression analysis of DDF. 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. This approach was an extension of the use of logistic regression in DIF detection.

Swaminathan and Rogers (1990) studied the use of logistic regression to detect DIF. Their motivation was to identify a cost effective method for detecting both uniform and nonuniform DIF since the MH had been found to be unable to detect nonuniform DIF.

The logistic regression model for each item is expressed by

$$\ln\left[\frac{p_e}{(1-p_e}\right] = \beta_0 + \beta_1 X_e + \beta_2 G_e + \beta_3 (XG)_e, \qquad (13)$$

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. For clarity, an item subscript can be included in the model.

Implementation of this method requires estimating several models. The first model is estimated with only X_e in the model, and this model is referred to as the 1st model. The 2nd model includes both X_e and G_e. The final model, the 3rd model, includes X_e, G_e, and $(XG)_e$. Each model is analyzed separately so that a chi-square statistic is estimated based on the log-likelihood ratio. There is no evidence of DIF if the coefficients β_2 and β_3 are not significantly different than zero. All coefficients in the model are in the scale of the log-odds ratio, and therefore can be interpreted in the same manner as $\hat{\beta}_{MH_i}$.

The chi-square statistics are used to test for uniform and nonuniform DIF. The first application of this test is to test for both uniform and nonuniform DIF. To do this, the chi-square

difference test is used to test for a difference between the 3rd and 1st models, by $x^2_{DIF} = x^2_{3rd\ model} - x^2_{1st\ model}$. If the test is significant, based on a chi-square distribution with 2 degrees of freedom, there is evidence for uniform, nonuniform, or both, DIF.

To determine the type of DIF, further x^2_{DIF} tests are performed. For uniform DIF, the test is based on 1 degree of freedom and is determined by $x^2_{uniform\ DIF} = x^2_{2nd\ model} - x^2_{1st\ model}$. Non-uniform DIF is tested with the difference between the 3rd and 2nd models, based on 1 degree of freedom, determined by $x^2_{non-uniform\ DIF} = x^2_{3rd\ model} - x^2_{2nd\ model}$.

To evaluate the magnitude of the DIF, the magnitudes of $\hat{\beta}_2$ and $\hat{\beta}_3$ are examined in the same way that $\hat{\beta}_{MH_i}$ is interpreted under the MH method. The estimates of β_2 and β_3 can be multiplied by -2.35 to interpret the values in the delta metric, or taking the exponents of $\hat{\beta}_2$ and $\hat{\beta}_3$ will place the values on the $\hat{\alpha}_{MH_i}$ scale.

Abedi et al. (2008) applied the logistic regression method to examine DDF effects in reading assessments for students with disabilities. All of the items that were studied were multiple-choice items with four response options and only the distractors were studied. Unlike the DIF analysis described previously, where the dependent variable is a dichotomous indicator for a correct or incorrect response, the authors defined the dependent variable to serve as an indicator of distractor selection. Student responses were coded into two categories based on the popularity of the distractor. The most commonly selected distractor was one category and the two less common distractors were grouped into the second category. In this manner, one outcome was the selection of the distractor most commonly selected and the other outcome was the selection of one of the two less common distractors. A total test score on the assessment instrument of interest was standardized and then used as the ability measure and a grouping indicator was used to identify students with and without disabilities. Three models were considered. The first model studied only included the ability measure; the second model included the ability measure, the group indicator (disability status); and third model included the ability measure, the group indicator, and the interaction between the group indicator and ability measure. The improvement in model fit was judged with a test of the significance in the change in the Nagelkerke *R*-square value. The authors used an *R*-square minimum change of 0.003 as a criterion for flagging items as having DDF.

The results of the study indicated that DDF effects were more likely toward the end of the assessment. When DDF was found, it was the result of students with disabilities choosing the least common distractors.

Multinomial analysis of DDF. Kato et al. (2009) proposed the use of multi-step multinomial logistic regression 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. The authors proposed this study to address several limitations of the Abedi et al. (2008) study. Namely, this study provided a separate analysis of each response option, as well as omits, and investigated the DDF effects for students with disabilities by disability category. The related research questions focused on whether reading items exhibit DIF or DDF for students with disabilities, and, if so, whether there is a pattern of DIF or DDF across different disability groups.

For each response option (j) within an item, the model estimates a response characteristic curve (RCC) as a function of ability (z),

$$p_j(z) = \frac{\exp(a_j + b_j z)}{\sum_{j=1}^{J} \exp(a_j + b_j z)} \tag{14}$$

where j=1,2,…, J. The ability measure used in the study was a standardized scale score and is represented in the model by z, and a_j and b_j are the regression coefficients. The shape of the RCC is dependent on the magnitude of the regression coefficients. The slope, b_j, determines the order of the RCCs, with the correct response usually having the largest value. The intercept, a_j, determines the relative popularity of the response option, with larger mean values indicating that the response option is chosen more often.

Two models were analyzed and compared. The first model included the ability measure only under the assumption that there were no group differences. The second model included both the ability measure and group indicator, with one disability category being compared at a time to the reference group. The second model allowed the estimates to vary by group. Nagelkerke's R-square was calculated for each model to determine the variance explained by the models. The models were compared using the likelihood ratio test and the R-square values.

In their study of a statewide reading assessment in third and fifth grade, Kato et al. (2009) identified items as having potential DIF or DDF effects for students with selected disabilities if

the likelihood ratio test was significant at α=.01 and if the *R*-square difference was at least 0.003. Like Abedi et al. (2008) the authors used the *R*-square difference as a measure of effect size, so as not to rely solely on a significance test when sample sizes are so large.

After meeting the preliminary screening criteria, the mean absolute difference (MAD) between the focal group RCC and the reference group RCC was calculated for each of the identified items using

$$MAD_j = \frac{1}{n}\sum_{e=1}^{n}\left|p_{j,0}(z_e) - p_{j,1}(z_e)\right|, \qquad (15)$$

where z_e is the ability measure of examinee *e*, *n* is the total sample size, and the estimated RCCs are $p_{j,0}(z)$ and $p_{j,1}(z)$ for response option *j*. The MADs were used to contribute to the understanding of what may have caused the item to be identified for further analysis, with the assumption that response options with large MADs contributed the most. It also was assumed that if the largest MAD is found in the correct option, the differential functioning is explained by DIF effects. If the largest MAD is found in the distractors, the differential functioning is explained by DDF effects. Plots of the RCCs for items under review were used to visually inspect the direction and magnitude of the differential functioning between groups.

Many of the analyzed items were found to exhibit statistically significant DIF or DDF effects for the groups studied; however, after the *R*-square criterion was applied, the number of items selected for further analysis was reduced. While noting that the use of 0.003 was sufficient for their analysis, the authors suggest that caution should be used in setting the minimum *R*-square difference, due to the potential impact of differences in group sizes. Because their study investigated three subcategories of students with disabilities, with traditionally smaller sample sizes than those of students without disabilities, the group sizes were not equal. This was found to result in smaller *R*-square differences than would have been observed if the groups would have been closer in size.

Kamata and Williams (2006) proposed to study DDF through fitting a multinomial regression model with the addition of an interaction term, with the correct choice as the reference category. Consistent with Kato et al. (2009), the coefficient for the group indicator would indicate the magnitude of the uniform DDF. In addition, the coefficient for the interaction term would indicate the magnitude of nonuniform DDF. The proposed study was not completed and results were not reported.

Factorial modeling of DDF. Wang (2000a) proposed a factorial model of differential distractor functioning to allow for the examination of the effect of multiple grouping factors, as well as interactions between them. The proposed method expands the factorial procedure for studying DIF effects described by Wang (2000b).

The factorial procedure for studying DIF, as discussed by Wang (2000b), is based on a formulation of factorial analysis of variance (ANOVA) to DIF detection. Item difficulties serve as dependent variables and factors serve as the independent variables. DIF is represented by main effects and interaction effects. Beginning within the framework of the Rasch model, item parameters can be estimated as

$$\log\left(p_{ei} / q_{ei}\right)_k = \theta_e - \delta_{ik} \tag{16}$$

where p_{ei} is the probability person e will respond correctly to item i, q_{ei} is the probability of an incorrect response, θ_e is the ability of examinee e, and δ_{ik} is the difficulty of item i by group k, where $k = 1, 2, \ldots, K$. The group indicator, in this case, is referred to as factor A.

Within the framework of ANOVA, the item parameter δ_{ik} can be parameterized as $\delta_{ik} = \delta_i + \alpha_{ik}$ where δ_i is the mean difficulty of item i and α_{ik} is the DIF parameter for the effect of group k on item i. The item is said to exhibit DIF if α_{ik} is different from zero.

This model can be extended to study more than one factor and is then referred to as factorial ANOVA. For example, a gender factor (factor B, l=0, 1) may be added in addition to factor A, the group factor. This model would be estimated by

$$\log\left(p_{ei} / q_{ei}\right)_{kl} = \theta_e - (\delta_i + \alpha_{ik} + \beta_{il} + \alpha\beta_{ikl}) \tag{17}$$

where δ_i is the grand difficulty of item i, α_{ik} is the DIF main effect of factor A_k, β_{il} is the DIF main effect of factor B_l, and $\alpha\beta_{ikl}$ is the DIF interaction effect of factor A_k by factor B_l, on item i. The DIF parameters can be tested for significance using a test of the point estimate or a test of the model fit using the likelihood ratio test.

Wang (2000a) expanded this approach to a study of distractors, naming this approach the factorial procedure for investigating DDF in multiple-choice items. In his model, δ_{ij} is the distractibility parameter for distractor j, and the probability of a response to choice j in item i by a person with ability θ_e is

$$P_{ij}(\theta_e) = \frac{\exp(b_{ij}\theta_e + \delta_{ij})}{\sum_{j=0}^{J_i} \exp(b_{ij}\theta_e + \delta_{ij})} \tag{18}$$

where there are $J_i + 1$ choices in item i, with j=0,..., J_i. The correct response is reserved for j=0, and all other responses are reserved for distractors. In addition, b_{ij} is the score of the response to choice j of item i, where b_{i0}=1 and b_{ij}=0, for j=1,…, J_i.

The log-odds formulation of the model is

$$\log\left(\frac{P_{i1}}{P_{i0j}}\right) = \theta_e - \delta_{ij} \tag{19}$$

where P_{i1} is the probability of being correct (scoring 1) and P_{i0j} is the probability of being incorrect by choosing distractor j (j=1,…, J_i).

When the model is expanded to include multiple factors, the factorial procedure decomposes the distractibility parameters to account for the mean distractibility and DDF effects. To illustrate, Wang (2000a) provided a model with two grouping factors, factor A with K levels and factor B with L levels, and estimated the distractibility parameters, now noted as δ_{ijkl}, as

$$\delta_{ijkl} = \delta_{ij} + \alpha_{ijk} + \beta_{ijl} + (\alpha\beta)_{ijkl}, \tag{20}$$

where δ_{ij} is the grand mean distractibility of distractor j in item i across groups, α_{ijk} and β_{ijl} represent the main effects of factors A and B, respectively, and $(\alpha\beta)_{ijkl}$ is the interaction effect. If any of the DDF effects (main effects or interaction effect) are significant, there is evidence for DDF.

So as not to rely purely on statistical hypothesis testing when sample sizes are large, it was proposed that differences in item difficulty estimates between groups should be considered. Using an acceptable method proposed in the literature, a difference of 0.25 logits was used to identify distractors with substantial DDF.

Wang (2000a) conducted a simulation study to assess parameter recovery of the factorial procedure. The author concluded that the procedure yielded unbiased parameter estimates, using the software ConQuest; however, there was overestimation of the asymptotic error variances for the DDF parameters and the mean-deviation parameters. In addition, the author concluded that this procedure may not be robust in cases where a distractor is rarely selected because there may not be enough information to estimate the distractibility parameter.

Item Response Theory (IRT) Methods

IRT methods for detecting DIF are parametric methods that make use of a latent variable in the definition of null DIF.

Differential alternative functioning. Thissen et al. (1989) proposed that the multiple-choice model (Thissen & Steinberg, 1984) provides a framework for examining differential attractiveness of distractors for members of different groups. The multiple-choice model was derived from the nominal response model (Bock, 1972) and its extension by Samejima (1979). Thissen et. al (1993) extended the use of the multiple-choice model to a test of what they called differential alternative functioning (DAF; Thissen, Steinberg, & Wainer, 1993). Under this framework, the multiple-choice model fits a trace line for each alternative and DAF occurs when trace lines differ between groups. Specifically, the trace line for each alternative j in item i with J_i alternatives is estimated by the following model,

$$T_i(j) = \frac{\exp(a_j\theta + c_j) + d_j \exp(a_0\theta + c_0)}{\sum_{h=0}^{J_i} \exp(a_h\theta + c_h)}. \qquad (21)$$

In this model, $j = 1, \ldots, J_i$ and a_j is the alternative-discrimination parameter, c_j is the alternative-intercept parameter, and d_j (representing "don't know") is the alternative-guessing parameter. For each item, there are $J_i + 1$ a_j and c_j parameters, and one d_j parameter. As described by Thissen et al. (1989), "the idea is that some proportion d_j guess each of the observable response alternatives and that proportion is combined with the examinees who chose those alternatives intentionally" (p. 163). The response category j=0 is reserved for the latent category, d_j.

To demonstrate this approach in comparison to the log-linear approach, Thissen et al (1993) re-examined an item on the SAT found to exhibit DDF by Green et al. (1989). Using the likelihood ratio test, the fit of an unconstrained model allowing parameters to differ was compared to a constrained model where the parameters are equal between the focal and reference groups. Like the results of the log-linear analysis, the DAF analysis results indicated a difference in alternative selection by group. The difference also was evident in the trace lines, which were found to be consistent with the plots produced by the log-linear method.

Relationship of DDF to DIF

Kato et al. (2009) concluded that "DIF might result from DDF, that is, students at a particular ability level in one group tended to choose a particular distractor more often than their counterparts in the other group. This fact suggests that item bias can be avoided by considering not only the behavior of correct responses but also that of distractors" (p. 39). Similarly, Mapuranga et al. (2008) stated that analysis of DDF "has the potential of providing supporting data and analysis to corroborate or refute proposed reasons why subgroup response differences may or may not be construct relevant, or to determine whether DIF might be attributable to specific features of an item (such as a specific distractor)" (p. 14). While the relationship was not clearly articulated, Penfield (2008) stated that "if all *J* DDF effects are zero then it must be the case that the DIF effect is zero (i.e., if all *J* DDF effects are zero, then the conditional between-group difference in the probability of correct response must be zero). Note, however, that it is also possible to have a zero DIF effect despite nonzero DDF effects, for example, if the DDF effects have opposite signs and cancel one another" (p.251).

