2009-12-18

A Comparison of Adjacent Categories and Cumulative DSF Effect Estimators

Karina Alvarez Gattamorta
University of Miami, kgattamorta@miami.edu

UNIVERSITY OF MIAMI

A COMPARISON OF ADJACENT CATEGORIES AND CUMULATIVE DSF EFFECT ESTIMATORS

By

Karina A. Gattamorta

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida

December 2009

UNIVERSITY OF MIAMI

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

A COMPARISON OF ADJACENT CATEGORIES AND CUMULATIVE DSF EFFECT ESTIMATORS

Karina A. Gattamorta

Approved:

_____ Randall D. Penfield, Ph.D. Associate Professor of Educational and Psychological Studies	_____ Terri A. Scandura, Ph.D. Dean of the Graduate School
_____ Nicholas D. Myers, Ph.D. Assistant Professor of Educational and Psychological Studies	_____ Kent Burnett, Ph.D. Associate Professor of Educational and Psychological Studies
_____ Batya Elbam, Ph.D. Associate Professor of Teaching and Learning	

GATTAMORTA, KARINA A. (Ph.D., Educational and Psychological Studies)
A Comparison of Adjacent Categories and Cumulative DSF Effect Estimators. (December 2009)

Abstract of a dissertation at the University of Miami.

Dissertation supervised by Professor Randall D. Penfield
No. of pages in text. (69)

The study of measurement invariance in polytomous items that targets individual score levels is known as differential step functioning (DSF; Penfield, 2007, 2008). DSF methods provide specific information describing the manifestation of the invariance effect within particular score levels and therefore serve a diagnostic role in identifying the individual score levels involved in the item's invariance effect.

The analysis of DSF requires the creation of a set of dichotomizations of the item response variable. There are two primary approaches for creating the set of dichotomizations to conduct a DSF analysis. The first approach, known as the adjacent categories approach, is consistent with the dichotomization scheme underlying the generalized partial credit model (GPCM; Muraki, 1992) and considers each pair of adjacent score levels while treating the other score levels as missing. The second approach, known as the cumulative approach, is consistent with the dichotomization scheme underlying the graded response model (GRM; Samejima, 1997) and includes data from every score level in each dichotomization. To date, there is limited research on how the cumulative and adjacent categories approaches compare within the context of DSF, particularly as applied to a real data set. The understanding of how the interpretation and practical outcomes may vary given these two approaches is also limited. The current study addressed these two issues.

This study evaluated the results of a DSF analysis using both the adjacent categories and cumulative dichotomization schemes in order to determine if the two approaches yield similar results and interpretations of DSF. These approaches were applied to data from a polytomously scored alternate assessment administered to children with significant cognitive disabilities. The results of the DSF analyses revealed that the two approaches generally led to consistent results, particularly in the case where DSF effects were negligible. For steps where significant DSF was present, the two approaches generally guide analysts to the same location of the item. However, several aspects of the results rose questions about the use of the adjacent categories dichotomization scheme. First, there seemed to be a lack of independence of the adjacent categories method since large DSF effects at one step are often paired with large DSF effects in the opposite direction found in the previous step. Additionally, when a substantial DSF effect existed, it was more likely to be significant using the cumulative approach over the adjacent categories approach. This is likely due to the smaller standard errors that lead to greater stability of the cumulative approach. In sum, the results indicate that the cumulative approach is preferable over the adjacent categories approach when conducting a DSF analysis.

To Francesca: whether in utero or sleeping soundly just a few feet away, you have been with me while I have written every word of this dissertation. You have been my motivation and inspiration.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

Measurement invariance is defined as the independence of group membership and item response after conditioning on ability (Millsap & Meredith, 1992). A particular violation of invariance known as differential item functioning (DIF) exists when individuals with the same level of ability but belonging to different groups have different chances of success on an item (Camilli & Shepard, 1994; Penfield & Camilli, 2007). The presence of DIF poses a threat to the validity of scores because it may imply that an item is biased for a subpopulation of examinees. As such, the investigation of DIF is an important component of test fairness and the validation process (AERA/APA/NCME, 1999; Camilli, 2006). While methods for evaluating DIF in dichotomous items have been well documented over the past two decades (Camilli & Shepard, 1994; Holland & Thayer, 1988; Lord, 1980; Penfield & Camilli, 2007), the use of performance-based assessment has led to a growing area of interest in DIF in polytomous items.

The assessment of DIF in polytomous items is conceptually more complex than DIF detection in dichotomous items. When considering invariance in polytomous items, the form of invariance can change across score levels. For example, it is possible for one score level to exhibit a lack of invariance, but not others. Additionally, within a single item, a lack of invariance can favor the reference group (i.e., the group expected to be at an advantage) at one score level and the focal group (i.e., the group expected to be at a disadvantage) in another score level (Gattamorta, 2008; Penfield, Alvarez, & Lee, 2009).

Despite the complexity of invariance in polytomous items, most widely used DIF evaluation approaches are based on item-level investigations of invariance (Penfield & Lam, 2000). These item-level approaches provide the analyst with a single index of DIF effect as they measure the overall invariance effect aggregated across all score levels.

Examples of this approach include the standardized mean difference (Dorans & Schmitt, 1991), Mantel's chi-square statistic (Mantel, 1963), the generalized Mantel-Haenszel statistic (Somes, 1986), polytomous SIBTEST (Chang, Mazzeo, & Roussos, 1996), logistic discriminant function analysis (Miller & Spray, 1993), the cumulative common log-odds ratio estimator (Penfield & Algina, 2003), and Cox's B (Camilli & Congdon, 1999; Cox, 1958). Because these approaches provide a single item-level measure of DIF, they do not inform the analyst as to which score levels are manifesting the DIF effect. As a result, they provide limited information to help guide item revision.

Recently, researchers have proposed addressing invariance at each individual score level rather than aggregating across score levels. This approach has the advantage of informing which score levels are responsible for the lack of invariance. The study of invariance in polytomous items that targets individual score levels is known as differential step functioning (DSF; Penfield, 2007, 2008; Penfield, Gattamorta, & Childs, 2009). The evaluation of DSF begins by decomposing each polytomous item into $J = r - 1$ step functions where r is the number of score levels in a given item. Each step function describes the probability of advancing or "stepping" from each score level to a successively higher score level. DSF exists when there is a between-group difference in one or more of the item's step functions. Analysis of DSF has been undertaken using a variety of approaches including IRT-based approaches under the graded response model (Cohen, Kim, & Baker, 1993) and the partial credit model (Penfield, Myers, & Wolfe, 2008), logistic regression (French & Miller, 1996), and using a common odds ratio approach (Penfield, 2007).

All DSF approaches look at between-group differences in step functions. The step function can be defined in several ways (Mellenbergh, 1995; Penfield, 2008). Two ways to define the step functions used in the context of DSF analyses are the adjacent categories approach and the cumulative approach. The adjacent categories approach defines each of the *J* step functions in a manner consistent with the generalized partial credit model (GPCM; Muraki, 1992). Under the GPCM, the *j*th step function specifies the probability of successfully advancing from score level $j-1$ to score level j. The cumulative approach defines the step functions in a manner consistent with the graded response model (GRM; Samejima, 1997). Under the GRM, the *j*th step function specifies the probability of successfully advancing from $0, 1, \ldots, j$ to $j+1, \ldots, J$. It is important to recognize that because these two models define the step function differently, there may be substantial differences in the results obtained from a DSF analysis conducted using one approach over the other. In other words, the resulting interpretation of the step-level parameters may ultimately depend on the approach used to define the step function. Additional information about the adjacent categories and cumulative approaches is provided in the next section.

Although DSF research has used both the cumulative (Penfield, 2008; Penfield, Alvarez, et al., 2009; Penfield, Gattamorta, et al., 2009) and adjacent categories (Alvarez & Penfield, 2007; Penfield, et al., 2008) approaches, little research has compared the two. Penfield (2008) conducted a simulation study comparing the statistical properties of the odds ratio DSF effect estimators under the adjacent categories and cumulative approaches. Previous research comparing the GRM and GPCM has also been conducted within the context of vertical scaling (Bishop & Omar, 2002), ability estimation in

computerized adaptive testing (Wang & Wang, 2002) and a person fit test for IRT models (Glas & Dagohoy, 2007). Nevertheless, there is limited research on how these two approaches compare within the context of DSF as applied to a real data set and limited understanding of how the interpretation and practical outcomes may vary given these two approaches. The adjacent categories and cumulative approaches define step functions differently and thus may give different results and interpretations of invariance. Also, the two approaches may have different statistical properties. An understanding of how these two approaches compare is critical to researchers and practitioners that use DSF as part of the test validation process.

Given the importance of whether the two approaches yield consistent results, this study aims to determine the extent to which the cumulative and adjacent categories approaches to defining the step function lead to similar results and interpretation of invariance when applied to a real educational data set. To answer this question, a DSF analysis comparing the performance of English language learners (ELLs) to non-ELLs on a performance-based assessment was conducted and the results using both the adjacent categories and cumulative dichotomization schemes were evaluated in order to determine if the two approaches yield similar results and interpretations of DSF.

Chapter 2: Two Conceptions of DSF

Let us define the response to the polytomous item by Y, where Y can assume values from 0, 1, … , J. As previously discussed, step functions are defined differently under the adjacent categories and cumulative approaches. In the adjacent categories approach, the jth step function specifies the probability of successfully advancing from $Y = j - 1$ to $Y = j$. A polytomous item with four score levels (i.e., $j = 0, 1, 2,$ and 3) would consist of three steps: step 1 would be defined by successfully advancing from a 0 to a 1, step 2 would be defined by successfully advancing from a 1 to a 2, and step 3 would be defined by successfully advancing from a 2 to a 3. Under the cumulative approach, the jth step function specifies the probability of successfully advancing from $Y < j$ to $Y \geq j$. Following the example of an item with four score levels, step 1 would be defined by successfully advancing from a 0 to a 1, 2, or 3; step 2 would be defined by successfully advancing from a 0 or a 1 to a 2 or a 3; and step 3 would be defined by successfully advancing from a 0, 1, or 2 to a 3.

It may be the case that despite their differences, the cumulative and adjacent categories approaches will yield relatively similar results (i.e., identify a similar DSF pattern of classification categories across steps), particularly under the condition of constant DSF. This may occur due to some overlap in the way steps are defined under the two approaches. It is possible, however, that substantial differences between the two approaches may exist, particularly under the condition of DSF that is isolated to a single score level. Consider, for example, an item where DSF exists only at a single step and a single score level, j, is responsible for the DSF effect. In this case, under the adjacent categories dichotomization scheme, only the step from $j - 1$ to j will be implicated. Under

the cumulative dichotomization scheme, the score level responsible for DSF is implicated in all steps and as such, estimates of DSF in all steps can be impacted. In the condition where only a single score level is responsible for the DSF effect, the adjacent categories dichotomization scheme is at an advantage since it is likely to be more precise in its detection rate and interpretation of an invariance effect since it will not be clouded by information from other score levels. A difference between the two approaches is not only expected with regards to the interpretation of invariance, but also with regards to power. It is expected that step functions defined using the cumulative approach will result in more powerful results due to the larger sample size compared to those step functions defined using the adjacent categories approach.

Although several approaches for evaluating DSF have been proposed, this paper focuses on Penfield's (2007, 2008) odds ratio approach which compares the odds of successfully advancing at the *j*th step for reference and focal group members with the same observed score. Consider a test with possible score levels (e.g., raw, summated score) denoted by $k = 1, 2, \ldots, S$, where the total score serves as an appropriate proxy for ability. A ratio of the odds of success at the *j*th step for the reference group over the odds of success for the focal group is estimated using:

$$\widehat{\alpha}_j = \frac{\sum_{k=1}^{S} A_{jk} D_{jk} / T_{jk}}{\sum_{k=1}^{S} B_{jk} C_{jk} / T_{jk}}, \quad (1)$$

where A_{jk} and B_{jk} represent the number of reference group members that were successful and unsuccessful at the *j*th step, C_{jk} and D_{jk} represent the number of focal group members that were successful and unsuccessful at the *j*th step, and T_{jk} represents the sum of A_{jk}, B_{jk}, C_{jk}, and D_{jk}. Note, this is equivalent to the Mantel-Haenszel odds ratio for dichotomous items, such that each step is treated as a dichotomy. Because a polytomous

item is dichotomized differently under the cumulative and adjacent categories approaches, A_{jk}, B_{jk}, C_{jk}, D_{jk}, and T_{jk} differ across the two approaches. To illustrate these differences, consider a polytomous item with four score levels, 0, 1, 2, and 3. Under the adjacent categories approach, A_{jk} for step 1 represents reference group members that earned a score of 1, B_{jk} represents reference group members that earned a score of 0, C_{jk} represents focal group members that earned a score of 1, D_{jk} represents focal group members that earned a score of 0, and T_{jk} represents the total number of examinees that earned a score of 0 or 1. Under this approach, examinees that earned scores of 2 or 3 would not be considered in this step. Under the cumulative approach, A_{jk} for step 1 represents reference group members that earned a score of 1, 2, or 3, B_{jk} represents reference group members that earned a score of 0, C_{jk} represents focal group members that earned a score of 1, 2, or 3, D_{jk} represents focal group members that earned a score of 0, and T_{jk} represents the total number of examinees.