Penfield (in-press) further studied how uniform and non-uniform DIF effects are a function of DDF effects as modeled under the NRM. Using the model parameterization, two conditions were found to cause uniform DIF. The first condition leading to uniform DIF is when there is a constant DDF effect across all distractors. The second condition is when there is a constant slope parameter, or a_j, across all distractors.

To study how DDF effects are related to non-uniform DIF, Penfield (in-press) conducted a numeric investigation by looking at the variation of the conditional DIF effect as a function of simulated DDF item effects. In this study, the conditional DIF effect was expressed as the conditional log-odds ratio, noted as $\Lambda(\theta)$, and calculated by

$$\Lambda(\theta) = \ln\left[\frac{P(Y=1)\big|\theta,G=1 \,\big/\, 1-P(Y=1)\big|\theta,G=1}{P(Y=j)\big|\theta,G=0 \,\big/\, P(Y=j)\big|\theta,G=0}\right] \tag{22}$$

where Y=1 indicates a correct response, Y=j indicates the selection of a distractor, and G is the grouping variable. When $\Lambda(\theta)$=0, there is no difference in the conditional logs-odds ratio between groups.

A distinction was made in the study between non-uniform DIF and crossing DIF, where crossing DIF is the portion of the non-uniform DIF effect that cancels due to differences in sign. Three items with three distractors each were parameterized under the NRM, with the slope and intercept parameters, a_j and c_j, having specified values. To quantify the overall measure of nonuniform DIF, $\Lambda(\theta)$ was evaluated at each 0.1 interval across the θ range of -4.0 to 4.0, with a total of 81 observation points denoted as $\Lambda(\theta_o)$, where *o* represents each observation point (e.g., $\theta_1 = -4.0$, $\theta_2 = -3.9$). The overall degree of non-uniform DIF (NON) was then represented by the standard deviation of $\Lambda(\theta_o)$ as follows,

$$NON = \sqrt{\frac{\sum_{o=1}^{81}\left[\Lambda(\theta_o) - \mu\right]^2}{81}} \tag{23}$$

where μ is the mean value of $\Lambda(\theta_o)$. Under this formulation, a large value of NON indicates the presence of non-uniform DIF because it is based on the variation of the conditional DIF effect across ability. In the absence of defined criteria, the author proposed that a large non-uniform DIF effect occurs when NON>0.3.

To evaluate crossing DIF (CRS) effects, Penfield (in-press) used the absolute value of the sum of the signed values of $\Lambda(\theta_o)$ divided by the sum of the unsigned values of $\Lambda(\theta_o)$ in the following formulation,

$$CRS = 1 - \frac{\left|\sum_{o=1}^{81}\Lambda(\theta_o)\right|}{\sum_{o=1}^{81}\left|\Lambda(\theta_o)\right|}. \tag{24}$$

In this case, a CRS near zero indicates minimal crossing DIF. A CRS near one indicates a presence of crossing DIF, caused by the cancellation of the conditional signed DIF effects across ability. A large crossing DIF effect was said to occur if CRS>0.5.

A simulation study was conducted for three items under 12 different conditions of DDF, including DDF effects equal in sign but differing in magnitude across distractors and DDF effects that varied in sign and/or magnitude. The results of the study yielded several findings (Penfield, in-press):

- The relationship between a particular set of DDF effects and the resulting DIF effect depends on the particular parameters underlying the item,

- Crossing DIF effects can only occur in the presence of DDF effects that vary in sign, and
- Large non-uniform DIF effects are likely an indication of DDF effects that vary in sign. (p. 18)

Additional results were discussed in the context of practical implications. It was noted that uniform DIF is likely the result of a factor in the item stem or correct option, since the DDF effect is constant across all distractors. On the contrary, non-uniform DIF is likely the result of biasing factors across the distractors.

While these findings are significant, the results are limited to interpretations of data that also fit the NRM. Penfield (in-press) suggests extending this work to determine if these relationships exist under different models for multiple-choice items.

CHAPTER 3

METHODS

This chapter provides a description of the data collection measure, the study sample, and the data analysis procedures for the purpose of comparing three methods for detecting DDF. The comparisons were conducted in such as way as to answer two questions:

1. Is the magnitude and pattern of the DDF effect constant across all methods?
2. Does the pattern of DDF effects support the DIF findings?

Data Collection Measure

Data collected from a 2006 administration of the Grade 3 Florida Comprehensive Assessment Test® (FCAT) Mathematics were used in this study. 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 (referred to as 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.

FCAT Background

The FCAT is Florida's statewide criterion-referenced assessment test in the subjects of reading, mathematics, science, and writing. Both reading and mathematics are assessed in grades 3-10, while science and writing are assessed once at the elementary, middle, and high school levels. FCAT results are used in the calculation of Florida's school grades and in the calculation of Annual Yearly Progress (AYP) under the federal No Child Left Behind (NCLB) Act. 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.

The Grade 3 FCAT Mathematics test design includes 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. These reviews include:

- A review for potential gender, racial, ethnic, linguistic, religious, geographic, or socioeconomic bias; and
- A review for community sensitivity to identify items that may not be acceptable by communities of different cultural, regional, philosophical, political, or religious backgrounds.

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 model is expressed as

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]}, \tag{25}$$

where a_i is the item discrimination parameter, b_i is the item difficulty parameter, and c_i is the pseudo-guessing parameter. FCAT scale scores are reported on a scale of 100-500, as well as developmental scale scores in the range of 0-3000.

The standardized mean difference (SMD; Zwick, Donoghue, & Grima, 1993) statistic was used to evaluate the effect size of DIF results for FCAT Mathematics. The Florida Department of Education uses the magnitude of the SMD statistic to categorize the DIF effect. This rating system includes seven categories, with the first three categories (1-3) representing a small performance difference between the focal and reference group members. These statistics are calculated for the following comparisons: females versus males, African-American students versus White students, and Hispanic students versus White students.

2006 Grade 3 FCAT Mathematics

Approximately 204,000 students participated in the 2006 Grade 3 FCAT Mathematics, with a distribution of 45% White, 25% Hispanic, 23% African-American, 2% Asian, 4% Multiracial, and less than 1% American Indian (Florida Department of Education, 2007). Males represented 51.53% of the cohort, while females represented 48.21%. Approximately 53% of these students were served by free or reduced-priced lunch programs. All 40 items on the assessment were classified as having a low DIF rating, as measured by the SMD. This finding indicates that, for the comparison groups studied, the DIF effect was found to be negligible. It is expected that the DIF effect would be negligible for these comparison groups, since most items with larger DIF would have been removed during the test construction process.

Preliminary Data Analyses

Missing Data and Violation of Assumptions

To ensure the validity of the results, steps were taken to identify and resolve issues with the data file. An analysis of the extent of missing data, as well as the identification of possible problematic observations, was conducted. Cases that did not have a reported scale score or did not match documented demographic characteristics (e.g., free or reduced price lunch indicator) were deleted from the data file. Missing data in item response fields were treated according to the DIF or DDF analysis procedures. In the case of DIF analyses, no responses are coded as incorrect. While a separate analysis of no responses is available through the DDF analysis methods used in this study, separate analyses were not conducted unless the no response rate exceeded 2.5%.

DIF Analyses

Both Dorans et al. (1992) and Penfield (in-press) argued that the value of DDF analyses is primarily found as a supplement to a DIF analysis to investigate where in the test item the DIF may be occurring. 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 nonuniform DIF.

As described previously, the logistic regression model for each item is expressed by

$$\ln\left[\frac{p_e}{(1-p_e)}\right] = \beta_0 + \beta_1 X_e + \beta_2 G_e + \beta_3 (XG)_e, \qquad (26)$$

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 matching criterion was the FCAT scale score.

The first model was estimated with only X_e in the model. The 2nd model included both X_e and G_e. The 3rd model included X_e, G_e, and $(XG)_e$. Each model was analyzed separately so that a chi-square statistic was estimated based on the log-likelihood ratio.

The chi-square statistic was used to test for uniform and nonuniform DIF. To test for both uniform and nonuniform DIF, the chi-square difference test was used to test for a difference between the 3rd and 1st models, by $x^2_{DIF} = x^2_{3rd\ model} - x^2_{1st\ model}$.

When the test was significant, based on a chi-square distribution with 2 degrees of freedom, a test for uniform DIF, based on 1 degree of freedom, was conducted by comparing the 2nd and 1st models, $x^2_{uniform\ DIF} = x^2_{2nd\ model} - x^2_{1st\ model}$. Non-uniform DIF was tested with the difference between the 3rd and 2nd models, based on 1 degree of freedom using $x^2_{DIF} = x^2_{3rd\ model} - x^2_{1st\ model}$.

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 $\hat{\alpha}_{MH_i}$ scale. These results were summarized and used to inform the answers to the two study research questions.

Magnitude and Pattern of DDF Effects

A systematic comparison of three methods for detecting DDF was conducted to determine if the magnitude and pattern of the DDF effect was constant across all three methods. The three methods that were compared are the standardized distractor analysis (STD) method, the odds-ratio (OR) method, and the multinomial logistic regression (MLR) method. Students classified as participating in free and reduced-priced lunch programs served as the focal group, while students not participating in these programs served as the reference group. This comparison group is important to study because participation in these programs is used as a proxy for economic disadvantage. Economically-disadvantaged students represent one subgroup of students that are included in the calculation of Adequate Yearly Progress decisions under the No Child Left Behind Act.

DDF Analyses

Standardized distractor analysis (STD). STD, as applied by Schmitt and Bleistein (1987), was used to estimate DDF effects.

The standardized 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

$$STD\ P\text{-}DIF\ (j) \quad = \sum_{s=1}^{S} w_s [P_{fs}(j) - P_{rs}(j)] \Big/ \sum_{s} w_s \ , \tag{27}$$

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 and reference group members, respectively, selecting distractor *j*. The standard weight, w_s, applied to both the focal and reference group proportions, was equal to n_{fs}, the number of examinees at *s* in the focal group. Following convention, this weighting factor was selected to apply the greatest weight to differences at score levels most obtained by the focal group. The FCAT scale score was used as the matching criterion (*s*) and score levels were not grouped. This method was used so that the information was maximized and consistent across all approaches.

A modification of the STD D-DIF for distractor *j* (Dorans, 1989) will be used to place the STD P-DIF values on the odds-ratio scale (i.e., the odds ratio will not be multiplied by -2.35ln). This standardization index was referred to by Wright (1986) as α_{STD}. The magnitude of the odds ratio was aligned to the ETS categories of A, B, and C, by Monahan et al. (2007) as described in Table 3.1. This classification was used in summarizing the magnitude of the DDF effects.

Table 3.1
Alignment of Odds Ratio to ETS Classification

Category	Odds Ratio Index
A	*0.65 ≤ Odds Ratio ≤ 1.53*
B	*Odds Ratio < 0.65 or Odds Ratio > 1.53*
C	*Odds Ratio < 0.53 or Odds Ratio > 1.89*

Odds Ratio (OR). 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 the correct answer and the *j*th distractor), the conditional odds ratio across all score levels was estimated by

$$\hat{\alpha}_j = \frac{\sum_{s=1}^{S} R_{1s} F_{js} / n_s}{\sum_{s=1}^{S} R_{js} F_{1s} / n_s}. \tag{28}$$

The FCAT scale score was used as the matching criterion (*s*) and score levels were not grouped. DDF effects were judged according to the ETS classification categories, as aligned to the odds-ratio index. Monahan et al. (2007) provided the equivalencies found in Table 3.1.

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

$$\ln\left[\frac{p_{je}}{(1-p_{je})}\right] = \beta_{j0} + \beta_{j1} X_e + \beta_{j2} G_e \tag{29}$$

where *j*=1,2,…, *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). 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. The results were summarized using the classification system provided in Table 3.1.

Comparison of Results

The results of the three methods were summarized by item and distractor, using several approaches. 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, a Pearson correlation

coefficient was used. These comparisons were made to the STD approach because it is a recognized 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 summarized for use 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, 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. Correlation indices were used to summarize the consistency of the data.

The expected results of this study were informed by studies that have investigated DIF and DDF detection methods. The STD and MH methods for detecting DIF are based on the same data table and have been found to have very highly related results. Their key differences are that the STD approach compares the probability of success while the MH method compares the odds of success, and the two methods use different weighting functions. Despite these differences that result in slight shifts in the scales, the correlation between the two indices has been found to be at least .94 (Wright, 1986).

Swaminathan and Rogers (1990) found that the logistic regression method for detecting uniform DIF is as powerful as the MH method and, in comparison, more powerful in detecting nonuniform DIF. A simulation study, using data generated with a three-parameter item response model, showed a logistic regression uniform DIF detection rate of about 75% accuracy in groups of 250 and 100% accuracy in groups of 500. Under nonuniform DIF, the MH method was completely unable to detect DIF while the logistic regression method performed with about 50% accuracy in groups of 250 and 75% accuracy in groups of 500. They noted that the MH method can be expected to be more successful in detecting nonuniform DIF when the interaction occurs at the low or high ends of the ability scale, instead of when crossing DIF occurs in the middle of the ability scale, as was simulated.

In discussing approaches to detecting DDF, Dorans et al. (1992) compared STD to DAF. When sample size is sufficient and the matching variable is reliable, the standardization approach is preferred because no parametric assumptions are required. On the other hand, when the sample size is not sufficient, an item response model fits the data, and the matching variable is DIF-free, the DAF method may be preferred.

Penfield (2008) proposed and studied the OR approach to detecting DDF under data generated by the NRM and the MCM. When groups of equal ability were studied under the NRM, the odds-ratio estimator of DDF was found to be relatively unbiased. Under the condition of unequal abilities, the OR estimator remained relatively unbiased, with the magnitude of the bias ranging from 0.05 to 0.10. The performance of the OR estimator was not as good when data were generated under the MCM. Under conditions of equal ability, there was a bias toward zero when the true DDF was non-zero. When the true DDF effect was 0.4, the bias was less than 0.1, but the magnitude of the bias increased to 0.3, in some cases, when the true DDF effect was 0.8. This bias increased under the condition of unequal ability distributions, with the bias reaching to 0.4 in some cases. Because of this bias, it was concluded that the OR estimator may serve as a conservative estimator of DDF effects when the data fit a model that incorporates guessing.

Based on the findings discussed in this section, it was expected that all three methods would yield similar results when the DDF effects are small and the sample size is sufficient. When the DDF effects increase in magnitude, it was expected that both the OR and STD approaches would underestimate the DDF effect.

Relationship Between DIF and DDF

Penfield (in-press) 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 can only occur in the presence of DDF effects that vary in sign. He proposed that these findings can "help target the particular item property responsible for the DIF effect" (Penfield, in-press). For items with significant uniform DIF, the item stem or correct option may be the source of the DIF effect. For items with significant nonuniform DIF, the distractors are likely the source of the DIF effect. This portion of the study was conducted to determine if the pattern of DDF effects inform the DIF findings, and therefore support Penfield's findings under different conditions and approaches.

Uniform DIF and DDF

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. The level of agreement between the pattern of DDF effects and indication of uniform DIF was summarized by method and then compared across all three methods.

Nonuniform and Crossing DIF and DDF

All items identified as having significant nonuniform DIF in the logistic regression analysis were plotted to determine if the nonuniform 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. The results of the DDF analyses indicating the pattern of DDF effects by method by item were compared to the DIF results for items with nonuniform and crossing DIF. The level of agreement between the pattern of DDF effects and indication of nonuniform and crossing DIF was summarized by method and then compared across all three methods.

Item Characteristics

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. Relationships were summarized using correlation indices.

Utility of DDF Effect Information

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 primary emphasis in the exploration was to determine if the DDF results could be explained by any plausible reasons, such as a distractor representing a common misconception. In other words, the exploration focused on the practical use of DDF results in the item revision process. The results of this exploration were summarized in a qualitative manner.

Summary

The methods described in this chapter are focused on the goal of discussing and making recommendations about DDF methodology choice and use for practitioners. Most statewide, high-stakes assessment programs use IRT methods and the results of this study will be applicable to these programs.

CHAPTER 4

RESULTS

Preliminary Data Analyses

Sample

Of the 206,678 cases in the 2006 Grade 3 FCAT Mathematics data file, 202,140 cases were used in this study. A total of 3,871 cases was 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. A variable in the file indicated whether the demographic data matched (including whether cases matched on gender, race, and whether the student received free or reduced-price lunch services). 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 FCAT scale scores were not reported. FCAT scale scores are required as the matching criterion.

The variable "Lunch" was originally coded as "Y" to indicate that the student participated in free or reduced-price lunch services and blank otherwise. This variable was recoded as "Y"=1 and blank=0. The focal group for this study is identified based on the Lunch status, with students with Lunch=1 serving as the focal group and those with Lunch=0 serving as the reference group. Table 4.1 provides a summary of the final sample.

The mean scale score for the focal group was 304 and the mean scale score for the reference group was 347. Table 4.2 describes the overall achievement of both groups and the total group using the mean FCAT scale score and standard deviation. FCAT achievement also is interpreted using achievement levels, with Level 1 the lowest level and Level 5 the highest level. Table 4.3 describes the achievement of both groups and the total group by achievement level.

Missing Data

Other than the cases with missing scale scores and inconsistencies in Lunch status, all cases were used in the analyses. 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.

Table 4.1
Descriptive Statistics for Gender and Race by Lunch Status

	Lunch = 1		Lunch =0		Total Sample	
	n	%	*n*	%	*n*	%
Male	56470	51.7	47931	51.6	104401	51.6
Female	52711	48.3	45028	48.4	97739	48.4
White	29722	27.2	62267	67.0	91989	45.5
Black	37791	34.6	9331	10.0	47122	23.3
Hispanic	35762	32.8	14326	15.4	50088	24.8
Mulitracial	3994	3.6	3802	4.1	7746	3.8
Asian	1660	1.5	2924	3.1	4584	2.3
American Indian	302	.3	309	.3	611	.3

Table 4.2
Mean FCAT Scale Score by Lunch Status

	n	Mean FCAT Scale Score	Standard Deviation
Lunch=1	109181	304.41	62.42
Lunch=0	92959	347.01	59.63
Total Sample	202140	324.00	64.73

Table 4.3
Percent in each Achievement Level by Lunch Status

	Level 1	Level 2	Level 3	Level 4	Level 5
Lunch=1	17.8	20.8	36.7	20.2	4.6
Lunch=0	5.7	10.0	31.7	35.8	16.9
Total Sample	12.2	15.8	34.4	27.3	10.2

Measurement Instrument

The standard error of measurement (SEM) for the 2006 Grade 3 FCAT Mathematics ranged across achievement levels, with larger SEM values at the lowest and highest achievement level cut scores. Table 4.4 summarizes the SEM by cut score as reported by the Florida

Department of Education (2007). The IRT marginal reliability was .927, and the reliability as measured by Cronbach's alpha was .900 (Florida Department of Education, 2007).

Table 4.4
SEM by Achievement Level Cut Score: 2006 Grade 3 FCAT Mathematics

	Level ½	Level 2/3	Level 3/4	Level 4/5
SEM	21	16	15	20

DIF Analyses

All 40 multiple-choice items in the data file were analyzed. All of the items were dichotomously scored by the Florida Department of Education, with a correct response coded as 1 and an incorrect response coded as 0. The FCAT scale score, included in the data file, served as the matching criterion for each of the three models estimated. The first model was estimated with only the matching criterion, the FCAT scale score, in the model. The second model included both the matching criterion and the group indicator (Lunch status). The final model, the 3rd model, included the FCAT scale score, Lunch status, and the interaction between the FCAT scale score and Lunch status. Each model was analyzed separately so that a chi-square statistic was estimated based on the log-likelihood ratio. These chi-square statistics were used to test for uniform and nonuniform DIF. 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 odds-ratio scale. In addition, when significant, the magnitude of the nonuniform DIF was evaluated at the four scale scores which are the cut scores separating the five FCAT Grade 3 achievement levels. The four cut scores, from lowest to highest, are: 253, 294, 346, and 398.

Evidence for DIF existed for all forty items. Of the forty items, thirty-four items had statistical evidence of small, uniform DIF (classified as A DIF). Thirty of these items also had statistical significance for nonuniform DIF, and seventeen of these items displayed medium to large nonuniform DIF effects across the ability scale. The results for six items indicated only significant nonuniform DIF. For those items with significant nonuniform DIF, the magnitude of the nonuniform DIF at the four FCAT cut scores was determined in the odds-ratio scale. The results of the analyses are provided in Table 4.5.

Table 4.5
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
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 = nonuniform 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 "-".

The predicted probability of achieving a correct response was saved for each case and graphed for both the focal and reference groups. As shown in Table 4.5, many of items exhibited crossing DIF. Evidence of crossing DIF can be seen when the odds ratio shifts from lower odds to greater odds, or vice versa, across the four cut scores. Figure 4.1 shows a graph of item #38. This graph supports the evidence of crossing DIF found in Table 4.5.

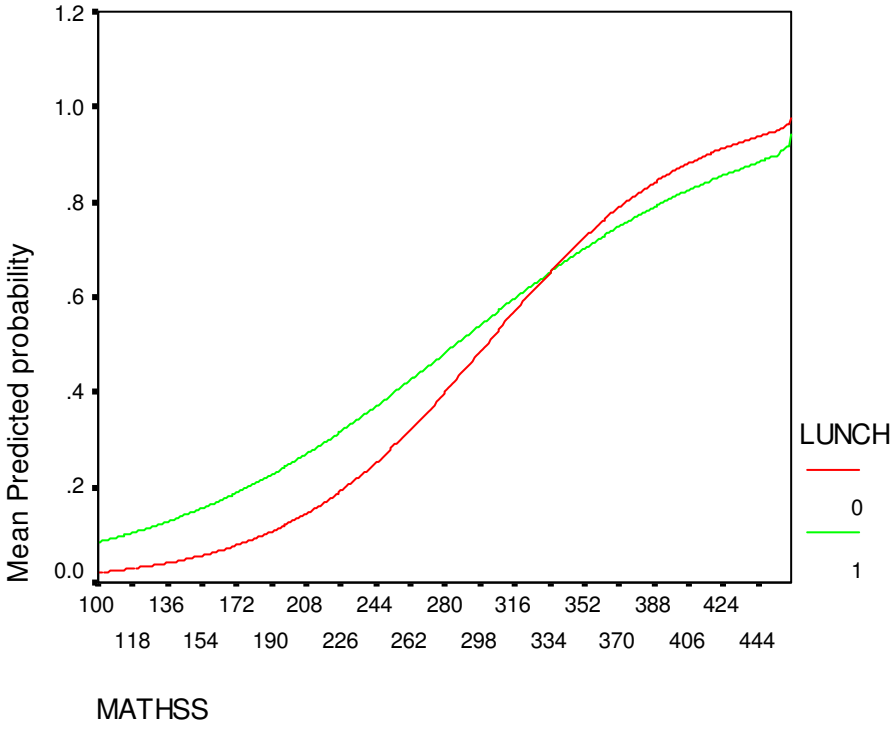


Figure 4.1. DIF results for item #38.

Magnitude and Pattern of DDF Effects

A comparison of the standardized distractor analysis (STD), odds ratio (OR), and multinomial logistic regression (MLR) methods for detecting DDF was conducted to determine if the magnitude and pattern of the DDF effects is constant across all three methods. Students classified as participating in free or reduced-price lunch programs served as the focal group, while students not participating in these programs served as the reference group.

STD Results

STD was conducted for all 40 test items, with separate estimates for each distractor, resulting in 120 estimates. The FCAT scale score was used as the matching criterion. While the effect-size estimates did range in size, none of the distractors were classified as having B or C DIF effects (see Table 4.6 for the classification criteria).

Table 4.6
Classification of DDF Effect-Size Estimates in Odds-Ratio Scale

Category	Odds Ratio Index
C+	*Odds Ratio > 1.89*
B+	*Odds Ratio > 1.53*
A+	*1.00 < Odds Ratio ≤ 1.53*
A-	*0.65 ≤ Odds Ratio < 1.00*
B-	*Odds Ratio < 0.65*
C-	*Odds Ratio < 0.53*

Table 4.7 provides the estimates by item in the odds-ratio scale. The estimates in the table 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 ($\alpha_{STD_{31-1}} = 1.1964$).

OR Results

DDF analyses were conducted using the OR approach for all 40 test items, with separate estimates for each distractor, again resulting in 120 estimates. The FCAT scale score was used as the matching criterion. While the effect-size estimates did range in size, no item options were classified as having B or C DIF effects. Table 4.8 provides the estimates by item in the odds-ratio scale. The estimates in the table 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

Table 4.7
STD Results by Item and Distractor

Item	Distractor 1	Distractor 2	Distractor 3
1	0.888	1.107	1.010
2	1.048	0.921	0.928
3	0.892	0.910	1.102
4	1.106	0.946	0.875
5	1.019	0.988	0.985
6	0.977	0.962	0.962
7	1.048	1.002	1.110
8	0.981	1.001	0.830
9	1.068	0.850	0.907
10	1.037	1.067	0.995
11	1.105	1.132	1.123
12	0.884	1.091	1.044
13	1.041	1.105	0.934
14	0.981	0.997	0.968
15	1.029	0.908	1.035
16	1.068	0.963	0.958
17	1.017	1.128	1.043
18	0.966	1.035	1.011
19	1.024	1.010	0.991
20	1.003	1.087	1.047
21	0.886	1.049	1.088
22	1.079	1.102	1.000
23	1.051	0.944	1.002
24	1.007	0.978	1.057
25	1.045	0.977	1.014
26	0.812	1.022	1.121
27	0.998	1.058	0.965
28	1.130	0.863	0.837
29	0.832	0.772	0.877
30	1.182	0.896	1.032
31	1.196	1.159	0.934
32	1.141	1.038	0.932
33	1.004	1.062	0.952
34	1.171	0.995	1.009
35	0.987	1.006	0.989
36	0.956	1.106	0.984
37	0.985	0.905	0.947
38	1.173	1.048	1.103
39	1.031	0.996	0.976
40	0.935	0.904	0.918

Note: Distractor 2 of item 29 is shaded to indicate that it was the least attractive to the reference group and distractor 1 of item 31 is shaded to indicate that it was the most attractive to the reference group.

odds of choosing the correct response over the studied distractor. This interpretation is different than that of the standardized 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 ($\lambda_{38\text{-}1} = .797$).

MLR Results

DDF analyses were conducted using the MLR approach for all 40 test items, with separate estimates for each distractor, resulting in 120 estimates. The FCAT scale score was again used as the matching criterion. The first model analyzed 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.

While the effect-size estimates varied, no distractors were classified as having B or C DIF effects. Table 4.9 provides the estimates by item in the odds-ratio scale. The estimates in the table 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. Thus these values will often be larger than 1 when the odds-ratio indices in Table 4.8 are below 1. 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 [$\exp(\hat{\beta}_{17-22}) = 1.230$].

Table 4.8
OR Results by Item and Distractor

Item	Distractor 1	Distractor 2	Distractor 3
1	1.135	0.889	0.964
2	0.975	1.088	1.098
3	1.153	1.178	0.939
4	0.990	1.142	1.197
5	0.983	0.969	1.068
6	1.023	1.096	1.045
7	0.938	0.952	0.862
8	1.111	1.060	1.263
9	0.977	1.249	1.155
10	0.920	0.872	0.952
11	0.841	0.808	0.824
12	1.112	0.896	0.970
13	0.949	0.913	1.063
14	1.032	1.006	1.068
15	1.020	1.087	1.007
16	0.949	1.096	1.061
17	0.882	0.804	0.850
18	1.040	0.967	0.990
19	0.977	0.993	0.997
20	0.976	0.912	0.952
21	1.195	0.993	0.980
22	0.859	0.851	0.934
23	0.987	1.025	0.972
24	0.998	1.014	0.945
25	0.957	0.974	0.957
26	1.227	0.999	0.933
27	0.982	0.946	1.042
28	1.002	1.285	1.262
29	1.280	1.388	1.190
30	0.854	1.142	1.046
31	0.807	0.812	0.984
32	0.882	0.931	1.043
33	0.970	0.945	1.080
34	0.809	0.982	0.916
35	1.029	1.007	1.019
36	1.016	0.897	0.994
37	1.084	1.148	1.107
38	0.797	0.906	0.845
39	0.960	0.991	1.041
40	1.126	1.168	1.210

Note: Distractor 2 of item 29 is shaded to indicate that it was the least attractive to the reference group and distractor 1 of item 38 is shaded to indicate that it was the most attractive to the reference group.

Table 4.9
MLR Results by Item and Distractor

Item	Distractor 1	Distractor 2	Distractor 3
1	0.872	1.077	0.954
2	0.967	0.885	0.859
3	0.829	0.792	1.072
4	0.988	0.860	0.829
5	0.952	1.033	0.892
6	0.941	0.928	0.959
7	0.992	1.028	1.139
8	0.884	0.852	0.701
9	0.998	0.741	0.836
10	0.974	1.104	1.037
11	1.167	1.182	1.215
12	0.876	1.149	0.998
13	1.027	1.051	0.887
14	0.935	0.973	0.895
15	0.968	0.867	0.919
16	0.996	0.844	0.869
17	1.115	1.230	1.147
18	0.936	1.019	0.957
19	0.910	0.980	0.890
20	0.991	1.181	1.037
21	0.791	0.973	1.058
22	1.141	1.087	1.004
23	0.959	0.886	1.005
24	0.975	0.940	1.040
25	1.025	1.001	0.969
26	0.697	0.930	1.005
27	0.988	1.058	0.907
28	1.017	0.771	0.788
29	0.707	0.654	0.776
30	1.184	0.813	0.937
31	1.187	1.200	0.967
32	1.163	1.077	0.893
33	0.992	1.022	0.883
34	1.133	1.028	1.013
35	0.910	0.867	0.868
36	0.951	1.118	0.951
37	0.820	0.797	0.903
38	1.157	1.041	1.105
39	0.949	0.982	0.908
40	0.824	0.791	0.732

Note: Distractor 2 of item 29 is shaded to indicate that it was the least attractive to the reference group and distractor 2 of item 17 is shaded to indicate that it was the most attractive to the reference group.