The natural logarithm of $\widehat{\alpha}_j$ for the jth step is denoted by $\hat{\lambda}_j$, where positive values of $\hat{\lambda}_j$ indicate DSF favoring the reference group, negative values of $\hat{\lambda}_j$ indicate DSF favoring the focal group, and a value of zero is indicative of no DSF. The resulting estimator is consistent in scale and direction with the Mantel-Haenszel common log-odds ratio applied to dichotomous items, and as such can be interpreted in a manner consistent with the ETS classification scheme (Zieky, 1993) whereby $\left|\hat{\lambda}_j\right| < 0.43$ corresponds to a small DSF effect, $0.43 \leq \left|\hat{\lambda}_j\right| < 0.64$ corresponds to a moderate DSF effect, and $\left|\hat{\lambda}_j\right| \geq 0.64$ corresponds to a large DSF effect (Penfield, 2007; Penfield, Alvarez, et al., 2009). For the remainder of this paper, the log-odds ratio estimated using the adjacent categories dichotomization scheme will be referred to as *AC-LOR* and the log-odds ratio

estimated using the cumulative dichotomization scheme will be referred to as *CU-LOR*. An estimator of the standard error of $\hat{\lambda}_j$ can be obtained using:

$$SE(\hat{\lambda}_j) = \sqrt{\frac{\sum_{k=1}^{S} T_k^{-2}(A_{jk}D_{jk}+\hat{\alpha}_j B_{jk}C_{jk})(A_{jk}+D_{jk}+\hat{\alpha}_j B_{jk}+\hat{\alpha}_j C_{jk})}{2\left(\sum_{k=1}^{S}\frac{A_{jk}D_{jk}}{T_k}\right)^2}}. \quad (2)$$

This standard error is essentially the same as the standard error used for the dichotomous Mantel-Haenszel log-odds ratio (Camilli & Shepard, 1994; Penfield & Camilli, 2007). Dividing $\hat{\lambda}_j$ by its estimated standard error yields the test statistic

$$z(\hat{\lambda}_j) = \frac{\hat{\lambda}_j}{SE(\hat{\lambda}_j)}. \quad (3)$$

Under the null hypothesis of no DSF, this test statistic is distributed approximately as standard normal.

Chapter 3: Method

Data from performance-based Grade 5 Science and Grade 8 Science assessments administered to students with significant cognitive disabilities were analyzed in this study. Each test consisted of 16 items that were scored polytomously each with possible scores of 0, 1, 2, 3, 6, or 9. Figure 1 describes the administration and scoring method used for each item.

Of the Grade 5 students, 230 were classified as ELLs and 1,672 as non-ELLs. ELL status was not available for 412 students (17.8% of the Grade 5 sample). Of the Grade 8 students, 144 were classified as ELLs and 1,954 were classified as non-ELLs. ELL status was not available for 356 students (14.5% of the Grade 8 sample). Students from both grades who did not have ELL status information were excluded from the analysis.

Grade 5 non-ELLs earned a mean Science total score of 87.96 (*SD* = 38.66) and grade 5 ELLs earned a mean Science total score of 86.48 (*SD* = 43.52). A comparison of the mean differences between ELLs and non-ELLs using Cohen's *d* indicates that on average, both groups scored similarly on this measure (*d* = -0.04). Grade 8 non-ELLs earned a mean Science total score of 92.32 (*SD* = 37.72) and Grade 8 ELLs earned a mean Science total score of 97.50 (*SD* = 43.86). A comparison of the mean differences using Cohen's *d* indicates that, on average, both ELLs and non-ELLs scored similarly on this measure (*d* = -0.13). The scores on these measures were found to be highly reliable for both grades (Cronbach's $\alpha > .97$). Given that the mean total scores were similar and highly reliable for ELLs and non-ELLs in both grades, total score was determined to be an adequate stratifying variable for the DSF analysis conducted in this study.

Analysis

The comparison of the cumulative and adjacent categories approaches for DSF was conducted using the odds ratio estimator described in Equation 1. The analysis was implemented using DIFAS 4.0 (Penfield, 2005). DIFAS provides estimates of the step-level log-odds ratio DSF effect estimator (*CU-LOR* using the cumulative approach and *AC-LOR* using the adjacent categories approach), the standard error (*SE*) estimator of the DSF effect estimate, and a ratio of each DSF effect estimate over its respective standard error estimator (*z*). In order to deal with sparse data, a decision was made to exclude DSF effect estimates with estimated *SE*s greater than 1.0. This cut-off was chosen since an item with a true DSF effect of zero and a *SE* of one can ultimately result in an estimated DSF effect ranging anywhere from small to large. As a result of this cut-off, roughly 10% of the DSF effects estimated using both the cumulative and adjacent categories approaches were not included in subsequent analyses. In total, 135 steps across the 32 items were included in this study.

Computing *AC-LOR* in DIFAS requires the score levels for the items to consist of consecutive integers. As a result, the data were recoded so as to have equal intervals between score levels. Score levels of 0, 1, 2, and 3, remained the same, score level 6 was recoded to a score of 4, and score level 9 was recoded to a score of 5. The original total score, however, was retained and used as the stratifying variable.

The resulting *CU-LOR* and *AC-LOR* estimates were compared to determine whether the two approaches yielded similar results. This comparison involved an analysis of both the classification categories based on effect sizes using ETS guidelines (i.e., small, moderate, and large DSF effects) and the quantitative differences between the two

estimates. The process of comparing the results occurred in two phases. The first phase consisted of an initial comparison of *CU-LOR* and *AC-LOR* of all steps of all items. This phase consisted of three sets of analyses. First, the mean difference and mean absolute difference between *CU-LOR* and *AC-LOR* estimates were compared for each grade in order to determine, on average, to what extent the results obtained under the cumulative and adjacent categories dichotomization schemes differed. A dependent sample *t*-test was conducted to determine whether the difference between *CU-LOR* and *AC-LOR* was significantly different from zero. The second analysis in this phase examined the distribution of differences to determine the percentage of steps with differences between *AC-LOR* and *CU-LOR* that exceeded 0.25. Differences greater than 0.25 in magnitude would indicate that choosing one approach over another would likely lead to different decisions about flagging a step. The third analysis concerned the classification categories obtained using the two approaches. For this analysis, a three-by-three table was constructed identifying the number of steps classified as having small, moderate, and large DSF effects using both approaches ($|\hat{\lambda}_j| < 0.43$ is small; $0.43 \leq |\hat{\lambda}_j| < 0.64$ is moderate; and $|\hat{\lambda}_j| \geq 0.64$ is large). A χ^2 test of significance was used to determine whether a statistically significant relationship existed between the categories of DSF obtained using both approaches across steps for all items.

The second phase of the analysis focused on items that were flagged as displaying one or more DSF effects for either *AC-LOR* or *CU-LOR* that were: (a) moderate or large, and (b) statistically different from zero. It was important to focus solely on these items to determine how the two approaches compare for items where DSF does exist and therefore, differences obtained between the two approaches can lead to different

conclusions regarding invariance. The selected items were compared and examined further using methods analogous to those described in phase 1 (i.e., examining both quantitative differences and differences in categorical patterns of DSF) to determine the degree of similarity between the two approaches.

Because an issue of primary interest in this study is whether *AC-LOR* and *CU-LOR* lead to the same interpretation of DSF, items that yielded different DSF categories across affected steps were examined. For items demonstrating different patterns of significant DSF effects, a two-step process was implemented to determine to what extent the resulting interpretations were different under the two approaches. The first step was to identify items and steps where there were discrepancies between the patterns of DSF under the two approaches. This was done by comparing the classification categories across all steps (i.e., small, moderate, and large DSF) for each item under the two approaches. Items with different patterns of DSF classification categories were flagged for further review. The second step consisted of an analysis of the interpretation of invariance for each item. The magnitude and location of the DSF effect was compared across the two approaches to determine whether you reach the same conclusion about invariance.

Chapter 4: Results

The results below are presented in order of the two phases of the analyses described in the Method Section. The first phase targeted all 16 items from each grade level and provided an overall comparison of the values of the DSF effect estimates obtained using *CU-LOR* and *AC-LOR*. The second phase was identical to the first phase with the exception that only items with one or more significant DSF effect were included. This phase also investigated the interpretation of invariance for items where different patterns of DSF were identified according to *CU-LOR* and *AC-LOR*. Finally, general trends in the findings are discussed.

Phase I

Table 1 presents the results comparing *CU-LOR* and *AC-LOR*. *CU-LOR* estimates ranged from -1.51 to 0.92 across both 5th and 8th grades with means near zero and standard deviations around 0.40. *AC-LOR* estimates ranged from -1.50 to 1.84 across both 5th and 8th grades with means near zero and standard deviations near 0.55. These values indicate that on average, the DSF effect estimates were not found to favor one group in particular but there is some variability in the DSF effect estimates based on both *CU-LOR* and *AC-LOR* with some steps favoring non-ELLs and other steps favoring ELLs. It should be noted that *AC-LOR* resulted in a somewhat larger range than that obtained from *CU-LOR* estimates. This finding is not surprising given that *AC-LOR* yields less stable results since smaller sample sizes are used to calculate the estimates.

The difference between *CU-LOR* and *AC-LOR* for Grade 5 ranged from -0.41 to 0.86 with a mean of 0.10. This difference was statistically significant [$t(72) = 2.87$, $p = .005$] indicating that for the Grade 5 data, the difference between the two estimates was

significantly different from zero. The mean absolute difference between *CU-LOR* and *AC-LOR* for Grade 5 was determined to be 0.24. This is considered a small difference given that a difference of this magnitude may result in steps being categorized differently under the two approaches. For Grade 8, the mean difference between the two estimates ranged from -1.83 to 1.60 with a mean of -0.01. This was not significantly different from zero [$t(61) = -0.19$, $p = .849$]. This non-significant difference may be misleading. Upon closer inspection of the LORs, there was a large spread of positive and negative LOR differences. When calculating the mean LOR difference, the positive and negative differences cancel out resulting in a mean LOR difference near zero. However, the mean absolute difference between *CU-LOR* and *AC-LOR* for Grade 8 was 0.36. This is slightly larger than what was observed in the Grade 5 data and can be considered a moderate difference since a difference of this magnitude may lead to different decisions about flagging a step based on what approach is used.

A comparison of the percentage of steps that yielded LOR differences greater than 0.25 in absolute magnitude was examined since a difference of this size may result in different classification categories of the magnitude of DSF between the two estimates. In the Grade 5 data, 34% of steps (25 of the 73 steps compared) resulted in LOR differences greater than 0.25. For six of these steps, *AC-LOR* was larger than *CU-LOR*, and for 19 of the steps, *CU-LOR* was larger than *AC-LOR*. In the Grade 8 data, 55% of steps (34 of the 62 steps compared) resulted in LOR differences greater than 0.25. For 19 of the steps, *AC-LOR* was larger than *CU-LOR*, and for 15 of the steps, *CU-LOR* was larger than *AC-LOR*. It was also of interest to determine whether large differences were more prevalent with large DSF effects. Thus, the percentage of steps that yielded LOR differences

greater than 0.25 in absolute magnitude was examined by DSF effect magnitude. Table 2 shows the number and percentage of steps with LOR differences by DSF magnitudes for all items. This table shows that in general, the difference between the LORs increases as DSF effects increase. Additionally, while *CU-LOR* often yielded larger estimates of DSF than *AC-LOR* for the Grade 5 data, in Grade 8, this pattern is not repeated.