Comparison of Results

To judge the consistency of the DDF effect-size estimates for each distractor from the OR and STD approaches, and the MLR and STD approaches, measures of correlation were used. To determine if the relationships were linear, plots were first examined. Figures 4.2 through 4.4 show the linear relationship between the indices from each approach after being transformed to the log-odds scale.

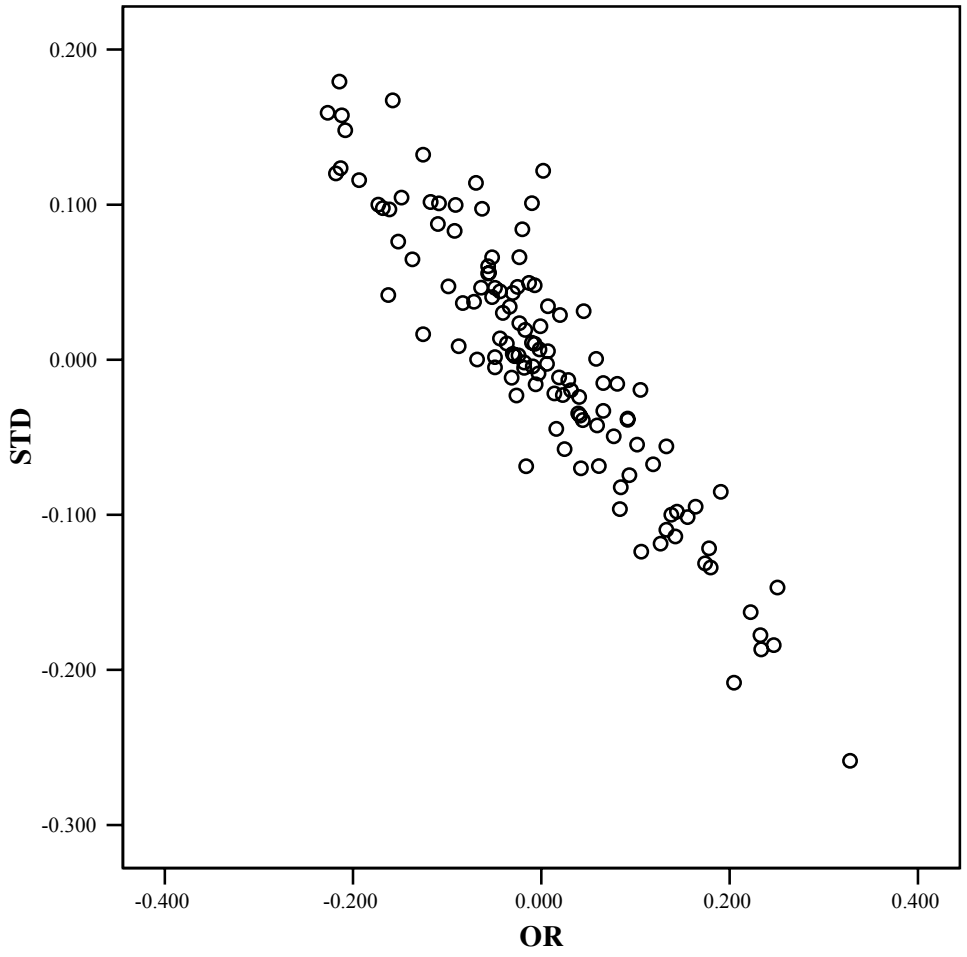


Figure 4.2: Plot of STD and OR Effect-Size Estimates for all Distractors

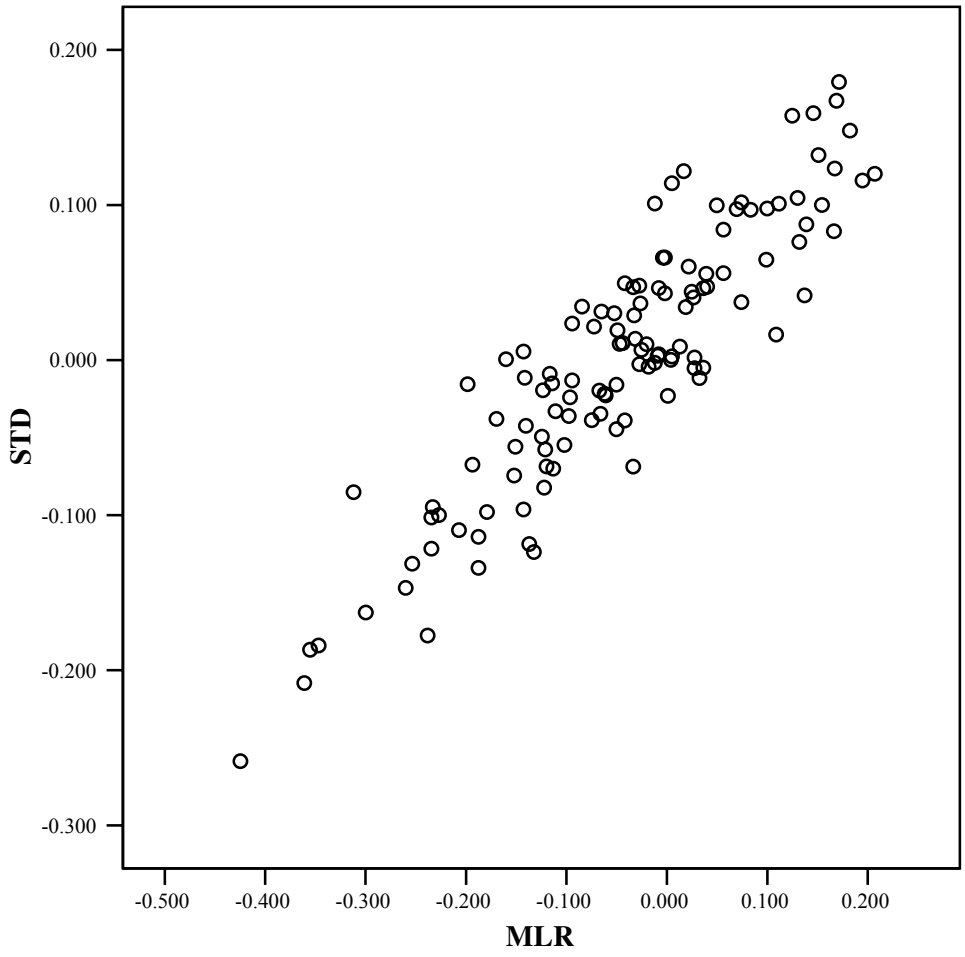


Figure 4.3: Plot of STD and MLR Effect-Size Estimates for all Distractors

The linear relationships identified in the plots also were consistent with values of Spearman rank correlations and Pearson correlations. The Spearman rank correlations compare the ranks of the estimates from one approach to the ranks generated under a second approach. Like the Spearman rank correlations, the Pearson correlations were all significant and indicate a strong linear relationship. The patterns of both measures of correlation were similar, with the highest correlation between the OR and MLR approaches. Both of these approaches are based on contrasts between the distractor and correct answer. The correlations are summarized in Table 4.10.

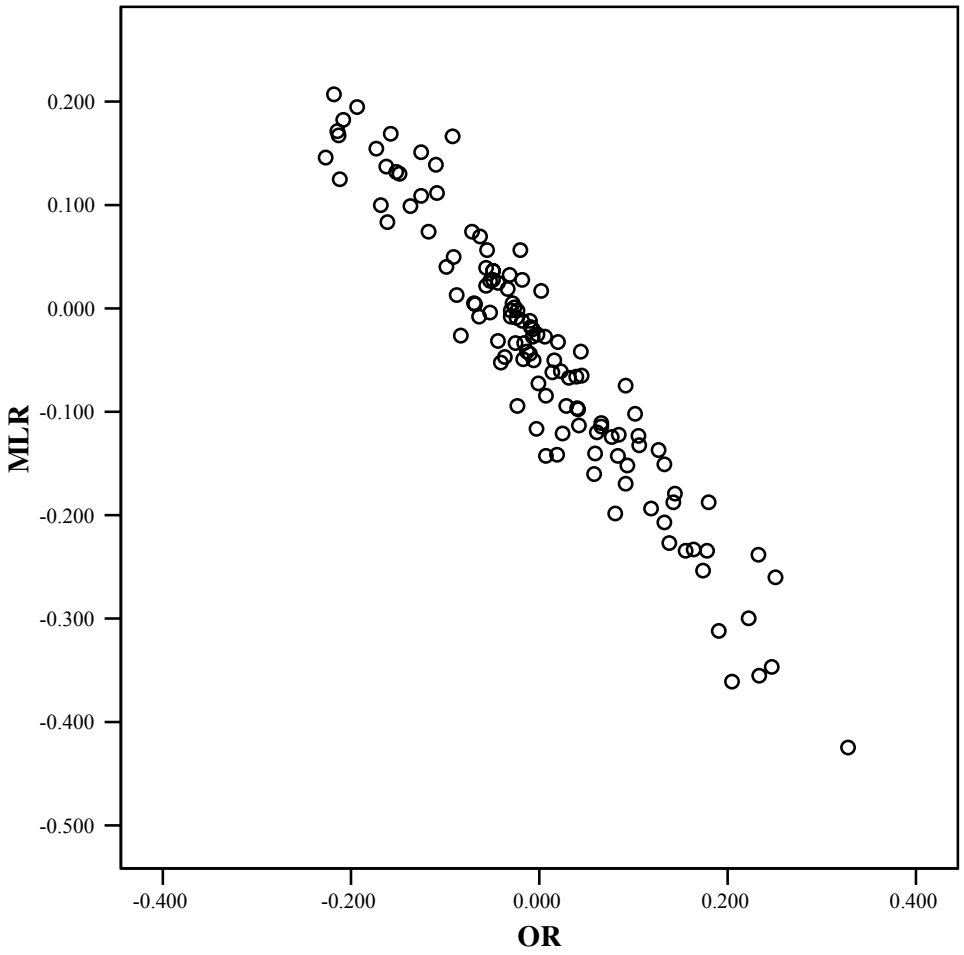


Figure 4.4: Plot of OR and MLR Effect-Size Estimates for all Distractors

Table 4.10
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

** Correlation is significant at the 0.01 level (2-tailed).

The DDF effect-size estimates were transformed to the log-odds scale so that the consistency between the standardization approach and other two approaches could be examined

through summarizing the 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 4.11 summarizes the correlation between the effect-size means. All of the methods have very high correlations, indicating a strong linear relationship in the effect-size means.

Table 4.11
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

** Correlation is significant at the 0.01 level (2-tailed).

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. Table 4.12 shows the Spearman and Pearson correlations.

Table 4.12
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

** Correlation is significant at the 0.01 level (2-tailed).

As mentioned previously, items were classified as having divergent distractor-level effects if at least two of the three DDF effects within each item included a positive and negative effect, after transforming the indices to the log-odds scale. The significance of a linear relationship between the methods in the identification of divergent DDF effects was tested with a Phi correlation. As provided in Table 4.13, both the OR and MLR methods were found to have a statistically significant relationship with the STD approach. The strongest relationship was between the OR and STD methods. The OR and MLR methods were found to have a positive correlation of medium magnitude.

Table 4.13
Phi Correlation Between Indicators of DDF Divergence

Index	STD	OR	MLR
STD	1	.616**	.423**
OR		1	.451**
MLR			1

** Correlation is significant at the 0.01 level. *n=40.*

Patterns in the direction of DDF effect-size estimates were further examined by coding the effect-size estimates as positive or negative using the log-odds scale indices, and when available, noting the significance of the estimates. 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. Agreement is indicated with shading. 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 conceptualization.

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. Table 4.14 shows the pattern for the OR and STD 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%. As shown in Table 4.15, differences included those items found

Table 4.14
DDF Effect-Size Patterns: STD and OR

	Distractor 1		Distractor 2		Distractor 3	
Item	STD	OR	STD	OR	STD	OR
1	-	++	+	--	+	--
2	+	-	-	++	-	++
3	-	++	-	++	+	--
4	+	-	-	++	-	++
5	+	-	-	-	-	+
6	-	+	-	++	-	+
7	+	--	+	--	+	--
8	-	++	+	+	-	++
9	+	-	-	++	-	++
10	+	--	+	--	-	--
11	+	--	+	--	+	--
12	-	++	+	--	+	--
13	+	--	+	--	-	++
14	-	++	-	++	-	++
15	+	+	-	++	+	+
16	+	--	-	++	-	++
17	+	--	+	--	+	--
18	-	++	+	-	+	-
19	+	-	+	-	-	-
20	+	-	+	--	+	-
21	-	++	+	-	+	-
22	+	--	+	--	+	--
23	+	-	-	+	+	-
24	+	-	-	+	+	--
25	+	--	-	-	+	--
26	-	++	+	-	+	--
27	-	-	+	--	-	++
28	+	+	-	++	-	++
29	-	++	-	++	-	++
30	+	--	-	++	+	+
31	+	--	+	--	-	-
32	+	--	+	--	-	+
33	+	-	+	--	-	+
34	+	--	-	-	+	--
35	-	+	+	+	-	+
36	-	+	+	--	-	-
37	-	++	-	++	-	++
38	+	--	+	--	+	--
39	+	-	-	-	-	+
40	-	++	-	++	-	++

Note: Patterns in the direction of DDF effect-size estimates were coded as positive or negative using the log-odds scale indices. 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. Agreement is indicated with shading.

Table 4.15
DDF Effect-Size Patterns: STD and MLR

	Distractor 1		Distractor 2		Distractor 3	
Item	STD	MLR	STD	MLR	STD	MLR
1	-	--	+	++	+	--
2	+	--	-	--	-	--
3	-	--	-	--	+	++
4	+	-	-	--	-	--
5	+	--	-	++	-	--
6	-	-	-	--	-	-
7	+	-	+	++	+	++
8	-	--	+	--	-	--
9	+	-	-	--	-	--
10	+	-	+	++	-	++
11	+	++	+	++	+	++
12	-	--	+	--	+	--
13	+	--	+	--	-	--
14	-	--	-	--	-	--
15	+	--	-	--	+	--
16	+	-	-	--	-	--
17	+	++	+	++	+	++
18	-	--	+	+	+	--
19	+	--	+	-	-	--
20	+	-	+	++	+	+
21	-	--	+	-	+	++
22	+	++	+	++	+	+
23	+	--	-	--	+	+
24	+	-	-	--	+	++
25	+	+	-	+	+	--
26	-	--	+	--	+	+
27	-	-	+	++	-	--
28	+	+	-	--	-	--
29	-	--	-	--	-	--
30	+	++	-	--	+	--
31	+	++	+	++	-	--
32	+	++	+	++	-	--
33	+	-	+	+	-	--
34	+	++	-	+	+	+
35	-	--	+	--	-	--
36	-	--	+	++	-	-
37	-	--	-	--	-	--
38	+	++	+	++	+	++
39	+	--	-	-	-	--
40	-	--	-	--	-	--

Note: Patterns in the direction of DDF effect-size estimates were coded as positive or negative using the log-odds scale indices. 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. Agreement is indicated with shading.

Table 4.16
DDF Effect-Size Patterns: OR and MLR

	Distractor 1		Distractor 2		Distractor 3	
Item	OR	MLR	OR	MLR	OR	MLR
1	++	--	--	++	--	--
2	-	--	++	--	++	--
3	++	--	++	--	--	++
4	-	-	++	--	++	--
5	-	--	-	++	+	--
6	+	-	++	--	+	-
7	--	-	--	++	--	++
8	++	--	+	--	++	--
9	-	-	++	--	++	--
10	--	-	--	++	--	++
11	--	++	--	++	--	++
12	++	--	--	--	--	--
13	--	--	--	--	++	--
14	++	--	++	--	++	--
15	+	--	++	--	+	--
16	--	-	++	--	++	--
17	--	++	--	++	--	++
18	++	--	-	+	-	--
19	-	--	-	-	-	--
20	-	-	--	++	-	+
21	++	--	-	-	-	++
22	--	++	--	++	--	+
23	-	--	+	--	-	+
24	-	-	+	--	--	++
25	--	+	-	+	--	--
26	++	--	-	--	--	+
27	-	-	--	++	++	--
28	+	+	++	--	++	--
29	++	--	++	--	++	--
30	--	++	++	--	+	--
31	--	++	--	++	-	--
32	--	++	--	++	+	--
33	-	-	--	+	+	--
34	--	++	-	+	--	+
35	+	--	+	--	+	--
36	+	--	--	++	-	-
37	++	--	++	--	++	--
38	--	++	--	++	--	++
39	-	--	-	-	+	--
40	++	--	++	--	++	--

Note: Patterns in the direction of DDF effect-size estimates were coded as positive or negative using the log-odds scale indices. Estimates that were significant at .05 were coded as "++" or "--" and estimates not found to be significant were coded as "+" or "-". Agreement is indicated with shading.

to be significant by the MLR model. Table 4.16 shows that the pattern of the OR and MLR estimates are slightly more consistent at 75.8%.

Relationship Between DIF and DDF

Uniform DIF and DDF

Using the logistic regression DIF results, 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 4.16 shows the DIF effect-size estimates in the log-odds scale for items 3, 20, 28, and 36. In addition, the DDF effect-size range by item and method are summarized in Table 4.17. The effect-size range by method (i.e., the STD Range, OR Range, and MLR Range) was determined by taking the difference between the highest and lowest DDF effects estimated for each item. None of the items exhibited equal DDF effects across the item distractors, as evidenced by the non-zero range estimates. The nonexistence of equal DDF effects across each item's distractors was consistent across methods.

Table 4.17
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 constant effect across all distractors. This is important to study because Penfield (in-press) modeled uniform DIF under the NRM and two conditions were found to cause uniform DIF. The first condition was when there was a constant DDF effect across all distractors and the second condition was when there was a constant slope parameter across all distractors. For these four items, no patterns were found to indicate that smaller ranges were indicative of only uniform DIF (see Figure 4.5).

Patterns in the direction of DDF effect-size estimates for these items were examined by using the coding discussed previously. For the purpose of this analysis, OR and MLR estimates

that were not found to be significant are not included. Given this, Table 4.18 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.

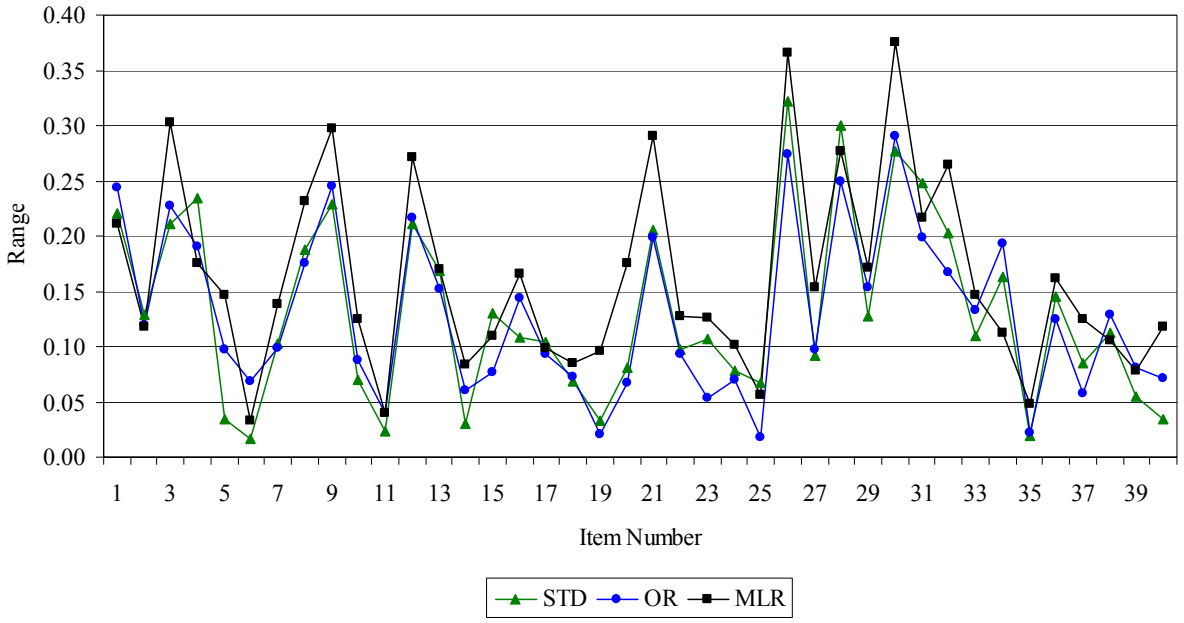


Figure 4.5: DDF Effect-Size Ranges by Item and Method

Table 4.18
Distractor Effect-Size Patterns for Items with Uniform DIF Only

	STD			MLR			OR		
Item	1	2	3	1	2	3	1	2	3
3	-	-	+	--	--	++	++	++	--
20	+	+	+		++			--	
28	+	-	-		--	--		++	++
36	-	+	-	--	++			--	

Note: 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.

Nonuniform and Crossing DIF and DDF

All items identified as having significant nonuniform DIF in the logistic regression analysis were plotted to determine if the nonuniform DIF also is 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.

To aid in the interpretation of the nonuniform DIF effect sizes across the ability scale, the DIF estimates were categorized according to Table 4.19.

Table 4.19
Coding of DIF Nonuniform Effect-Size Estimates

Category	Odds Ratio Index
C+	*Odds Ratio > 1.89*
B+	*Odds Ratio > 1.53*
A+	*1.00 < Odds Ratio ≤ 1.53*
A-	*0.65 ≤ Odds Ratio < 1.00*
B-	*Odds Ratio < 0.65*
C-	*Odds Ratio < 0.53*

Two items were found to have significant nonuniform DIF with no crossing DIF. Figures 4.6 and 4.7 display the graphs of these items. As shown in Table 4.20, the DDF pattern was consistent across all three methods for these items.

Table 4.20
DDF Patterns: Nonuniform DIF Effect-Size Estimates

	FCAT Cut Score				STD			MLR			OR		
Item	253	294	346	398	1	2	3	1	2	3	1	2	3
4	A-	A-	A-	A-	+	-	-		--	--		++	++
17	A+	A+	A+	A+	+	+	+	++	++	++	--	--	--

Note: 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 nonuniform 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. Tables 4.21 and 4.22 show the patterns, while Table 4.23 summarizes the consistency across methods in comparison to the STD. Items are grouped by DIF pattern to assist in the determination of possible patterns. Using this grouping, the consistency percent was determined and is summarized in Table 4.24. The consistency percent was calculated as the percent agreement across all three methods by DIF pattern. The highest levels of consistency were found when the divergence in nonuniform DIF effects did not occur at the FCAT level 3/4 cut score (at a cut score of 346).

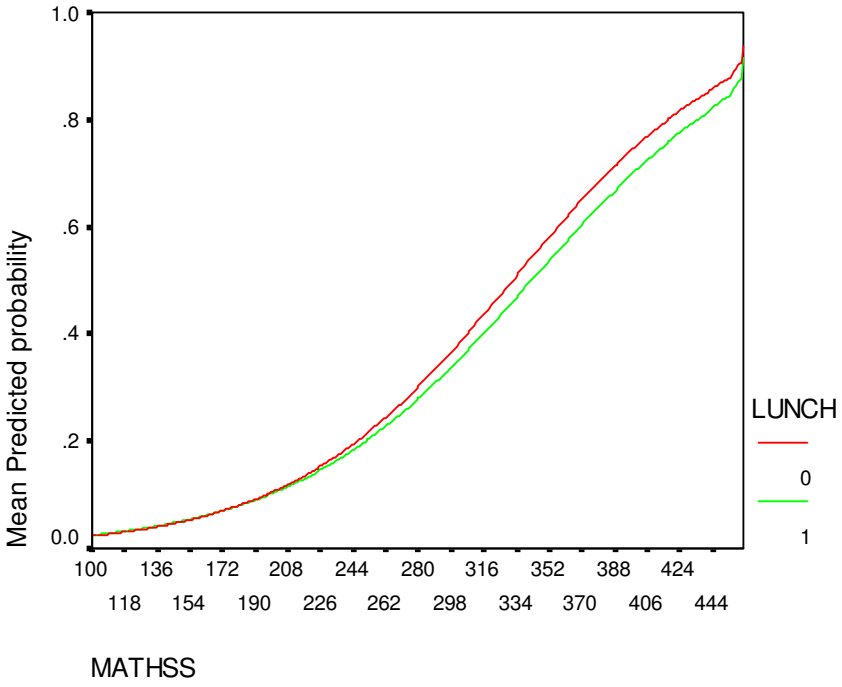


Figure 4.6: Item #4 Nonuniform DIF

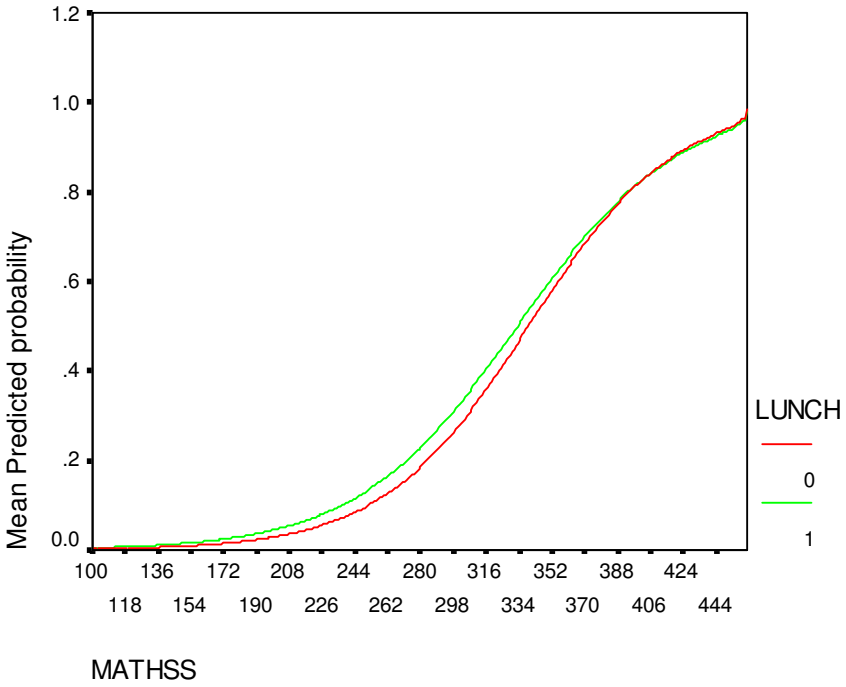


Figure 4.7: Item #17 Nonuniform DIF

Table 4.21
Items with Crossing DIF: STD and OR Patterns

	FCAT Cut Score				STD			OR		
Item	253	294	346	398	1	2	3	1	2	3
27	A+	A+	A+	A+	-	+	-		--	++
11	A+	A+	A+	A-	+	+	+	--	--	--
12	A+	A+	A+	A-	-	+	+	++	--	--
31	A+	A+	A+	A-	+	+	-	--	--	
22	B+	A+	A+	A-	+	+	+	--	--	--
25	C+	B+	A+	A-	+	-	+	--		--
7	A+	A+	A-	A-	+	+	+	--	--	--
9	A+	A+	A-	A-	+	-	-		++	++
14	A+	A+	A-	A-	-	-	-	++	++	++
15	A+	A+	A-	A-	+	-	+		++	
35	A+	A+	A-	A-	-	+	-	+		+
13	A+	A+	A-	A-	+	+	-	--	--	++
1	A+	A+	A-	B-	-	+	+	++	--	--
18	A+	A+	A-	B-	-	+	+	++		
19	A+	A+	A-	B-	+	+	-			
23	B+	A+	A-	A-	+	-	+			
10	B+	A+	A-	A-	+	+	-	--	--	--
38	B+	A+	A-	A-	+	+	+	--	--	--
24	B+	A+	A-	B-	+	-	+			--
26	B+	A+	A-	B-	-	+	+	++		--
16	B+	A+	A-	C-	+	-	-	--	++	++
39	A+	A-	A-	A-	+	-	-			
34	A+	A-	A-	A-	+	-	+	--		--
21	A+	A-	A-	A-	-	+	+	++		
5	A+	A-	A-	A-	+	-	-			
33	A+	A-	A-	A-	+	+	-		--	+
30	A+	A-	A-	B-	+	-	+	--	++	
2	A+	A-	A-	B-	+	-	-		++	++
37	A+	A-	A-	C-	-	-	-	++	++	++
8	A-	A-	A-	B-	-	+	-	++		++
29	A-	A-	B-	B-	-	-	-	++	++	++
40	A-	A-	B-	C-	-	-	-	++	++	++
6	A-	A+	A+	A+	-	-	-		++	
32	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. 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. Shading occurs when the OR pattern is consistent with the STD patterns.