In addition to examining the values of the LOR estimates, the classification categories obtained for the estimates were also compared. Table 3 presents three-by-three classification tables for the DSF effect sizes (small, moderate, and large) obtained using *CU-LOR* and *AC-LOR* for both Grade 5 and 8. The diagonal of each table represents the number of steps that were classified identically across the two approaches. The lower off-diagonal elements represent steps where *AC-LOR* yielded a more severe classification of DSF as compared to *CU-LOR* and the upper off-diagonal elements represent steps where *CU-LOR* yielded a more severe classification of DSF as compared to *AC-LOR*. For the Grade 5 data, 53 of the 73 steps (73%) yielded identical DSF effect classification categories under the two approaches. *AC-LOR* yielded a larger classification category of DSF than *CU-LOR* for 19 steps, and one step yielded a larger classification category using *CU-LOR* than *AC-LOR*. A chi-square test of independence of the *CU-LOR* and *AC-LOR* classifications was significant [$\chi^2(4) = 28.91$, $p < .001$].

For the Grade 8 data, 41 of 62 steps (66%) yielded identical DSF effect classification categories under the two approaches. *AC-LOR* yielded a larger classification category of DSF than *CU-LOR* for 14 steps and *CU-LOR* yielded a larger classification category of DSF than *AC-LOR* for seven steps. The chi-square test of

independence for Grade 8 indicated that there is a statistical relationship between the categories of DSF obtained using the two approaches [$\chi^2(4) = 20.41$, $p < .001$].

Examining the results of phase 1, it can be seen that overall, the two approaches often generated consistent results, particularly when the magnitude of DSF was negligible. However, for items where the DSF effect was not small, *AC-LOR* often resulted in larger DSF effects than *CU-LOR*. It should be noted, however, that these classification categories are based on DSF effect size only, and does not take statistical significance into consideration.

Phase II

The second phase of the analysis focused only on the items where one or more steps were found to exhibit a significant ($\alpha = .05$) DSF effect by either *CU-LOR* or *AC-LOR* in Phase I and this effect was moderate or large (note: all significant steps were also moderate or large). These items were examined in isolation to determine how the two approaches compare for items where DSF does exist and therefore, differences obtained between the two approaches can lead to different conclusions regarding invariance. For the items that were flagged, three methods were implemented for examining how similar or different the two approaches were. The first comparison focused on the pattern of significant versus non-significant DSF effects. The second comparison examined raw differences in effect sizes across all steps. The third comparison examined differences in the classification categories of DSF identified by each approach.

Let us first focus on the pattern of significance across the various steps for Grade 5. Table 4 provides the results of the DSF analysis for Grade 5. Six items were flagged as containing significant DSF effects in one or more steps. Of the six items, three (Items 3,

14, and 16) had the same pattern of significant DSF effects using both *CU-LOR* and *AC-LOR*. For each of these items, the same step was flagged by both approaches. For the remaining three items, *CU-LOR* and *AC-LOR* yielded different patterns of significant DSF effects. For Item 6, two steps were flagged using *AC-LOR* only and no steps were flagged using *CU-LOR*. For Item 7, Step 3 was flagged using *CU-LOR* and no steps were flagged using *AC-LOR*. Lastly, Item 11 resulted in one step that was flagged using *AC-LOR* only (Step 2), and one step (Step 3) that was flagged using both *CU-LOR* and *AC-LOR*.

The DSF analysis of the Grade 8 data (see Table 5) resulted in nine flagged items containing one or more significant DSF effects. Examination of the pattern of significance across steps revealed that of the nine items, none resulted in the same pattern of DSF across the two approaches. Four items resulted in one or more steps flagged using *CU-LOR* only. One item was flagged using *AC-LOR* but not *CU-LOR*. In Item 2, Steps 3 and 4 were significant using *CU-LOR* only and Step 5 was significant under both approaches. In Item 9, Step 2 was significant using *AC-LOR* only and Step 5 was significant using both approaches. In Item 12, Step 2 was significant using *CU-LOR* only and Step 5 was significant using both approaches. Lastly, in Item 14, Step 2 was significant using *AC-LOR* only and Step 3 was significant using *CU-LOR* only. The results of this first method of comparison provides evidence that when analyzing the pattern of significant DSF effects, the two approaches often identified different patterns of DSF. Moreover, steps with moderate to large DSF effects were more likely to be significant using *CU-LOR* than *AC-LOR*.

Second, let us discuss differences in raw effect sizes. The mean difference between *CU-LOR* and *AC-LOR* estimates for the six items containing a significant DSF effect for Grade 5 was 0.08. This difference was not statistically significant [$t(25) = 1.09$, $p = .287$]. The distribution of the differences between *AC-LOR* and *CU-LOR* indicated that there was a large variability ($SD = 0.35$) yet these differences cancelled each other resulting in a non-significant mean difference. The mean absolute difference between *AC-LOR* and *CU-LOR* for Grade 5 was 0.28 ($SD = 0.22$). These results were consistent with what was found in the first phase indicating that there is a small to moderate difference in the DSF magnitudes between the two approaches. An important consideration is the presence of differences greater than 0.25 in magnitude since a difference of this size is likely to result in different classification categories of the magnitude of DSF between the two estimates. For the Grade 5 data, 42% of steps (11 of the 26 steps compared) resulted in LOR differences greater than this magnitude. *AC-LOR* was larger than *CU-LOR* in four of these steps, and *CU-LOR* was larger than *AC-LOR* in 7 of these steps. To explore whether differences between *CU-LOR* and *AC-LOR* estimates varied by DSF effect size, the percentage of steps that yielded LOR differences greater than 0.25 in absolute magnitude was examined by DSF effect magnitude (see Table 6). Because fewer items and therefore fewer steps were included in this analysis, the small number of steps in each cell makes it difficult to decipher a particular pattern. Therefore, the results of this analysis are inconclusive.

Next, we will discuss differences in raw effect sizes for the Grade 8 data. The mean difference between *AC-LOR* and *CU-LOR* estimates for the nine items containing a significant DSF effect for Grade was -0.03. This difference was not significantly different

from zero [$t(35) = -0.27$, $p = .789$]. As previously explained, while there was a great deal of variability in the distribution of differences between *AC-LOR* and *CU-LOR* (*SD* = 0.58), the positive and negative differences cancel each other resulting in a non-significant mean difference. The mean absolute difference between *AC-LOR* and *CU-LOR* for Grade was calculated to be 0.40 (*SD* = 0.42). This is slightly larger than what was observed both in the Grade 5 data as well as what was observed across all items indicating that it is possible for the results obtained under the two approaches to result in different interpretations of DSF. A comparison of the percentage of steps that yielded LOR differences greater than 0.25 in absolute magnitude revealed that 56% of steps (20 of the 36 steps compared) resulted in LOR differences greater than this magnitude. *AC-LOR* was larger than *CU-LOR* for 11 of these steps, and *CU-LOR* was larger than *AC-LOR* for 9 of these steps. To explore whether differences between *CU-LOR* and *AC-LOR* estimates varied by DSF effect size, the percentage of steps that yielded LOR differences greater than 0.25 in absolute magnitude was examined by DSF effect magnitude (see Table 6). Similar to the Grade 5 data, the small number of steps in each cell makes it difficult to decipher a particular pattern and the results of this analysis are inconclusive. The results of this second method of comparison provides evidence that when considering the magnitude of the DSF effect and not whether the effect is statistically significant, the two approaches are likely to yield differing conclusions about invariance.

The third method of comparison exploring classification categories obtained for the estimates will now be discussed. As shown in Table 7, for Grade 5 data, 16 of 26 steps (62%) yielded identical DSF effect classification categories under the two approaches. *AC-LOR* yielded a larger classification category of DSF than *CU-LOR* for

nine steps and *CU-LOR* yielded a larger classification category than *AC-LOR* for one step. This pattern remains consistent with the results from Phase I of the analysis in that while the two estimates generally resulted in the same categorization of DSF, when differences did exist, *AC-LOR* tended to yield more severe categorization than *CU-LOR*. A chi-square test of significance revealed that there is a statistical relationship between the categories of DSF obtained using the two approaches [$\chi^2(4) = 11.18$, $p = .025$].

The examination of the classification categories obtained for the Grade 8 data revealed that 24 of 36 (67%) steps yielded identical DSF effect classification categories under the two approaches. *AC-LOR* yielded a larger classification category of DSF than *CU-LOR* for six steps and *CU-LOR* yielded a larger classification category of DSF than *AC-LOR* for six steps. A chi-square test of significance revealed significant results indicating that there is a statistical relationship between the categories of DSF obtained using the two approaches [$\chi^2(4) = 15.34$, $p = .004$]. These results provide evidence that while the two approaches generally classify the DSF effect similarly, in cases where the two approaches lead to differing categorizations of DSF, *AC-LOR* tends to yield a more severe classification.

Lastly, the interpretation of DSF was investigated for items where different patterns of DSF effects were found between *AC-LOR* and *CU-LOR*. In Grade 5, three items were found to display different patterns of significant DSF effects (Items 6, 7, and 11). For Item 6, no steps were flagged as containing significant DSF effects using *CU-LOR*. However, using *AC-LOR* two steps were flagged indicating that earning a score or 3 was easier for ELLs and earning a score of 9 was more difficult for ELLs. On Item 7, while no DSF was found using *AC-LOR*, *CU-LOR* revealed that earning a score of 3 was

more difficult for ELLs. Lastly, on Item 11, the results based on both approaches found that earning a score of 3 was more difficult for ELLs. However, *AC-LOR* also flagged the second step of this item as favoring ELLs indicating that earning a score of 2 was easier for ELLs.

The Grade 8 sample contained nine items that were found to display different DSF patterns. For Item 2, results based on *CU-LOR* indicated that earning scores of 3, 6, and 9 was more difficult for ELLs. *AC-LOR*, however, only found earning a score of 9 to be more difficult for ELLs. On Item 4, *CU-LOR* did not result in any significant DSF effects. However, the results of the DSF analysis using *AC-LOR* indicated that earning a score of 3 was more difficult for ELLs. On Item 5, the results based on *AC-LOR* did not reveal any significant DSF effects. On the other hand, earning scores of 6 or 9 was more difficult for ELLs based on *CU-LOR*. Both *CU-LOR* and *AC-LOR* revealed that scoring at the independent level for Item 9 was more difficult for ELLs. On this item, *AC-LOR* also found that earning a score of 2 was easier for ELLs. On Item 12, both approaches found that earning a score of 9 was easier for ELLs. Additionally, according to *CU-LOR*, earning a score of 2 was also easier for ELLs. On Item 13, *CU-LOR* found earning a score of 6 was easier for ELLs. Results based on *AC-LOR*, however, did not reveal any significant DSF effects for this item. For Item 14, *AC-LOR* and *CU-LOR* each flagged one step as containing a significant DSF effect; however, the step flagged was not the same across the two approaches. According to the results based on *CU-LOR*, earning a score of 3 was easier for ELLs, whereas according to the results based on *AC-LOR*, earning a score of 2 was more difficult for ELLs. The two remaining items yielded significant DSF effects solely using *CU-LOR*: on Item 15, earning a score of 3 was found

to be easier for ELLs, and on Item 16, earning a score of 2 was found to be easier for ELLs. The results above provide strong evidence that the interpretation of invariance and ultimately the decisions regarding item revision or removal are severely impacted by the method used to dichotomize the step function.

Chapter 5: Implications for Practice

In examining DSF effects in Tables 4 and 5 three important trends were observed that have direct implications for applied DSF analyses. The first trend is that the results obtained under the two approaches were often similar, particularly in the case of negligible DSF. Close examination of the results leads to findings that are more similar than originally expected. For example, on 11 of the 15 items found to display significant DSF effects, when attempting to identify the source of noninvariance, both approaches guide the analyst to the same location of the item. Eight items lead to an examination of the Participatory level, two items lead to an examination of the Supported level, and three items lead to an examination of the Independent level. Only on four of the 15 items do the adjacent categories and cumulative approaches lead the analyst to focus on different parts of the item. This is not entirely surprising given that the two approaches use a different dichotomization scheme to define the step functions, and therefore, the steps created under each approach do not mean the exact same thing.