Table 4.22

Items with Crossing DIF: STD and MLR Patterns

Item	FCAT Cut Score				STD			MLR		
	253	294	346	398	1	2	3	1	2	3
27	A+	A+	A+	A+	-	+	-		++	--
11	A+	A+	A+	A-	+	+	+	++	++	++
12	A+	A+	A+	A-	-	+	+	--	--	--
31	A+	A+	A+	A-	+	+	-	++	++	--
22	B+	A+	A+	A-	+	+	+	++	++	+
25	C+	B+	A+	A-	+	-	+			--
7	A+	A+	A-	A-	+	+	+		++	++
9	A+	A+	A-	A-	+	-	-		--	--
14	A+	A+	A-	A-	-	-	-	--	--	--
15	A+	A+	A-	A-	+	-	+	--	--	--
35	A+	A+	A-	A-	-	+	-	--	--	--
13	A+	A+	A-	A-	+	+	-	--	--	--
1	A+	A+	A-	B-	-	+	+	--	++	--
18	A+	A+	A-	B-	-	+	+	--		--
19	A+	A+	A-	B-	+	+	-	--		--
23	B+	A+	A-	A-	+	-	+	--	--	
10	B+	A+	A-	A-	+	+	-		++	++
38	B+	A+	A-	A-	+	+	+	++	++	++
24	B+	A+	A-	B-	+	-	+		--	++
26	B+	A+	A-	B-	-	+	+	--	--	
16	B+	A+	A-	C-	+	-	-		--	--
39	A+	A-	A-	A-	+	-	-	--		--
34	A+	A-	A-	A-	+	-	+	++		
21	A+	A-	A-	A-	-	+	+	--		++
5	A+	A-	A-	A-	+	-	-	--	++	--
33	A+	A-	A-	A-	+	+	-			--
30	A+	A-	A-	B-	+	-	+	++	--	--
2	A+	A-	A-	B-	+	-	-	--	--	--
37	A+	A-	A-	C-	-	-	-	--	--	--
8	A-	A-	A-	B-	-	+	-	--	--	--
29	A-	A-	B-	B-	-	-	-	--	--	--
40	A-	A-	B-	C-	-	-	-	--	--	--
6	A-	A+	A+	A+	-	-	-		--	
32	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 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. Shading occurs when the MLR pattern is consistent with the STD pattern.

Table 4.23

Consistency of DDF Pattern in Comparison to STD by Nonuniform DIF Pattern

	FCAT Cut Score				Consistency (1=yes, 0=no)		
Item	253	294	346	398	STD/MLR	STD/OR	STD/MLR/OR
27	A+	A+	A+	A+	1	1	1
11	A+	A+	A+	A-	1	1	1
12	A+	A+	A+	A-	0	1	0
31	A+	A+	A+	A-	1	1	1
22	B+	A+	A+	A-	1	1	1
25	C+	B+	A+	A-	0	1	0
7	A+	A+	A-	A-	1	1	1
9	A+	A+	A-	A-	1	1	1
14	A+	A+	A-	A-	1	1	1
15	A+	A+	A-	A-	0	1	0
35	A+	A+	A-	A-	0	1	0
13	A+	A+	A-	A-	0	1	0
1	A+	A+	A-	B-	0	1	0
18	A+	A+	A-	B-	0	1	0
19	A+	A+	A-	B-	0	1	0
23	B+	A+	A-	A-	0	1	0
10	B+	A+	A-	A-	0	0	0
38	B+	A+	A-	A-	1	1	1
24	B+	A+	A-	B-	1	1	1
26	B+	A+	A-	B-	0	1	0
16	B+	A+	A-	C-	1	1	1
39	A+	A-	A-	A-	0	1	0
34	A+	A-	A-	A-	1	1	1
21	A+	A-	A-	A-	1	1	1
5	A+	A-	A-	A-	0	1	0
33	A+	A-	A-	A-	1	1	1
30	A+	A-	A-	B-	0	1	0
2	A+	A-	A-	B-	0	1	0
37	A+	A-	A-	C-	1	1	1
8	A-	A-	A-	B-	0	1	0
29	A-	A-	B-	B-	1	1	1
40	A-	A-	B-	C-	1	1	1
6	A-	A+	A+	A+	1	1	1
32	A-	A+	A+	A+	1	1	1

Note: Shading occurs when the MLR and OR patterns are consistent with the STD pattern.

Table 4.24
Percent Consistent by Nonuniform 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

Note: The six observed nonuniform DIF effect patterns are summarized 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 nonuniform, crossing DIF. Figures 4.8, 4.9, and 4.10 graph the mean probability of a correct response by FCAT scale score for these items.

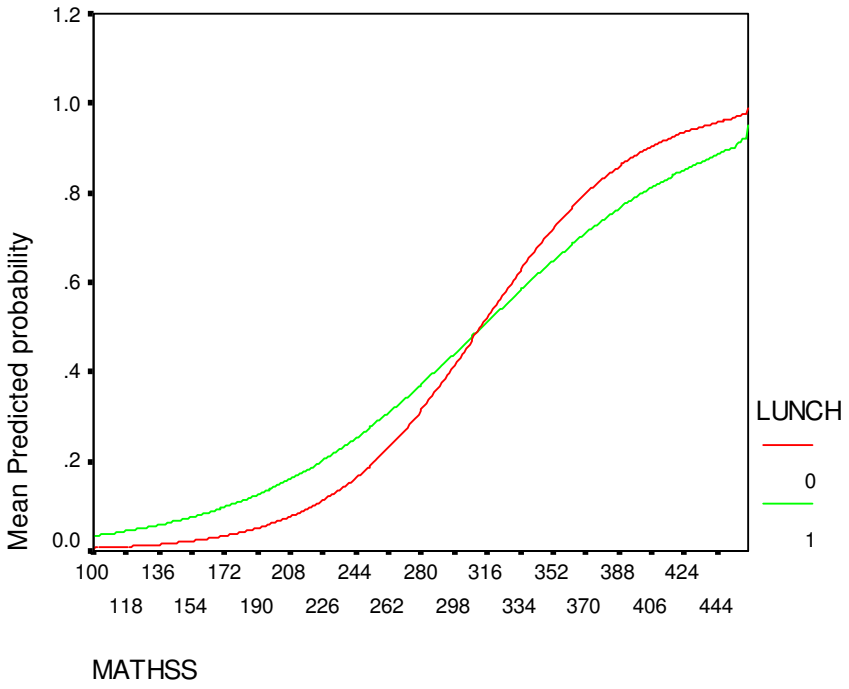


Figure 4.8: Item #16 Crossing DIF

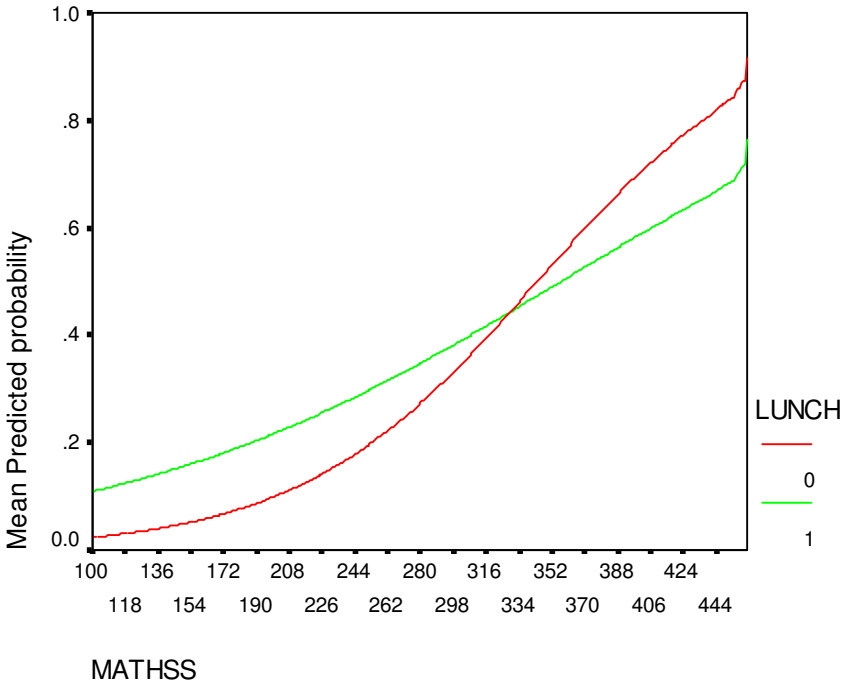


Figure 4.9: Item #24 Crossing DIF

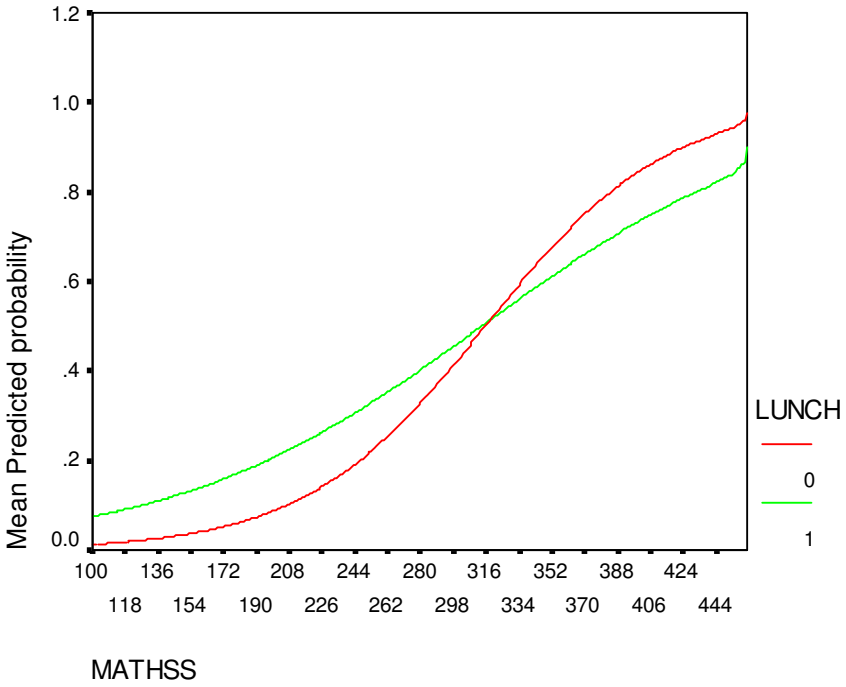


Figure 4.10: Item #26 Crossing DIF

Table 4.25
Substantial Nonuniform 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.

As can be seen in Table 4.25, the STD and OR methods are consistent in identifying divergent DDF effects when considering both significant and nonsignificant departures from 0. The results of the comparison between the MLR and OR are not as consistent and do not indicate divergent DDF in all cases (see the MLR results for item 16).

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. While these classifications primarily are applied to the item and the correct response, they also apply to the item distractors. The relationship between item characteristics and the DDF item summary statistics (i.e., DDF effect-size range, mean effect size) was explored first using scatterplots. Figure 4.11 shows the scatterplots relating item characteristics to DDF effect-size ranges. A review of the scatterplots revealed one possible influential observation. This item, item 29, was found to have moderate discrimination (a = .906), low difficulty (b= -1.173), and the highest guessing parameter (c=.562).

A Pearson correlation was calculated to determine if there was a significant relationship between item characteristics and DDF effect-size range, with and without item 29. Consistent with the scatterplots, there were no significant correlations (see Table 4.26). The reported results include item 29.

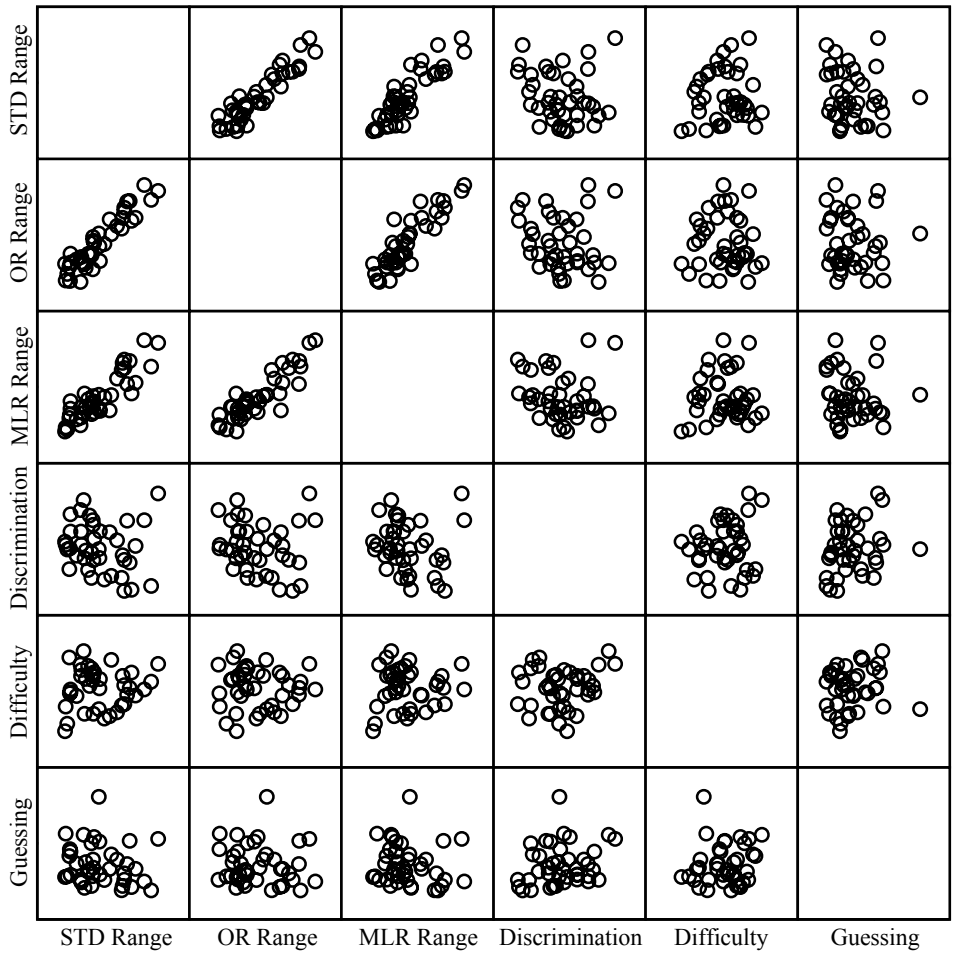


Figure 4.11: DDF Effect-Size Range by Item Characteristics

Table 4.26
Correlations Between DDF Effect-Size Range and Item Characteristics

	STD Range	OR Range	MLR Range	Discrimination	Difficulty	Guessing
STD Range	1	.929(**)	.882(**)	-.147	.089	-.133
OR Range	.929(**)	1	.888(**)	-.135	-.025	-.058
MLR Range	.882(**)	.888(**)	1	-.127	.052	-.114
Discrimination	-.147	-.135	-.127	1	.161	.225
Difficulty	.089	-.025	.052	.161	1	.107
Guessing	-.133	-.058	-.114	.225	.107	1

** Correlation is significant at the 0.01 level (2-tailed). *n*=40.

Figure 4.12 shows the scatterplots for item characteristics and DDF mean effect size. Prior to the removal of item 29, the Pearson correlation indicated a significant relationship

between the item guessing parameter and the DDF mean effect. After removing this item from the analysis, the Pearson correlation was no longer significant for this relationship. This finding was consistent across methods. See Table 4.27 for a summary of the correlations with item 29 included. Table 4.28 provides the results without item 29.