A second trend observed is that within the adjacent categories approach, there seemed to be a lack of independence between adjacent steps that potentially lead to spurious results. Specifically, when an item was found to have a large effect at one step, it was often the case that the previous step displayed a large effect in the opposite direction. Items 6 and 11 from the Grade 5 data and Items 4 and 14 of the Grade 8 data all evidence this trend. Item 6 from the Grade 5 data obtained an *AC-LOR* of 0.61 yielding a classification of moderate DSF on Step 5 and an *AC-LOR* of -0.71 yielding a classification of large DSF for Step 4. Moreover, these steps yielded small DSF effects based on *CU-LOR* (-0.21 and 0.20 for steps 4 and 5, respectively). Item 11 from the Grade 5 data earned an *AC-LOR* of 0.76 yielding a classification of large DSF on Step 3

and an *AC-LOR* of -0.94 yielding a classification of Large DSF for Step 2. The DSF effects based on *CU-LOR* were -0.08 (small) and 0.63 (moderate) for steps 2 and 3, respectively. Item 4 from the Grade 8 data yielded *AC-LORs* of 1.25 for step 3 and -1.28 for step 4. While both of these estimates corresponded to a classification of large DSF, *CU-LOR* yielded a small DSF effect for step 2 (0.31) and a large DSF effect for step 3 (0.73). Lastly, Item 14 yielded *AC-LOR*s of -1.22 (large DSF) for step 3 and 1.84 (large DSF) for step 2. *CU-LOR*s for these steps were 0.01 (small DSF) and -0.64 (large DSF) for steps 2 and 3, respectively. Based on the inconsistencies between *AC-LOR* and *CU-LOR* and the consistently atypical pattern observed with *AC-LOR*, it appears that the results based on *AC-LOR* may be biased representations of the actual invariance effects. As a result of the lack of clarity with *AC-LOR*, it is recommended that *CU-LOR* be used as a DSF effect estimator.

The last, and most important trend observed in this data is that non-negligible DSF effects were more likely to significant using the cumulative approach than the adjacent categories approach. Even large effects found using the adjacent categories approach were often not powerful enough to achieve statistical significance. This was the case for Item 7 in the Grade 5 sample and Items 13, 14, 15, and 16 of the Grade 8 sample. In other cases, steps that were found to possess moderate or large statistically significant DSF effects according to the cumulative approach were not flagged based on the adjacent categories approach. Examples of this trend are Items 2 and 5 of the Grade 8 data.

Based on these trends, it appears that the *CU-LOR* is preferable over the *AC-LOR*. This recommendation is made since several aspects of the results raise concern over the use of the adjacent categories dichotomization scheme when conducting a DSF analysis.

While the two approaches generally lead to similar results as to the location of the DSF effect, there seems to be a lack of independence of the *AC-LOR* since large DSF effects at one step are often paired with large DSF effects in the opposite direction found in the previous step. Additionally, when a substantial DSF effect existed, it was more likely to be significant using *CU-LOR* than *AC-LOR*. Therefore, while the two approaches yield quite similar results, the results based on *CU-LOR* appear more stable and more powerful.

Chapter 6: Discussion

The purpose of this study was to determine how using the cumulative dichotomization scheme versus the adjacent categories dichotomization scheme when conducting a DSF analysis could potentially impact the results and interpretation of invariance as applied to a real dataset. The results of this study indicate that the cumulative and adjacent categories dichotomization schemes yielded quite similar results overall. However, certain portions of the results obtained raised concern over the use of the adjacent categories dichotomization scheme for the purpose of a DSF analysis.

One of the aims of this study was to investigate how the two approaches compared under different DSF patterns (i.e., constant, convergent, and divergent DSF). For items where DSF was not present, the two approaches generally lead to consistent results. The dataset used in this study did not contain items with constant DSF and therefore, how the two approaches compare under this condition was not studied. Items with DSF effects that were isolated to a single score level were observed in this dataset as well as items displaying convergent DSF (i.e., the reference or focal group is favored at multiple score levels) and divergent DSF (i.e., the reference group is favored at one score level and the focal group is favored at another score level within the same item). For items where DSF is present, the two approaches often guided the analyst to focus on the same location of the item. However, a lack of independence of the *AC-LOR* appears to exist since large DSF effects at one step were often paired with large DSF effects in the opposite direction found in the previous step. Also, when a substantial DSF effect existed, it was more likely to be significant using *CU-LOR* than *AC-LOR*. For these reasons, concerns were raised over the results obtained using the adjacent categories dichotomization scheme. Overall, the results from this study are consistent with the

results found by Penfield (2008) indicating that the DSF effects estimated under the cumulative approach were more stable than those estimated under the adjacent categories approach.

An analysis of the classification categories by DSF effect size revealed that the adjacent categories dichotomizations scheme often yielded larger DSF effect estimates as compared to the estimates obtained using the adjacent categories dichotomization scheme. This leads to more steps flagged based only on effect size as exhibiting moderate to large DSF as compared to the results based on the cumulative approach. However, when examining items that displayed different patterns of DSF based on both effect sizes and significance tests more closely, more steps were found to display significant DSF under the cumulative approach. This is most likely because the results from the cumulative approach generally yielded smaller standard errors and therefore were more likely to be statistically significant as compared to the results from the adjacent categories approach. Additionally, the larger sample sizes included in the calculation of *CU-LOR* leads to an increase in power and stability. Based on the results of this study, the use of the cumulative dichotomization scheme is preferable over the use of the adjacent categories dichotomization scheme when using a DSF approach for investigating invariance in polytomous items.

There are some limitations to this study that should be discussed since they may affect the generalizability of the results. For starters, the context of this analysis is specific to a comparison of ELLs to non-ELLs with significant cognitive disabilities on Grade 5 and Grade 8 measures of science. It is unknown whether the results obtained would generalize to different populations and different types of assessments. Second,

several items suffered from sparse data, particularly at the lower score levels either making it impossible to calculate DSF estimates within these steps or, if estimated, resulting in DSF effects with large standard errors. Another limitation of this study is that the LORs were computed based on stratifying an observed score to estimate a target trait. However, in the case where an observed score is not a sufficient statistic for target trait, the resulting DSF effects can be biased, particularly if the distribution of the target trait differs between ELLs and non-ELLs (Penfield & Camilli, 2007). Lastly, because no items were found to display constant DSF in this dataset, it is unknown how these two approaches compare under this condition. The cause of constant DSF is generally due to an item-level problem. In this assessment, each item presents a different set of stimuli for the participatory, supported, and independent levels. As a result, there is little consistency across all steps of a particular item, therefore making it unlikely that the cause of invariance would be at the item level and that constant DSF would be found in any of these items.

Future research can help address these limitations. For example, studies similar to the one executed here should be conducted on a variety of datasets that use polytomous items in order to determine to what extent these results can be generalized across populations and assessment instruments. Second, it would be beneficial to perform a similar analysis on a data set with larger group sizes and thus more stable DSF effect estimates. In order to determine how the two approaches compare under the condition of constant DSF, the analysis would have to be conducted on a dataset consisting of polytomous items that have common item stems for all score levels. Lastly, this study investigates whether the results obtained under the GRM (*CU-LOR*) and GPCM (*AC-*

LOR) are consistent. However, it does not examine how the results compared to the continuation ratio dichotomization scheme, which has also been used in DSF analyses (Penfield, 2008). Future research can address this by including the continuation ratio dichotomization scheme to determine how estimates of DSF effects calculated under this approach will compare to those under the adjacent categories and cumulative approaches.

Despite these limitations, this study presents valuable information about how the cumulative and adjacent categories approaches compare when applied to a real dataset. While the cumulative approach appears to result in more stable and powerful results, as compared to the adjacent categories approach, it is still necessary to determine how the continuation ratio approach holds up. It is also necessary to study polytomous items with common item-level characteristics such as a common item stem as well as datasets of diverse student populations and assessment characteristics.

Appendix A

Literature Review

Fairness

Importance of fairness in testing. It is generally believed within the measurement community and today's society as a whole that tests should be thoughtfully developed and the testing process should be equitable for all students. The context and concept of fairness in education has evolved in relation to social and legal issues dating back to the mid-1800's (Camilli, 2006). Issues of fairness continue to be shaped by the particular social context in which they are embedded. Most concerns related to fairness have evolved with respect to issues of race, ethnicity, and gender. Fairness can also be considered in relation to other issues including special populations and linguistic diversity. The 1999 *Standards for Educational and Psychological Testing* (*Standards*; AERA, APA, NCME, 1999), states that fairness "is subject to different definitions and interpretations in different social and political circumstances" (p. 80).

Reliability and validity are concepts and tools that are prerequisites for test fairness (Camilli, 2006). Yet fairness in testing extends beyond reliability and validity. Camilli states:

> Fairness in testing refers to perspectives on the ways that scores from tests or items are interpreted in the process of evaluating test takers for a selection or classification decision. Fairness in testing is closely related to test validity, and the evaluation of fairness requires a broad range of evidence that includes empirical data, but may also involve legal, ethical, political, philosophical, and economic reasoning (p. 225).

Like validity, the presence of fairness must be supported with evidence that the conditions of testing are equitable and that scores on a test have the same meaning for different subgroups of the population (Camilli, 2006).

The *Standards* (AERA, APA, NCME, 1999) acknowledge while absolute fairness for each individual is impossible, tests can further societal justice related to fairness and equal opportunity when properly designed and more importantly, when considerations of fairness are embedded in the implementation of tests. The concern for fairness in testing and test use continues to be a pervasive issue in test development today (AERA, APA, NCME, 1999). Major testing companies including Educational Testing Service (ETS) and ACT have clearly developed guidelines documenting standards for practice used within their companies (ETS, 2002; ETS, 2008; ACT, 2008; ETS, 2009). Additionally, chapters devoted to fairness can also be found in the *Standards* and the fourth edition of *Educational Measurement* (Brennan, 2006).

There are four conceptions of fairness outlined in the *Standards*: (a) fairness as lack of bias, (b) fairness as equitable treatment in the testing process, (c) fairness as equality in outcomes of testing, and (d) fairness as opportunity to learn. Lack of bias, the first conception of fairness, is violated if "deficiencies in a test itself or the manner in which it is used result in different meanings for scores earned by members of different identifiable subgroups" (AERA, APA, NCME, 1999, p. 74). The *Standards* state that bias refers to "construct-irrelevant components that result in systematically lower or higher scores for identifiable groups of examinees" (AERA, APA, NCME, 1999, p. 76). Camilli and Shepard (1994) describe two basic statistical approaches for detecting bias. External methods use a criterion separate from the test itself to identify bias. Examples of these

external methods include differential prediction and differential selection. Internal methods use a criterion internal to the test. Item bias detection techniques including *differential item functioning* (DIF) are internal methods that are often used to describe bias found at the item level.

The second conception of fairness, equitable treatment in the testing process, requires consideration of the purpose and context of testing as well as the way in which test scores are used. It is not just the manner in which a test is designed that makes it fair or unfair, but the manner in which a test is used that can produce an unfair test (AERA, APA, NCME, 1999). A fair test requires that individual examinees are provided equal opportunity to demonstrate their knowledge within appropriate testing conditions and an equal opportunity to prepare for a test. Fairness as equitable treatment extends to the reporting of individual and group test results. Providing accurate and informative score reports is necessary but not sufficient. The *Standards* (AERA, APA, NCME, 1999) recommend that "confidentiality should be respected [and] scores should be disclosed only as appropriate" (p. 75).

Equality in outcomes of testing is the third conception of fairness. This is the most controversial conception since outcome differences among varying subpopulations do not always signify that a particular test is either biased or unfair. The testing community agrees that if a test is free of bias according to external and/or internal methods and test takers have been afforded fair treatment during the testing process, the test is fair. The *Standards* states that:

> …unequal outcomes at the group level have no direct bearing on questions of test fairness. There may be legal requirements to investigate certain differences in

> outcomes of testing among subgroups. Those requirements further may provide that, other things being equal, a testing alternative that minimizes outcome differences across relevant subgroups should be used (p. 76).

Although disparate impact is generally rejected by the measurement community as a conception of fairness, it is included in the *Standards* since societal and legal views of fairness have traditionally focused on the equality of outcomes in testing in their consideration of fairness (Phillips & Camara, 2006).

The final conception of fairness is a property of educational achievement, providing an opportunity to learn. Achievement testing measures what an examinee can do as a result of formal academic instruction. Low scores on an achievement test are often a result of not having had the opportunity to learn due to inadequate exposure to the assessed curriculum (Wang, 1998). Theoretically, opportunity to learn is conceptually essential but exceedingly difficult to measure (Herman, Klein, & Abedi, 2000).

Item bias in more detail. The remainder of this paper focuses on the first conception of fairness, lack of bias. Bias in educational testing has been a dominant theme in measurement since the mid-1960's (Cole & Moss, 1989). Bias occurs for a variety of reasons. Two potential sources of bias are discussed in the *Standards* (AERA, APA, NCME, 1999): content-related sources and response-related sources. Content-related sources of bias result from inappropriate selection of test content, language that is interpreted differently by members of different groups, and material that is emotionally disturbing or offensive to certain test takers. The presence of content-related sources of bias can be assessed via a sensitivity review (i.e., an inspection of the test itself) conducted by a diverse panel of experts (ACT, 2008; ETS, 2002).

Response-related sources of bias are construct-irrelevant score components that result when test items elicit responses or response processes other than those intended. Sources of response-related bias can include unclear testing instructions and, in the case of performance assessments, scoring rubrics that credit certain responses over other equally correct responses. Response-related bias is assessed via a comparison of the internal structure of test responses for different groups of test takers.

Differential item functioning. Judgmental methods for the review of tests and items by expert panels to identify content or language that can be differentially interpreted by subgroups of examinees are often supplemented by statistical procedures for identifying test items that operate differently among different population subgroups. One widely-used approach for evaluating bias is the framework of differential item functioning (DIF). DIF can be broadly defined as the conditional dependence of group membership and item response (Penfield & Camilli, 2007). The study of DIF traditionally involves the comparison of two groups. The reference group is the group historically advantaged (e.g., whites, males, or native English speakers) whereas the focal group is the group historically disadvantaged (e.g., racial minorities, females, or English language learners). DIF exists when members of the reference and focal groups with the same level of ability do not have the same chance of success on an item. While DIF is often expected to favor the reference group, it can also favor the focal group. Often, within the same assessment instrument, some DIF items will favor the reference group while other DIF items will favor the focal group.

The presence of DIF poses a threat to the validity of test scores. If DIF is present in an item, then there is the implication that, in addition to the target ability an item

measures, a secondary factor related to group membership affects performance on this item. This phenomenon is known as multidimensionality. In some cases, the secondary factor causing the multidimensionality may be related to the construct being measured. In other cases, however, the secondary factor can also be related to the method of administration or some other unintended item property. Only in the second scenario would the item be considered biased.

DIF detection methods vary depending on whether the item under consideration is dichotomous or polytomous. Dichotomous items are items with two possible score options (i.e., correct and incorrect). DIF detection techniques for these items are well-established (Camilli & Shepard, 1994; Penfield & Camilli, 2007) and include item response theory (IRT) approaches, odds ratio approaches, and logistic regression approaches. Polytomous items are items with more than two possible score options (i.e., rating scale items or performance-based items). Investigation of DIF in polytomous items is generally more difficult due to the presence of multiple score categories, which allows for the nature (i.e., magnitude and sign) of the DIF effect to change across the score categories.

DIF in Dichotomous Items

Definition of DIF in dichotomous items. DIF in dichotomous items pertains to the conditional dependence of group membership and correct or incorrect response conditioning on ability (Camilli & Shepard, 1994; Penfield & Camilli, 2007). Because there are only two outcomes (correct or incorrect) the conditional dependence can be examined using the between-group differences in the probability of correct response. The relationship between group membership and item response can change across the ability

continuum. This leads to a distinction between uniform and nonuniform DIF. Uniform DIF exists when the conditional dependence is in the same direction and size across the ability continuum. Thus the same group is at an advantage regardless of the ability level. Nonuniform DIF, on the other hand, exists when the conditional dependence shifts in magnitude or direction across the ability continuum. If the conditional dependence shifts in size, the magnitude of the advantage changes across the ability continuum. For example, one group can have a relatively small advantage at the lower end of the ability continuum and a relatively large advantage at the higher end of the ability continuum. If the conditional dependence shifts in direction, this implies that one group is favored at low ability levels while the other group is favored at high ability levels.

Methods for evaluating DIF. There are a variety of DIF detection procedures for dichotomous items many of which share common theoretical interpretations founded in IRT. IRT consists of several parametric models representing the probability of correct response as a function of ability and one or more item-level parameters including difficulty, discrimination, and guessing. The three-parameter logistic model represents the probability of correct response given a specified level of ability (θ) as

$$P(Y = 1|\theta) = c + (1 - c)\frac{\exp[Da(\theta - b)]}{1+\exp[Da(\theta - b)]} \quad (4)$$

where c represents the guessing parameter, b represents the item difficulty parameter, a represents the item discrimination parameter, and D is a scaling constant equal to 1.7. In the two-parameter logistic model, a more restricted version of the three-parameter model, c is set to equal zero. In the one-parameter logistic model, the most restrictive of IRT models, c is equal to zero and a has a constant value across all items. The function generated by these parametric IRT models results in an S-shaped curve that

mathematically represents the probability of correct response as a function of θ. These curves are known as item characteristic curves (ICCs). An example of an ICC following the three-parameter logistic model is shown in Figure 1. The item parameters for the displayed item are as follows: $a = 1.7$, $b = 0$, and $c = 0.2$.

Comparison of ICCs for reference and focal groups allows analysts to determine if after conditioning on ability, the two groups have the same probability of success on that item at each level of ability. If the ICC of the reference and focal groups differ, then one or more of the item parameters must be different between the groups. Visual inspection of ICCs can facilitate one's conceptualization and assessment of the DIF effect. For example, if the ICC for the focal group is shifted to the right as compared to the ICC for the reference group, but the lines do not cross, then the item has a greater difficulty for the focal group (i.e., the difficulty parameter, b, is higher for the focal group), but item's discrimination parameter (a) is the same for the two groups. Figure 2 displays the ICCs for the reference and focal groups for such an item, which is an example of uniform DIF ($a_R = 1.7$, $b_R = 0$, $c_R = 0$, $a_F = 1.7$, $b_F = 1.5$, and $c_F = 0$). If the ICCs for the two groups cross, then the discrimination parameter is not the same for the reference and focal groups. Figure 3 demonstrates this phenomenon, which is an example of nonuniform DIF ($a_R = 1.75$, $b_R = -0.5$, $c_R = 0$, $a_F = 0.75$, $b_F = -0.25$, and $c_F = 0$).

Within the IRT framework, there are two different interpretations of DIF: between-group differences in an item's ICCs, and between-group differences in an item's parameters. Each of these interpretations results in statistical approaches for detecting DIF. The signed area index (Rudner, Getson, & Knight, 1980) is an index that quantifies the difference between the ICCs of the reference and focal groups. The signed area index

is useful in the case of uniform DIF, but can be misleading when between-group differences in the *a* parameter exist (Penfield & Camilli, 2007). Researchers developed the unsigned area index (Penfield & Camilli, 2007) to address this issue. The signed area and unsigned area indices have multiple drawbacks that limit their applicability and utility (Penfield & Camilli, 2007). For example, large sample sizes are often required for their direct estimation and they do not offer a method for testing the null hypothesis of no DIF (Penfield & Camilli, 2007).

The second interpretation of DIF within the IRT framework is a between-group difference in the item parameters. One index available to identify divergence in item parameters is an investigation of the difference in an item's *b* parameter for the reference group as compared to the focal group (Camilli & Shepard, 1994). Researchers conduct statistical tests of the null hypothesis of no DIF using several methods. The first is a test of equal *b* parameters (Lord, 1980), which is useful when the data follows the one-parameter logistic model. When the data is better described by a two-parameter or three-parameter logistic model, Lord's chi-square test (Lord, 1980) can be used since it simultaneously tests differences between *b*- and *a*- parameters. A third test of DIF can be conducted using a likelihood ratio test which examines the relative fit of two models (Camilli & Shepard, 1994). The studied item's parameters are constrained to be equal in the compact model and allowed to vary between the reference and focal groups in the augmented model. If the augmented model is determined to improve model fit, then it is retained and DIF is said to exist. The likelihood ratio test is a more accurate assessment of DIF than Lord's chi-square and the *b*-parameter difference tests (Camilli & Shepard, 1994) since it can test for differences in both the *a*- and *b*- parameters and it estimates the

variance-covariance matrix for item parameter estimates more accurately than Lord's chi-square (Kim & Cohen, 1995).

In addition to IRT-based methods, other DIF detection methods are available including nonparametric methods such as contingency table approaches that use observed scores such as "total score" as a proxy for latent ability. Several contingency table approaches for studying DIF have been developed including proportion difference measures (Dorans & Kulick, 1986) and the Mantel-Haenszel common log odds ratio estimator (Holland & Thayer, 1988). These nonparametric approaches do not rely on explicit measurement models and are generally accessible to people without technical measurement backgrounds. Consequently, they are of increased practicality in applied settings as compared to IRT methods. Camilli and Shepard (1994, chap. 4) described several advantages of contingency table methods including the ability to implement these approaches with small sample sizes and their increased appeal in applied settings such as test development.

The most widely used contingency table approach, the Mantel-Haenszel common log odds ratio estimator, (Holland and Thayer, 1988; Mantel & Haenszel, 1959) combines the odds ratios, α_j, across trait levels with the formula for a weighted average. The odds ratios are computed by examining the odds that a reference group member will answer an item correctly over the odds that a focal group member will answer an item correctly. If the odds that a member of the reference group will answer an item correct is 1.50 and the odds that a member of the focal group will answer that item correct is .50, then the odds ratio is computed by the ratio of the two, resulting in an odds ratio of 1.5/0.5 = 3.0. The interpretation of this odds ratio is that a member of the reference group

has an odds of answering correctly that is three times greater than the odds of a member of the focal group to answer correctly. For each item, an odds ratio is computed at each score level "*k*" across the "*M*" score levels and the *S* odds ratios are combined in a weighted aggregate. The formula used for computation is given by

$$\hat{\alpha}_{\mathrm{MH}} = \frac{\sum_{k=1}^{M} \frac{A_k D_k}{T_k}}{\sum_{k=1}^{M} \frac{B_k C_k}{T_k}}, \tag{5}$$

where A_j is the number of examinees at the j^{th} total score in the reference group that answered an item correctly, B_j is the number of examinees at the j^{th} total score in the reference group that answered an item incorrectly, C_j is the number of examinees at the j^{th} total score in the focal group that answered an item correctly, D_j is the number of examinees at the j^{th} total score in the focal group that answered the item incorrectly, and T_j is the total number of examinees at the j^{th} total score that responded to the item. A unique property of $\hat{\alpha}_{\mathrm{MH}}$ is that it is theoretically proportional to IRT's *b*-parameter difference under the one-parameter logistic model. However, this equivalence does not hold under the two- or three-parameter logistic models (Penfield & Camilli, 2007).