Table 4.27
Correlations Between DDF Mean Effect and Item Characteristics

	STD Mean	OR Mean	MLR Mean	Discrimination	Difficulty	Guessing
STD Mean	1	-.961(**)	.951(**)	.062	.065	-.395(*)
OR Mean	-.961(**)	1	-.967(**)	-.104	-.034	.295
MLR Mean	.951(**)	-.967(**)	1	-.020	.021	-.450(**)
Discrimination	.062	-.104	-.020	1	.161	.225
Difficulty	.065	-.034	.021	.161	1	.107
Guessing	-.395(*)	.295	-.450(**)	.225	.107	1

**Correlation is significant at the 0.01 level (2-tailed). *Correlation is significant at the 0.05 level (2-tailed). *n*=40.

Table 4.28
Correlations Between DDF Mean Effect and Item Characteristics (Without Item 29)

	STD Mean	OR Mean	MLR Mean	Discrimination	Difficulty	Guessing
STD Mean	1	-.967(**)	.946(**)	.065	-.041	-.117
OR Mean	-.967(**)	1	-.960(**)	-.109	.044	.083
MLR Mean	.946(**)	-.960(**)	1	-.030	-.066	-.269
Discrimination	.065	-.109	-.030	1	.161	.279
Difficulty	-.041	.044	-.066	.161	1	.246
Guessing	-.117	.083	-.269	.279	.246	1

**Correlation is significant at the 0.01 level (2-tailed). *n*=39.

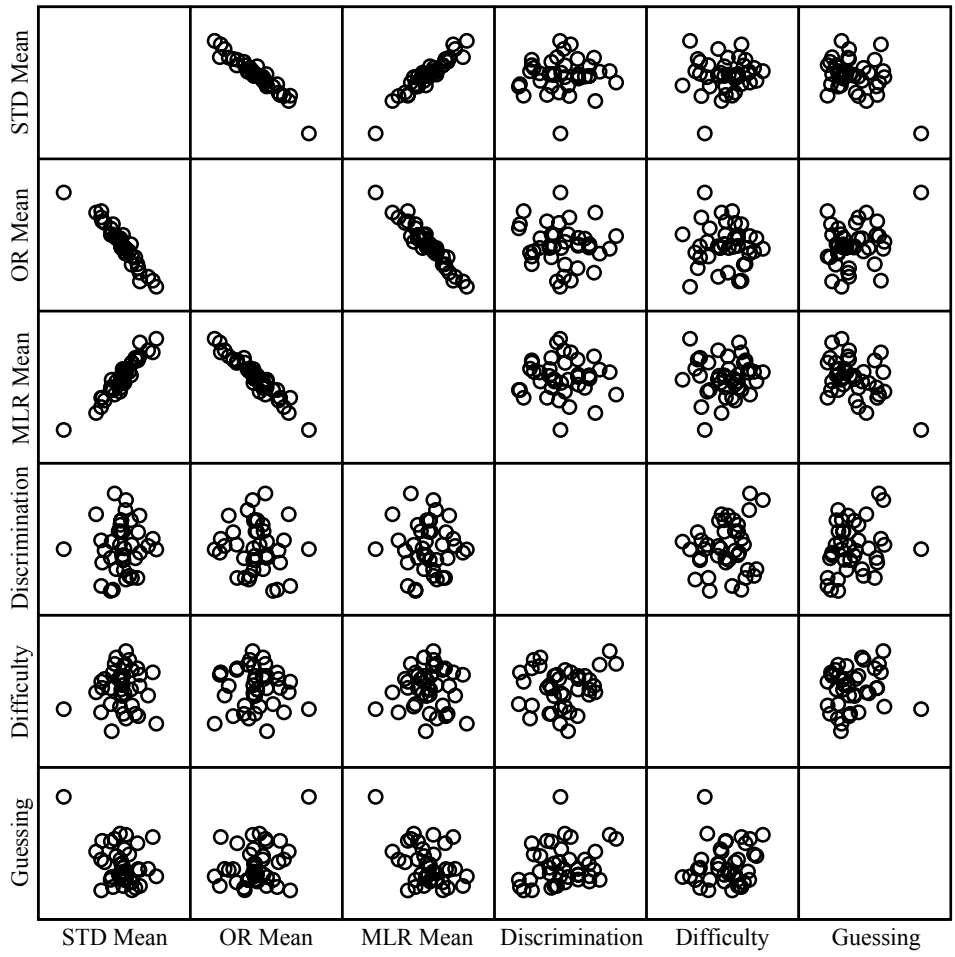


Figure 4.12: DDF Mean Effect Size by Item Characteristics

Utility of DDF Effect Information

The three items identified as having large crossing DIF were studied to explore the utility of the information provided by the DDF analyses. This exploration included discussion with a Florida Department of Education content expert to see if any plausible reasons for the DDF effects could be identified, such as the distractor representing a common misconception or common error that would make a distractor more or less attractive to a particular group. This exploration focused on the practical use of DDF results in the item revision process. The results of this exploration are summarized in a qualitative manner.

Item 16

Item 16 was classified as a moderate complexity item with a focus on measuring a student's understanding of content related to length. As noted in Figure 4.13, 59% of the

students tested selected 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 are identified in Table 4.29. As identified by the content expert, options F and G had answers that were too small and option I presented a distractor whose answer was much too large. Other than these obvious distractor characteristics, there were no plausible reasons for the identified behavior. This item had relatively high discrimination (a=1.134), moderate difficulty (b=0.078), and moderate guessing (c=.208). Given the high discrimination, and low response rates to the first two distractors, it is likely that the differential attractiveness of distractor 3 resulted from the focal group not understanding the item content.

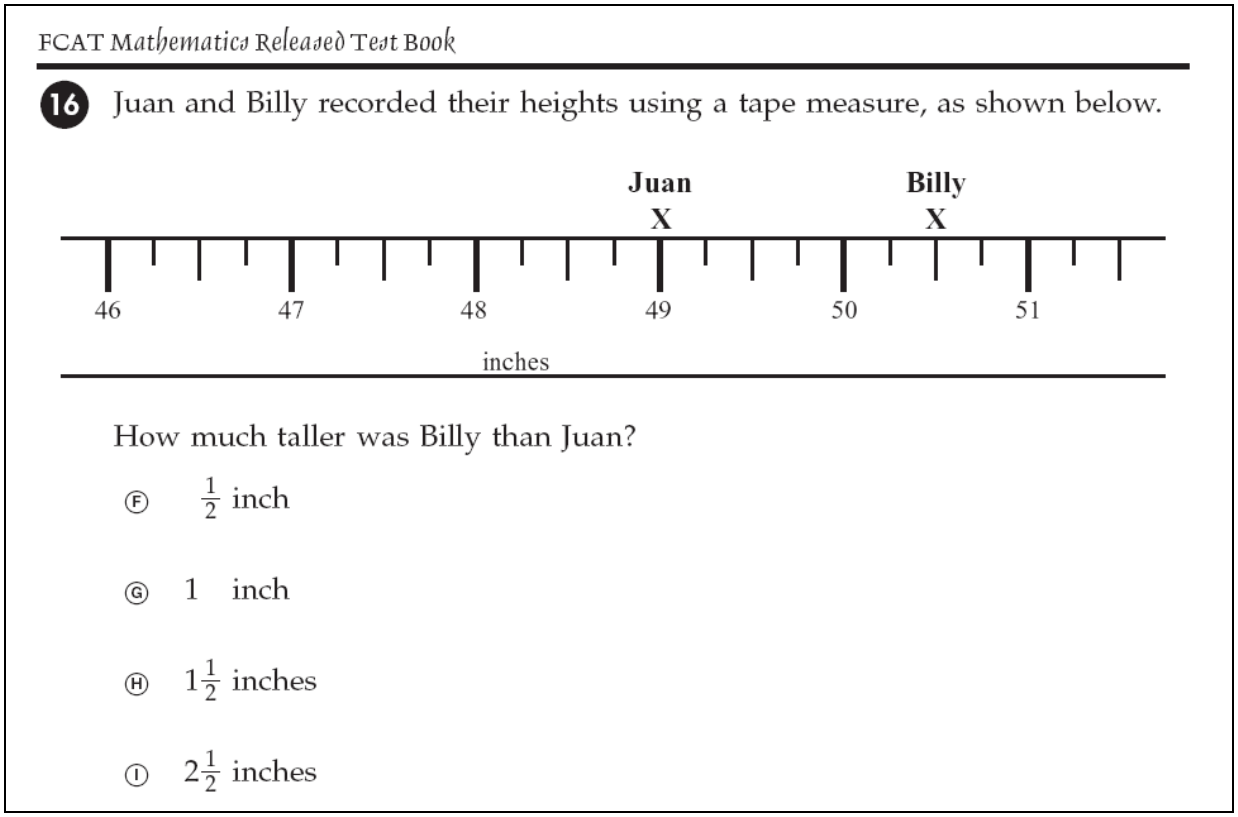


Figure 4.13: Item #16 Content (Source: Florida Department of Education, 2006)

Table 4.29
Item #16: DDF Pattern by Method

Item	MLR			STD			OR		
	1	2	3	1	2	3	1	2	3
16	-	--	--	+	-	-	--	++	++

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.

Item 24

Item 24 was classified as a moderate complexity item with a focus on measuring a student's understanding of content related to fraction size. As noted in Figure 4.14, 49% of the students tested selected the correct response.

All methods found distractor 2 (option G) less attractive to the reference group in comparison to the focal group. In addition, all three methods found distractor 3 (option H) more attractive to reference group members. The patterns are summarized in Table 4.30.

Table 4.30
Item #24: DDF Pattern by Method

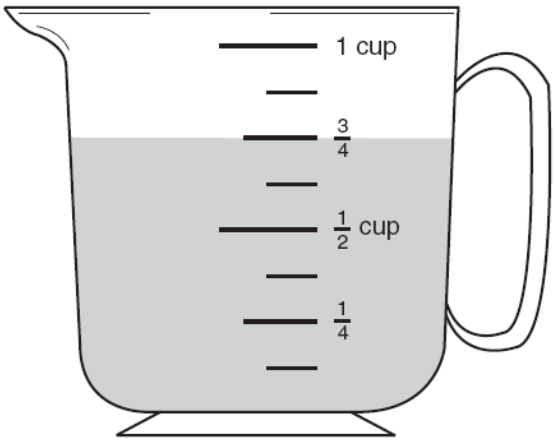
Item	MLR			STD			OR		
	1	2	3	1	2	3	1	2	3
24	-	--	++	+	-	+	-	+	--

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.

As identified by the content expert, options G and H represent common errors that reflect limited understanding of concepts related to fractions. While option F was identified as being differentially attractive to reference group members, this result was not significant for either the MLR or OR methods. Any differences found with regard to distractor 1 (option F) are likely due to not understanding the item content or item characteristics, as identified by item difficulty, discrimination, and guessing. There were no other plausible reasons for these differences. This

item had the second highest discrimination (a=1.367), the highest difficulty (b=1.011), and third highest guessing parameter (c=.348).

24 Jackie used $\frac{3}{4}$ cup of sugar to make cookies.



Which of the following is greater than $\frac{3}{4}$?

Ⓕ $\frac{1}{8}$ cup

Ⓖ $\frac{3}{8}$ cup

Ⓗ $\frac{5}{8}$ cup

Ⓘ $\frac{7}{8}$ cup

Figure 4.14: Item #24 Content (Source: Florida Department of Education, 2006)

Item 26

Item 26 was classified as a low complexity item with a focus on measuring a student's understanding of content related to the likelihood of an outcome. According to the item statistics, 55% of the students tested selected the correct response.

All methods found distractor 1 (option F) less attractive to the reference group in comparison to the focal group. In addition, while not significant across all methods, all three methods found distractor 3 (option I) more attractive to reference group members. The patterns are identified in Table 4.31. This item had the highest discrimination (a=1.429), relatively high difficulty (b=.525), and high guessing (c=.325).

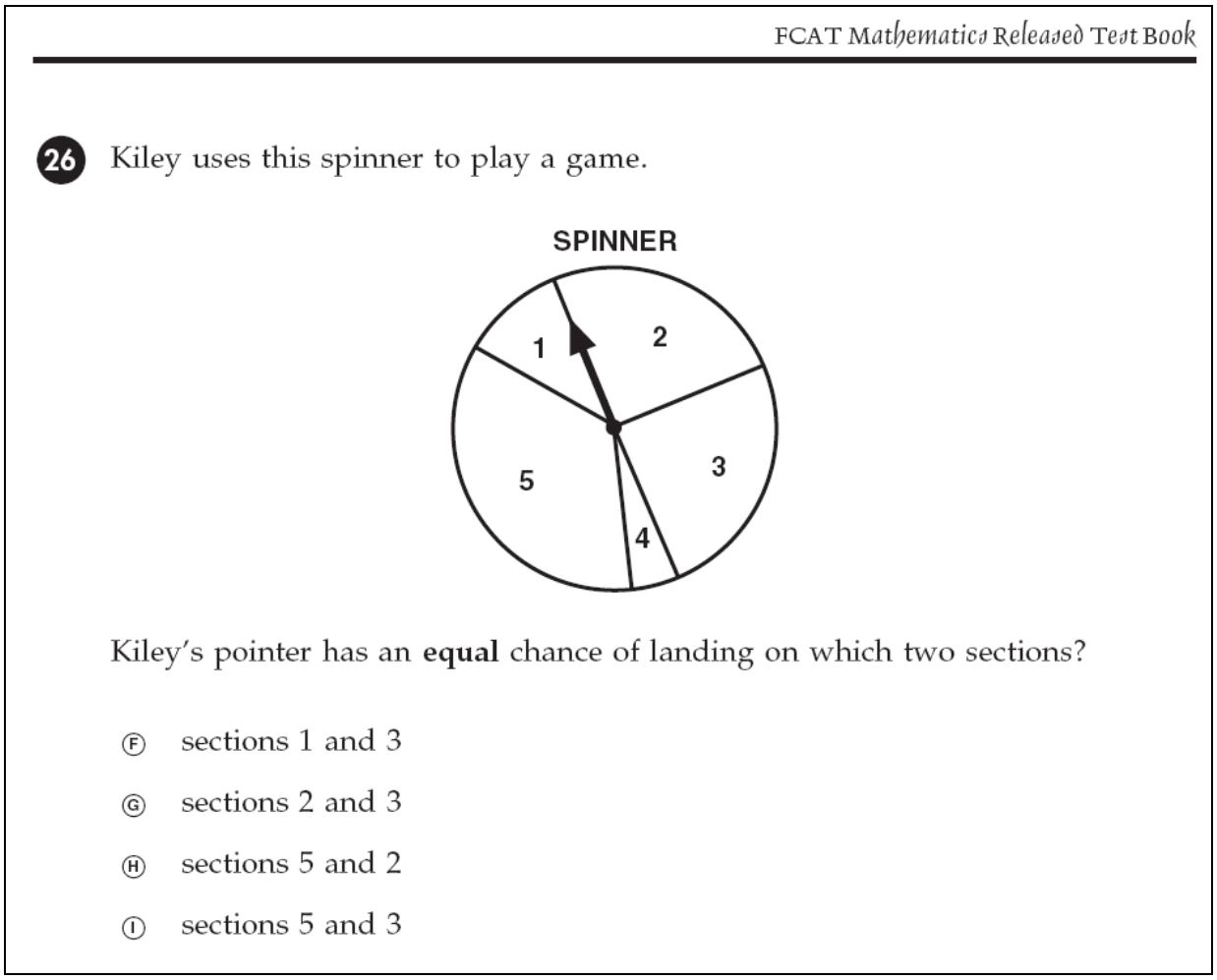


Figure 4.15: Item #26 (Source: Florida Department of Education, 2006)

The content expert stated that the distractors were meant to provide a variety of combinations representing unequal sections. Students who understood the content measured by this item, and successfully understood the question, would have selected the correct response. Students differentially attracted to options H and I may have misread the question and focused

on identifying the sections with the highest likelihood of being selected, although this explanation is not complete because the chance of choosing sections 2 and 3 are equal. Students differentially attracted to option I may not have had experience with spinners or content knowledge in this area and, as result, simply guessed.