The Mantel-Haenszel odds ratio is not a symmetric measure of DIF (i.e., values of DIF favoring the focal group can range from below 1 to zero and values of DIF favoring the reference group can range from above 1 to infinity). To improve its interpretation, $\hat{\alpha}_{\mathrm{MH}}$ can be transformed by the natural logarithm to yield

$$\hat{\lambda}_{\mathrm{MH}} = \ln(\hat{\alpha}_{\mathrm{MH}}). \tag{6}$$

The resulting index yields positive values when DIF favors the reference group and negative values when DIF favors the focal group. A value of zero is indicative of no DIF. A hypothesis test of no DIF can be tested using

$$z = \frac{\hat{\lambda}_{\mathrm{MH}}}{\sqrt{Var\left(\hat{\lambda}_{\mathrm{MH}}\right)}}, \tag{7}$$

where

$$Var\left(\hat{\lambda}_{\mathrm{MH}}\right) = \frac{\sum_{k=1}^{S} T^{-2}(A_k D_k + \hat{\alpha}_{\mathrm{MH}} B_k C_k)(A_k + D_k + \hat{\alpha}_{\mathrm{MH}} B_k + \hat{\alpha}_{\mathrm{MH}} C_k)}{2\left[\sum_{k=1}^{S} \frac{A_k D_k}{T_k}\right]^2}. \tag{8}$$

The following classification scheme, developed by ETS (Zieky, 1993), lists criteria for differentiating between small, moderate, and large DIF effects based on the results of the hypothesis test of no DIF and the magnitude of $\hat{\lambda}_{\mathrm{MH}}$. If $\hat{\lambda}_{\mathrm{MH}}$ is not significantly different from zero, and $\left|\hat{\lambda}_{\mathrm{MH}}\right| < 0.43$, the effect is considered small; if $\hat{\lambda}_{\mathrm{MH}}$ is significantly different from zero and either (a) $\left|\hat{\lambda}_{\mathrm{MH}}\right| < 0.64$, or (b) $\hat{\lambda}_{\mathrm{MH}}$ is not significantly greater than 0.43, then the effect is considered moderate, and if $\hat{\lambda}_{\mathrm{MH}}$ is significantly greater than 0.43 and $\left|\hat{\lambda}_{\mathrm{MH}}\right| \geq 0.64$, the effect is considered large. Although $\hat{\lambda}_{\mathrm{MH}}$ yields both a significance test and an effect size measure, the established criteria described above allows analysts to use the effect size on its own. Therefore, while the ETS scheme is based on a combination of effect size and significance test, the focus for this study will be on the effect sizes only. The $\hat{\lambda}_{\mathrm{MH}}$ has received widespread use due to its computational simplicity and good efficiency (Penfield & Camilli, 2007).

A third approach for detecting DIF in dichotomous items is logistic regression (Swaminathan & Rogers, 1990). Like IRT, it is a parametric approach, yet it uses the observed score as the matching criterion, like the Mantel-Haenszel. The logistic regression approach models the probability of correct response for a particular item as a function of observed test score (*X*), group membership (*G*), and the interaction of the two. This results in the following logistic model:

$$P(Y = 1|X, G) = \frac{\exp(\beta_0+\beta_1 X+\beta_2 G+\beta_3 XG)}{1+\exp(\beta_0+\beta_1 X+\beta_2 G+\beta_3 XG)}, \quad (9)$$

where the coefficient β_2 represents the effect of group membership and the coefficient β_3 represents the interaction between group membership and observed test score. In the case of no DIF, the parameters β_2 and β_3 would equal zero indicating that there is no effect of group membership on the probability of correct response to the studied item. In the case of uniform DIF, β_2 would be non-zero, but β_3 would be equal to zero indicating that there is an effect of group membership on the probability of correct response to the studied item, but that effect is stable across all levels of ability (i.e., observed test score). In the case of nonuniform DIF, the value β_2 may or may not be equal to zero and the value of β_3 would be non-zero indicating that there is an effect of group membership on the probability of correct response and thus the effect of group membership varies across different levels of ability.

There is a similarity between logistic regression and $\hat{\lambda}_{\mathrm{MH}}$; under the condition of uniform DIF (i.e., β_3 = 0), β_2 is equivalent to $\hat{\lambda}_{\mathrm{MH}}$ defined in Equation 3 (Penfield & Camilli, 2007). Nonetheless, logistic regression has the capability of measuring nonuniform DIF, whereas the $\hat{\lambda}_{\mathrm{MH}}$ does not.

Using logistic regression, it is possible to test the null hypothesis of no DIF by evaluating a series of nested models that differ with respect to the DIF parameters introduced in the model. Examining the extent to which the models fit the data provides information as to the form of DIF that exists. The most complex of the models (Model 3) includes test score (*X*), group membership (*G*), and the interaction of observed score and group membership (*XG*) as predictors of the probability of correct response and is given by Equation 6, which contains the exponent of

$$z = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG. \quad (10)$$

Simpler models introduce fewer parameters (i.e., Model 2 does not include the interaction, and Model 1 does not include the interaction or group membership). Once the models are established, their likelihoods can be compared by conducting the following χ^2 test with 1 degree of freedom:

$$\chi^2 = 2\ln\left[\frac{L(\text{Model 3})}{L(\text{Model 2})}\right], \quad (11)$$

where $L(\text{Model 2})$ corresponds to the likelihood of Model 2 and $L(\text{Model 3})$ corresponds to the likelihood of Model 3. A non-significant χ^2 indicates that β_3 is equal to zero, and therefore nonuniform DIF does not exist. Similarly, a second likelihood ratio test can be conducted whereby the likelihood of Model 2 is compared to the likelihood of Model 1. In this case, a non-significant χ^2 indicates that β_2 is also equal to zero, and therefore uniform DIF also does not exist.

There are multiple advantages to using logistic regression over other approaches including IRT and common odds ratio approaches (Penfield & Camilli, 2007). One advantage is the ability to model both uniform DIF and nonuniform DIF, providing a comprehensive and flexible DIF detection mechanism. A second advantage is the ability to successfully implement logistic regression despite small group sizes since an observed ability estimate (observed test score) is used. Nevertheless, this approach is based on strong assumptions including that *X* is a valid representation of ability and that the logistic regression model accurately represents the relationship between the observed test score and the probability of correct response (Camilli & Shepard, 1994).

Selecting the most appropriate DIF detection technique for dichotomous items is dependent on of the form of DIF present (Penfield & Camilli, 2007). For uniform DIF,

the Mantel-Haenszel approach is more powerful than logistic regression (Hidalgo & Lopez-Pina, 2004). For nonuniform DIF, the power of logistic regression exceeds that of the Mantel-Haenszel approach (Rogers & Swaminathan, 1993).

DIF in Polytomous Items

Definition of DIF in polytomous items. Polytomous items are items with multiple score levels such as rating scale items and performance-based items. When considering DIF in polytomous items, the form of DIF can vary across score levels. As a result, DIF in polytomous items is conceptually more complex than dichotomous DIF. It is possible for one score level to exhibit DIF, but not others. Additionally, within a single item, DIF can favor the focal group at one score level and the reference group in another score level.

The evaluation of DIF in polytomous items can be conducted using two different approaches: an item-level (omnibus) approach and an approach that targets individual score levels. The item-level approach addresses item-level invariance and measures the overall effect across all score levels. The item-level approach is most commonly used (Penfield & Lam, 2000) and provides the analyst with a single item-level index of DIF. Examples of this omnibus approach include the standardized mean difference (Dorans & Schmitt, 1991), Mantel's chi-square (Mantel, 1963), Generalized Mantel-Haenszel (Somes, 1986), polytomous SIBTEST (Chang, Mazzeo, & Roussos, 1996), logistic discriminant function analysis (Miller & Spray. 1993), cumulative common log-odds ratio estimator (Penfield & Algina, 2003), and Cox's B (Camilli & Congdon, 1999). These approaches provide a single item-level test of DIF and do not inform the analyst as to which score levels are manifesting the DIF effect, hence providing limited information

to help guide item revision. A second approach for investigating DIF in a polytomous item targets individual score levels. Within this framework, DIF exists if there is a between-group difference in the conditional probability associated with any score level. This approach has the advantage of informing which score level is manifesting the DIF effect. The study of DIF in polytomous items that targets individual score levels is known as differential step functioning (DSF; Penfield, 2007, 2008).

Differential step functioning. DSF provides information about the relative difficulty of advancing or "stepping" from each score level to a successively higher score level. Unlike item-level polytomous DIF approaches, DSF methods provide specific information describing the manifestation of the DIF effect within particular score levels. These techniques serve a diagnostic role in identifying the individual score levels involved in the item's DIF effect. Analysis of DSF has been undertaken using a variety of approaches including IRT-based approaches under the graded response model (Cohen, Kim, & Baker, 1993) and the partial credit model (Penfield, Myers, and Wolfe, 2008), logistic regression (French & Miller, 1996), and a common odds ratio approach (Penfield, 2007).

Methods evaluating DSF begin with the construction of a step function. A polytomous item having r score levels consists of $J = r - 1$ step functions. Each step function describes the probability of successfully advancing to the next highest score level and is defined using a two-parameter logistic model:

$$P(Y \geq j|\theta) = \frac{\exp[a(\theta - b_j)]}{1+\exp[a(\theta - b_j)]} \quad , \tag{12}$$

where b_j represents a difficulty parameter that is specific to step j and a is a discrimination parameter that is constant across all steps. An item with four score levels

will consist of three steps. The first step function describes the probability of advancing from a score of 0 to a score greater than 0; the second step function describes the probability of advancing from a score of 1 or less to a score greater than 1; and the last step function describes the probability of advancing from a score of 2 or less to a score of 3.

DSF exists in many forms including constant, convergent, and divergent DSF (Penfield, Alvarez, & Lee, 2009; Penfield, Gattamorta, & Childs, 2009). Constant DSF is observed when the step levels displaying a DSF effect are relatively equal in magnitude and sign. Figure 4 shows the trace lines for the reference and focal groups for an item displaying constant DSF (a = 1.7, b_1 = -1.5, b_2 = 0, b_3 = 1.5, and the DSF at each step = 0.3). The presence of constant DSF indicates that the factor responsible for the DSF effect is either a property of the item or it is shared across all steps. Convergent DSF describes the condition where affected steps display a DSF effect of the same sign, (i.e., favoring the same group) but of different magnitudes. Figure 5 shows the trace lines for the reference and focal groups for an item displaying convergent DSF (a = 1.7, b_1 = -2.1, b_2 = -0.5, b_3 = 2, DSF_1 = 0.1, DSF_2 = 0.8, and DSF_3 = 0.1). Convergent DSF occurs when the causal properties of DIF are manifested differently across score levels. Divergent DSF is characterized by affected steps that display DSF effects of opposite signs (i.e., the relative advantage shifts between the reference and focal groups for different steps). Figure 6 shows the trace lines for the reference and focal groups for an item displaying divergent DSF (a = 1.7, b_1 = -2.1, b_2 = -0.5, b_3 = 2, DSF_1 = -0.5, DSF_2 = 0.2, and DSF_3 = 0.5). The presence of divergent DSF implies that the causes of the DSF effects are

different for the affected score levels or that there is more than one causal property responsible for the effect.

Importance of DSF. There are several advantages to investigating both item-level and score-level (i.e., DSF) invariance estimates on the same data set. First, most item-level measures of DIF are relatively insensitive when the sign and/or magnitude of the invariance effect varies across the score levels (Penfield, 2007; Penfield, Alvarez, et al., 2009; Penfield, Gattamorta, et al., 2009). For example when one score level displays DSF favoring one group and another score level displays DSF favoring the other group, the DSF effects will cancel to yield a near-zero item-level DIF effect. A second example can occur when a DSF effect may be isolated at a single score level resulting in an insignificant item-level DIF effect despite the violation of measurement invariance in one of the score levels.

Second, DSF provides a more comprehensive picture of the violation of invariance than item-level DIF alone. For example, in the case when the magnitude and/or sign of the DSF effect changes across score levels, conducting an item-level DIF analysis alone may not provide insight as to what is causing lack of invariance in that item. In contrast, a DSF analysis allows the analyst to pinpoint precisely which score levels are responsible for the lack of measurement invariance and may also provide insight as to the causes of DIF. For instance, when a similar DSF effect is present across all score levels it would appear to be an item-level effect attributable to a factor that is common across the entire item such as the item stem or the content being measured by that item. However, if a DSF effect is isolated to one or a few score levels, the cause may be due to a problem specific to that particular score level (Penfield, Alvarez et al., 2009).

This can be demonstrated by considering a writing task where students are asked to respond to a particular prompt. In such an item, the presence of DSF at all steps would imply that the factor responsible for the non-invariance is inherent to the content of the prompt itself. In contrast, the presence of DSF isolated to particular score levels implies that the factor causing DIF may be isolated to one, or a few of the score levels. Using the writing task as an example, DSF may only be associated with score levels related to grammar, but not organization/structure of the paragraph.

Methods for evaluating DSF. Now that the definition of a step function has been established, it is possible to describe several approaches used to evaluate DSF effects. Several IRT-based approaches have been proposed for evaluating DSF in polytomous items. The first, Cohen et al.'s (1993) approach under the graded response model, considers both the signed and unsigned area between the item true score functions obtained for the reference and focal groups as well as a comparison of item parameters based on Lord's χ^2 (1980) for each of an item's J steps. Recently, Penfield et al. (2008) presented both parametric and nonparametric methods for modeling item-level and step-level invariance under the partial credit model. IRT approaches for investigating DSF are based on examining the between-group differences in the b_j parameters of the jth step function. Therefore, the DSF effect for the jth step can be defined as the between-group difference in the b_j parameter given by:

$$\Delta b_j = b_{jF} - b_{jR}. \tag{13}$$

Both of the IRT approaches described above examine between-group differences in step functions that parallel the dichotomous DIF examples described earlier. Once the

polytomous item is dichotomized to create the associated *J* step-level response variables, a separate analysis is conducted at each of the *J* steps.

French and Miller (1996) proposed an approach for investigating both uniform and nonuniform DSF based on logistic regression whereby multiple tests of the null hypothesis of no DIF are conducted at each step using a likelihood ratio test. French and Miller modeled DSF using the adjacent categories and cumulative approaches, as well as a third approach known as the continuation ratio approach. Logistic regression for polytomous items is equivalent to performing a dichotomous logistic regression (shown in Equations 6-8) for each of the step functions associated with a particular item.

DSF can also be examined by employing Penfield's (2007, 2008) odds ratio approach to compare the odds of successfully advancing at the *j*th step for reference and focal group members with the same observed score. A ratio of the odds of success for the reference group over the odds of success for the focal group is calculated using the equation:

$$\widehat{\alpha}_j = \frac{\sum_{k=1}^{S} A_{jk} D_{jk} / T_{jk}}{\sum_{k=1}^{S} B_{jk} C_{jk} / T_{jk}}, \tag{14}$$

where *k* represents a stratum of ability for each of the *S* strata, *j* represents the steps, A_{jk} and B_{jk} represent reference group members that were successful and unsuccessful at the *j*th step, respectively, and C_{jk} and D_{jk} represent focal group members that were successful and unsuccessful at the *j*th step, respectively, and T_{jk} represents the sum of A_{jk} , B_{jk} , C_{jk} , and D_{jk}.

The natural logarithm of $\widehat{\alpha}_j$ for the *j*th step is denoted by $\hat{\lambda}_j$, where positive values of $\hat{\lambda}_j$ indicate DSF favoring the reference group, negative values of $\hat{\lambda}_j$ indicate DSF

favoring the focal group, and a value of zero is indicative of no DSF. The resulting estimator is consistent in scale and direction with the Mantel-Haenszel common log-odds ratio, and as such can be interpreted in a manner consistent with the ETS classification scheme (Zieky, 1993) whereby $\left|\hat{\lambda}_j\right| < .43$ corresponds to a small DSF effect, $.43 \leq \left|\hat{\lambda}_j\right| < .64$ corresponds to a moderate DSF effect, and $\left|\hat{\lambda}_j\right| \geq .64$ corresponds to a large DSF effect. An estimator of the standard error of $\hat{\lambda}_j$ can be obtained using the formula:

$$SE\left(\hat{\lambda}_j\right) = \sqrt{\frac{\sum_{k=1}^{S} T_k^{-2}\left(A_{jk}D_{jk} + \hat{\propto}_j B_{jk}C_{jk}\right)\left(A_{jk} + D_{jk} + \hat{\propto}_j B_{jk} + \hat{\propto}_j C_{jk}\right)}{2\left(\sum_{k=1}^{S} \frac{A_{jk}D_{jk}}{T_k}\right)^2}}. \quad (15)$$

Dividing $\hat{\lambda}_j$ by its estimated standard error yields the test statistic

$$z\left(\hat{\lambda}_j\right) = \frac{\hat{\lambda}_j}{SE\left(\hat{\lambda}_j\right)}. \quad (16)$$

When observed test score is an adequate approximation for ability and the data fit the IRT model employed, the common log-odds ratio is approximately proportional to the between-group difference in the b_j values for the jth step, where the proportionality is determined by the item-level discrimination parameter, a. In the logistic regression framework, there is an equivalence of the estimated value of β_{j2} and the common log-odds ratio, $\hat{\lambda}_j$.

Adjacent categories vs. cumulative approaches. The measurement community uses two primary forms of step functions in widely used IRT models. The first approach defines each of the J step functions using the adjacent categories approach, which is consistent with the generalized partial credit model (GPCM; Muraki, 1992). Under the GPCM, the jth step function specifies the probability of successfully advancing from score level $j-1$ to score level j. A polytomous item with four score levels (i.e., 0, 1, 2, and 3) would

consist of three steps: step 1 would be defined by successfully advancing from a 0 to a 1, step 2 would be defined by successfully advancing from a 1 to a 2, and step 3 would be defined by successfully advancing from a 2 to a 3.

The second form of the step function is the cumulative approach, which is consistent with the graded response model (GRM; Samejima, 1997). Under the GRM, the *j*th step function specifies the probability of successfully advancing from $0, 1, \ldots, j$ to $j + 1, \ldots, J$. Following the example of an item with four score levels, under the GRM, step 1 would be defined by successfully advancing from a 0 to a 1, 2, or 3; step 2 would be defined by successfully advancing from a 0 or a 1 to a 2 or a 3; and step 3 would be defined by successfully advancing from a 0, 1, or 2 to a 3. It is important to recognize that because these two models define the step function differently, the resulting interpretation of the step-level parameters depends on the approach used to define the step function.