Table 4.31
Item #26: DDF Pattern by Method

Item	MLR			STD			OR		
	1	2	3	1	2	3	1	2	3
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.

Summary

No clear patterns emerged with these items to definitively 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. Presently, content experts are only provided with DIF results and these results do not provide indications of which distractors are problematic. The items reviewed in this section were operational items that had already been subjected to many qualitative reviews prior to serving as core items.

CHAPTER 5

DISCUSSION

Summary

This research study examined the consistency of DDF estimates from the standardized distractor analysis (STD), odds ratio (OR), and multinomial logistic regression (MLR) approaches. As the nation continues to place high stakes on results of statewide assessment programs, understanding the factors that contribute to DIF becomes increasingly important. DDF analyses have resurfaced in the literature as one way to understand these factors. Given this, it was determined that a study comparing these methods would be of value to statewide assessment programs.

The study was designed to provide insight into whether the magnitude and pattern of DDF effects were found to be consistent across all methods and whether the pattern of DDF effects supported the DIF findings.

The results of the three methods were summarized by item and distractor, using several approaches. 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, a Pearson correlation coefficient was used. The STD approach was treated as the standard for comparison, because it is a recognized 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 summarized for use 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, 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. Correlation indices were used to summarize the consistency of the data.

As expected, all methods yielded very similar results. The STD and OR methods for detecting DDF are based on the same data table and were found to have very highly related results. Key differences are that STD compares the probability of success while the OR method compares the odds of success, and the two methods use different weighting functions. Despite these differences that result in slight shifts in the ordering of items, the correlation between the

DDF effect-size estimates reached almost negative unity (-.922). Likewise, the correlation between the STD and MLR methods was very high (.907).

The consistency between STD and the other two approaches was examined through summarizing the mean effect size, the effect-size range, and the pattern of effects. All of the methods were found to have high correlations, indicating a strong linear relationship in the effect-size means. The range of each item's DDF effect-size estimates was found to be consistent among the three methods, with the OR and STD methods having the strongest linear relationship. With regard to divergent distractor-level effects, both the OR and MLR methods were found to have a statistically significant relationship with the STD approach.

Under the NRM, Penfield (2009) found that uniform DIF can occur when the DDF effect is constant across all distractors. In addition, he found that crossing DIF effects can only occur in the presence of DDF effects that vary in sign. He proposed that these findings can "help target the particular item property responsible for the DIF effect" (Penfield, 2009, p. 23). For items with significant uniform DIF, the item stem or correct option may be the source of the DIF effect. For items with significant nonuniform DIF, the distractors are likely the source of the DIF effect.

Using the logistic regression DIF results, items that had only significant uniform DIF effects were investigated for a constant DIF effect across all distractors based on the magnitude of the range. Four items were found to only have significant uniform DIF effects. None of these items exhibited equal DDF effects across the item's distractors, as evidenced by the non-zero range estimates. To partially compensate for error in the estimates, 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 constant effect across all distractors. No patterns were found to indicate that smaller ranges were indicative of only uniform DIF

All items identified as having significant nonuniform DIF in the logistic regression analysis were studied for 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 nonuniform DIF with no crossing DIF at those scale points. The DDF pattern was consistent across all three methods for these items.

Thirty-four items were found to have significant nonuniform DIF that crossed at some point on the ability scale. For all patterns of DIF under this condition, the STD and OR results

were consistent. The STD and MLR results were much less consistent. Three items exhibited large nonuniform, crossing DIF. For these items, the STD and OR methods were consistent in identifying divergent DDF effects when considering both significant and nonsignificant departures from 0. The results of the comparison between the MLR and OR were not as consistent and did not indicate divergent DDF in all cases.

Both scatterplots and Pearson correlations were used to investigate possible relationships between DDF item summary statistics (i.e., DDF effect-size range, mean effect size) and item discrimination, difficulty, and guessing. Consistent with the scatterplots, there were no significant correlations between item characteristics and DDF item summary statistics.

The DDF effects for the three items with large crossing DIF were qualitatively reviewed to determine the contributions that DDF analyses could make to a statewide assessment program. Based on this limited review, it appears that the contributions would be greatest during field testing. The items that were reviewed in this study were operational items that had already passed many qualitative reviews and the additional information provided by the DDF analysis did not lead to any definitive conclusions.

DDF Methodology Choice and Use by Practitioners

The results of the STD, OR, and MLR methods were found to be very highly related, based on correlation indices. Because of the strong relationship of the OR method to the accepted STD method, practitioners may prefer to conduct DDF analyses using the OR method. In comparing the OR and STD methods, OR appears to be the preferred approach because the OR is easily calculated and offers a test of the significance of the DDF effect. After a simple recoding of the data, the OR can be run with common software programs like SPSS that are capable of calculating the Mantel-Haenszel statistic. While conceptually simple to calculate, the STD method is not as easy to implement. For those not adept at writing computer code, the STD requires first calculating frequencies at each score level using a cross tabulation procedure and then using those frequencies in the calculation of the index. A combination of SAS and Excel were used for this purpose in this study. In addition, the STD does not offer a test of the significance of the DDF effect. The MLR approach also is available through SPSS and other similar programs that are capable of estimating multinomial regression models; however, the analyst must be concerned with model improvement and the interpretation of the results is more

complex. Further analysis is needed to determine the feasibility of using the MLR approach with the addition of an interaction term. This is discussed later as a possible future direction.

For practitioners, the value of DDF analyses appears to be early in the item development and revision process as a tool for item writers. The information provided can help target the particular distractor that may be differentially attractive to a specific group. This type of distractor-level information is valuable information to individuals reviewing items for possible future use on statewide assessments.

Limitations

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 programs that differ in the scoring model used and the steps used to screen items for placement on the statewide 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 maximize 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.

Future Directions

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 nonuniform, 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.

Conclusion

Comparisons of three methods for detecting DDF were made 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 the SPSS software package used in this study, unlike the STD method which must be programmed. Despite these and the other discussed differences, all three methods present a viable option for use in improving test items included in statewide assessment programs.

APPENDIX

HUMAN SUBJECTS APPROVAL MEMORANDUM

Office of the Vice President For Research
Human Subjects Committee
Tallahassee, Florida 32306-2742
(850) 644-8673 · FAX (850) 644-4392

APPROVAL MEMORANDUM

Date: 2/12/2010

To: Sharon Koon

Address: 9146 Old Chemonie Road, Tallahasse, Florida 32309
Dept.: EDUCATIONAL PSYCHOLOGY AND LEARNING SYSTEMS

From: Thomas L. Jacobson, Chair

Re: Use of Human Subjects in Research
A Comparison of Methods for Detecting Differential Distractor Functioning

The application that you submitted to this office in regard to the use of human subjects in the research proposal referenced above has been reviewed by the Human Subjects Committee at its meeting on 02/10/2010. Your project was approved by the Committee.

The Human Subjects Committee has not evaluated your proposal for scientific merit, except to weigh the risk to the human participants and the aspects of the proposal related to potential risk and benefit. This approval does not replace any departmental or other approvals, which may be required.

If you submitted a proposed consent form with your application, the approved stamped consent form is attached to this approval notice. Only the stamped version of the consent form may be used in recruiting research subjects.

If the project has not been completed by 2/9/2011 you must request a renewal of approval for continuation of the project. As a courtesy, a renewal notice will be sent to you prior to your expiration date; however, it is your responsibility as the Principal Investigator to timely request renewal of your approval from the Committee.

You are advised that any change in protocol for this project must be reviewed and approved by the Committee prior to implementation of the proposed change in the protocol. A protocol change/amendment form is required to be submitted for approval by the Committee. In addition, federal regulations require that the Principal Investigator promptly report, in writing any unanticipated problems or adverse events involving risks to research subjects or others.

By copy of this memorandum, the Chair of your department and/or your major professor is reminded that he/she is responsible for being informed concerning research projects involving human subjects in the department, and should review protocols as often as needed to insure that

the project is being conducted in compliance with our institution and with DHHS regulations.

This institution has an Assurance on file with the Office for Human Research Protection. The Assurance Number is IRB00000446.

Cc: Betsy Becker, Advisor
HSC No. 2010.3877

LIST OF REFERENCES

Abedi, J., Leon, S., & Kao, J. C. (2008). *Examining differential distractor functioning in reading assessments for students with disabilities* (CRESST Tech. Rep. No. 743). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Bock, R. D. (1972). Estimating item parameters and latent proficiency when the responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (pp. 234-240). Westport, CT: American Council on Education and Praeger Publishers.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education*, *2*, 217-233.

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 Associates.

Dorans, N. J., & 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). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355-368.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1988). *The standardization approach to assessing differential speededness* (Research Rep. No. 88-31). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement*, *29*, 309-319.

Florida Department of Education. (2005). *FCAT mathematics test item specifications, grades 3-5*. Tallahassee, FL: Author.

Florida Department of Education. (2006). *FCAT 2006 mathematics released test, grade 3.* Tallahassee, FL: Author.

Florida Department of Education. (2007). *FCAT reading and mathematics: Technical report for the 2006 FCAT test administrations*. Tallahassee, FL: Author. Retrieved from http://fcat.fldoe.org/pdf/releasepdf/06/FL06_Rel_G3M_AK_Cwf001.pdf

Florida Department of Education. (2008). *FCAT test design summary*. Tallahassee, FL: Author.

Florida Department of Education. (2009). *Frequently asked questions*. Retrieved from http://www.fldoe.org/faq/default.asp?ALL=Y&Dept=202

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement*, *26*, 147-160.

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

Kamata, A., & Vaughn, B. K. (2004). An introduction to differential item functioning analysis. *Learning Disabilities: A Contemporary Journal*, *2*, 49-69.

Kamata, A., & Williams, R. (2006). *Modeling a Differential Distractor Functioning (DDF) with Multinomial Regression*. Proposal submitted to the American Educational Research Association.

Kato, K., Moen, R. E., & 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*, *28*, 28-40.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 710-748.

Mapuranga, R., Dorans, N. J., & Middleton, K. (2008, August). *A review of recent developments in differential item functioning* (ETS Research Rep. No. RR-08-43). Princeton, NJ: Educational Testing Service.

Middleton, K., & Laitusis, C. C. (2007). *Examining test items for differential distractor functioning among students with learning disabilities* (ETS Research Rep. No. RR-07-43). Princeton, NJ: Educational Testing Service.

Monahan, P. O., McHorney, C. A., Stump T. E., & 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*, *32*, 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*, *45*, 247-269.

Penfield, R. D. (in-press). Modeling DIF effects using distractor-level invariance effects: Implications for understanding the causes of DIF. *Applied Psychological Measurement*.

Petersen, N. S. (1987). DIF procedures for use in statistical analyses. *Unpublished memorandum of September 25, 1987*. Princeton, NJ: Educational Testing Service.

Rivera, C., & Bleistein, C. A. (1988). *A comparison of Hispanic and white non-Hispanic students' omit patterns on the Scholastic Aptitude Test* (ETS Research Rep. No. RR-88-44). Princeton, NJ: Educational Testing Service.

Samejima, F. (1979). *A new family of models for the multiple choice item* (Research Rep. No. 79-4). Knoxville, TN: University of Tennessee.

Schmitt, A. P., & 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). Princeton, NJ: Educational Testing Service.

Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement*, *27*, 67-81.

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

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement*, *26*, 161-176.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*, 501-519.

Thissen, D., Steinberg, L., & Wainer, H. (1992). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, W.-C. (2000a). Factorial modeling of differential distractor functioning in multiple-choice items. *Journal of Applied Measurement*, *I*, 238-256.

Wang, W.-C. (2000b). The simultaneous factorial analysis of differential item functioning. *Methods of Psychological Research Online*, *5*, 57-76.

Wright, D.J. (1986). *An empirical comparison of the Mantel-Haenszel and standardization methods of detecting differential item performance*. A paper presented at the annual meeting of the National Council on Measurement in Education in San Francisco, April 1986.

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

BIOGRAPHICAL SKETCH

Sharon Koon was born in October 1966 in New Jersey, but raised in Michigan and later Florida. She earned her Bachelor's degree in Secondary Science and Mathematics Teaching in 1988 and her Master's Degree in Secondary Science Education in 1989, with both of these degrees earned at Florida State University. In 1989, Sharon joined the Florida Department of Education as a Research Associate in the Office of Policy Research and Improvement, focused on initiatives designed to improve mathematics and science education in Florida. Aside from a brief period of time spent teaching, Sharon has continued to work at the Florida Department of Education in areas related to curriculum, instruction, and assessment. Currently, she is a Policy, Research, and Accountability Coordinator in the Office of Assessment. In this role, she assists in the development and implementation of Florida's statewide assessment programs for both students and teachers.