Figure 1 Scoring and Administration Guide

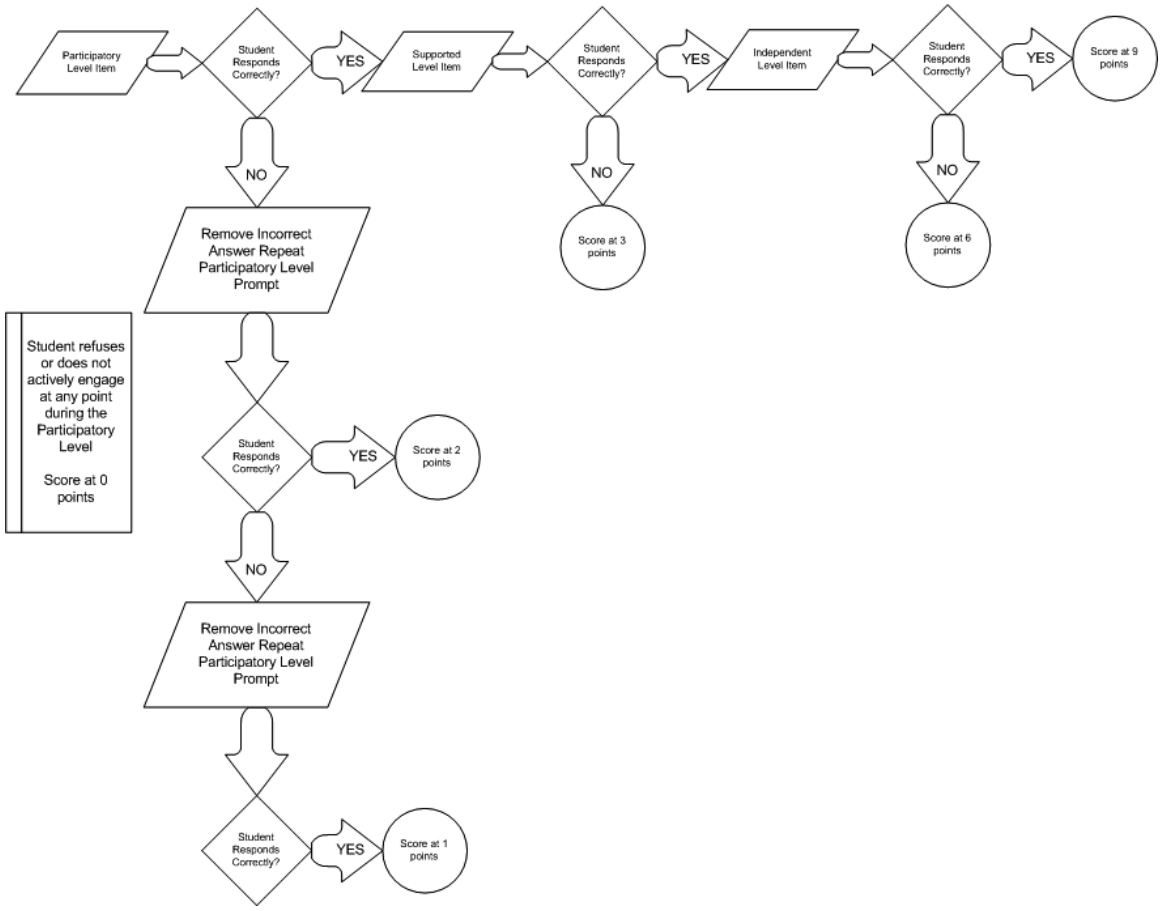

Figure 2 ICC following the 3PL Model

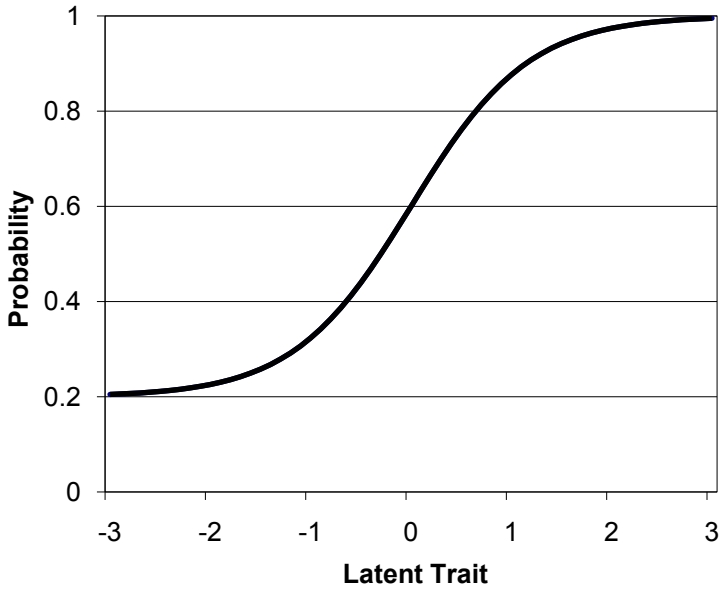

Figure 3 ICCs for Uniform DIF Item

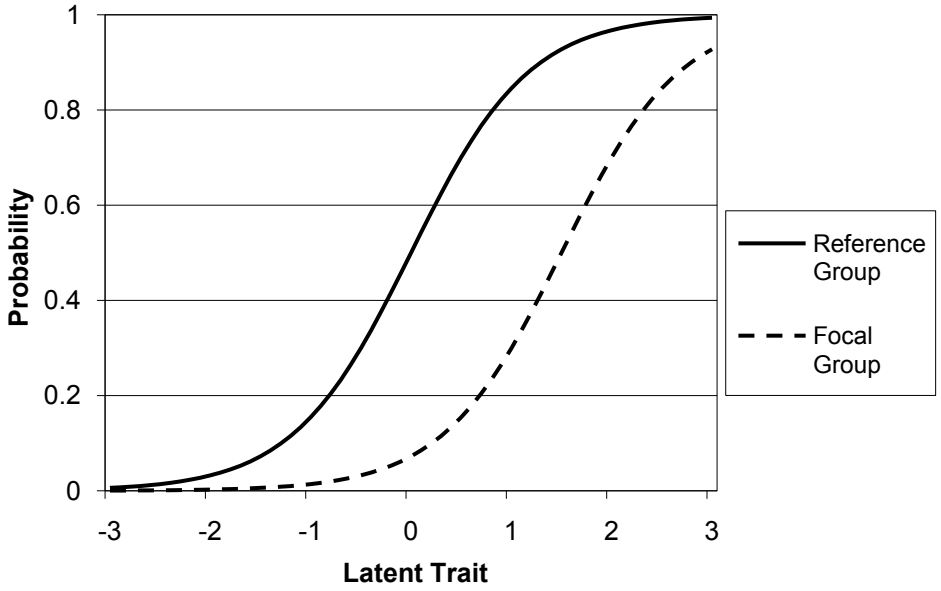

Figure 4 ICCs for Nonuniform DIF Item

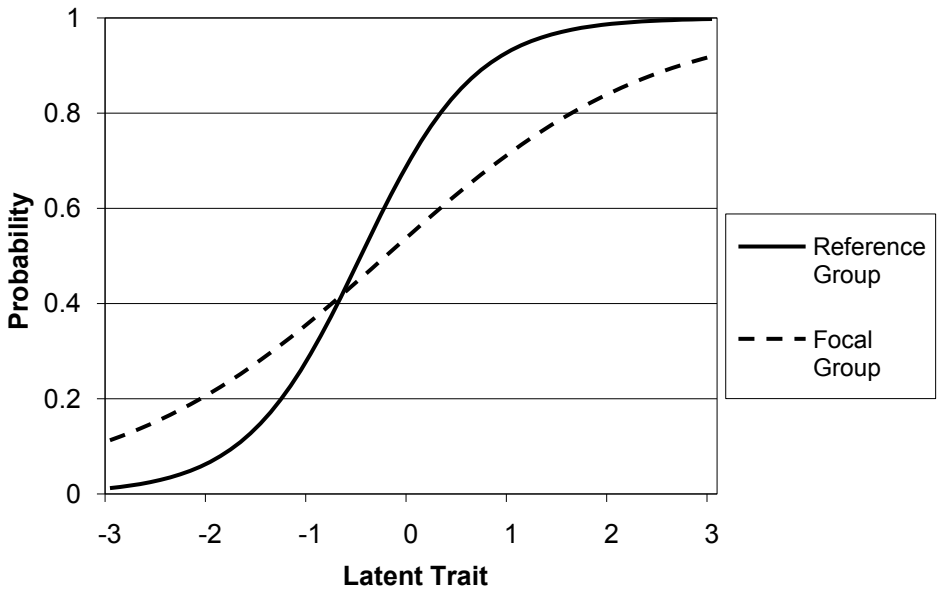

Figure 5 Trace Lines for a Polytomous Item Displaying Constant DSF

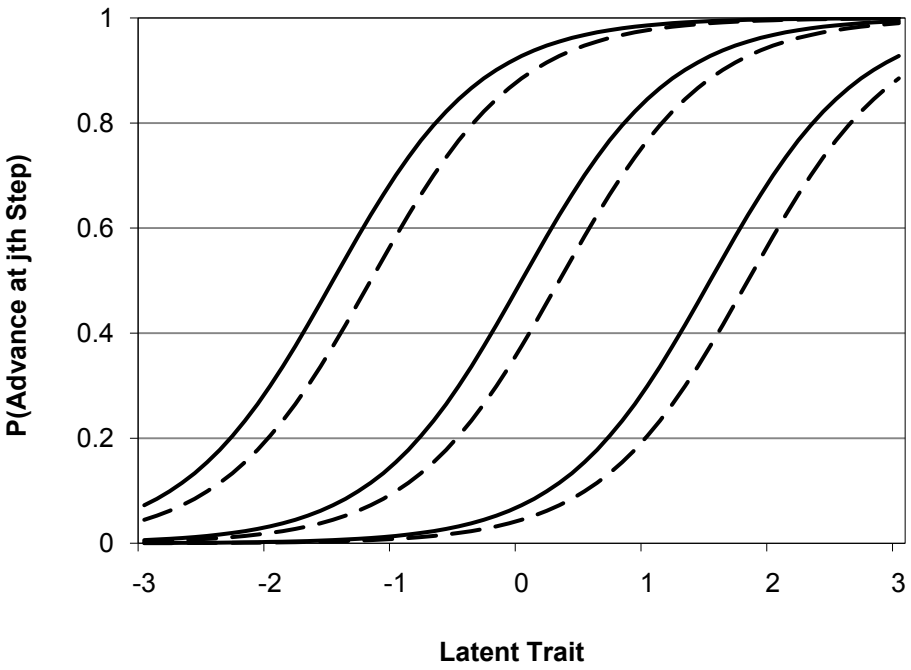

Figure 6 Trace Lines for a Polytomous Item Displaying Convergent DSF

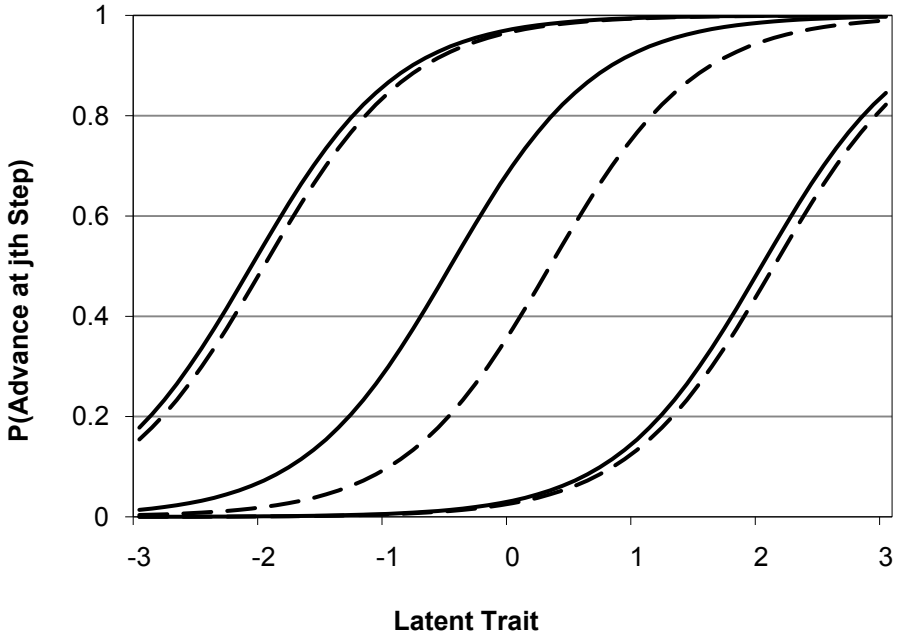

Figure 7 Trace Lines for a Polytomous Item Displaying Divergent DSF

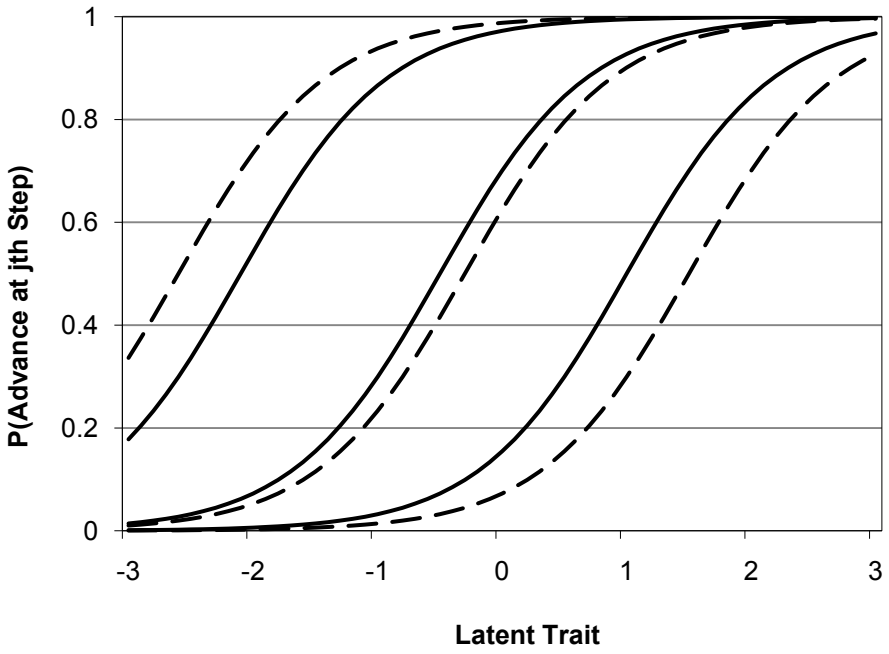

Table 1

Descriptive Results for the Comparison Between *CU-LOR* and *AC-LOR*

	Grade 5 (*N*=73)				Grade 8 (*N*=62)			
	Min	Max	Mean	*SD*	Min	Max	Mean	*SD*
CU-LOR	-0.92	0.92	0.01	0.32	-1.51	0.92	-0.06	0.50
AC-LOR	-1.50	0.85	-0.09	0.48	-1.47	1.84	-0.05	0.64
CU-LOR – AC-LOR	-0.41	0.86	0.10	0.30	-1.83	1.60	-0.01	0.50
|*CU-LOR – AC-LOR*|	0.01	0.86	0.24	0.21	0.00	1.83	0.36	0.35

Table 2

Percentage of Steps with Large LOR Differences by DSF Magnitude for all Items

	Magnitude of DSF Effect	$\lvert CU\text{-}AC\rvert \leq 0.25$	$(CU - AC) > 0.25$	$(AC - CU) > 0.25$	Total
Grade 5	Small	35 (71%)	10 (20%)	4 (8%)	49 (100%)
	Moderate	6 (75%)	0 (0%)	2 (25%)	8 (100%)
	Large	7 (44%)	9 (56%)	0 (0%)	16 (100%)
	Total	48 (66%)	19 (26%)	6 (8%)	73 (100%)
Grade 8	Small	17 (63%)	6 (22%)	4 (15%)	27 (100%)
	Moderate	5 (36%)	2 (14%)	7 (50%)	14 (100%)
	Large	5 (24%)	8 (38%)	8 (38%)	21 (100%)
	Total	27 (44%)	16 (26%)	19 (31%)	62 (100%)

Table 3

Classification of DSF Effects for all Items

			CU-LOR DSF Effect			
			Small	Moderate	Large	Total
AC-LOR DSF Effect	Grade 5	Small	49	1	0	50
		Moderate	6	1	0	7
		Large	7	6	3	16
		Total	62	8	3	73
	Grade 8	Small	31	5	2	38
		Moderate	6	1	0	7
		Large	6	2	9	17
		Total	43	8	11	62

Table 4

Items with Statistically Significant DSF Effects in Grade 5

Item		Step				
		1	2	3	4	5
3	CU-LOR	-0.59	-0.47	-0.24	-0.54*	0.12
	AC-LOR	-0.72	-0.27	0.14	-0.78*	0.34
6	CU-LOR	NA	-0.17	0.12	-0.21	0.20
	AC-LOR	NA	-0.16	0.36	-0.71*	0.61
7	CU-LOR	0.15	0.38	0.59*	0.29	0.23
	AC-LOR	-0.19	-0.26	0.65	-0.19	0.15
11	CU-LOR	NA	-0.08	0.63*	-0.07	0.13
	AC-LOR	NA	-0.97*	0.76*	-0.40	0.36
14	CU-LOR	NA	-0.33	0.19	-0.03	0.48*
	AC-LOR	NA	-1.05	0.21	-0.18	0.68*
16	CU-LOR	NA	-0.29	-0.18	-0.19	-0.92*
	AC-LOR	NA	-0.35	0.07	0.13	-1.00*

Note: Significant LOR are marked by an asterisk (*)

Table 5

Items with Statistically Significant DSF Effects in Grade 8

Item		Step				
		1	2	3	4	5
2	CU-LOR	NA	0.14	0.51*	0.71*	0.92*
	AC-LOR	NA	-0.17	0.39	0.15	1.19*
4	CU-LOR	NA	0.31	0.73	0.29	0.28
	AC-LOR	NA	-1.28	1.25*	0.08	0.28
5	CU-LOR	NA	0.69	0.41	0.61*	0.62*
	AC-LOR	NA	-0.06	-0.23	0.46	0.19
9	CU-LOR	NA	-1.08	-0.13	0.33	0.80*
	AC-LOR	NA	-1.47*	-0.07	0.17	0.82
12	CU-LOR	NA	-1.34*	-0.45	0.24	-0.58*
	AC-LOR	NA	-0.90	-0.11	0.62	-0.80*
13	CU-LOR	NA	0.26	-0.31	-0.64*	-0.30
	AC-LOR	NA	0.61	-0.14	-0.88	-0.22
14	CU-LOR	NA	0.01	-0.64*	-0.12	-0.01
	AC-LOR	NA	1.84*	-1.22	-0.02	0.09
15	CU-LOR	NA	-0.61	-1.04*	-0.05	-0.16
	AC-LOR	NA	0.78	-1.31	0.15	-0.19
16	CU-LOR	NA	-1.51*	-0.54	-0.42	-0.39
	AC-LOR	NA	-1.13	0.04	-0.14	-0.29

Note: Significant LOR values are marked by an asterisk (*)

Table 6

Percentage of steps with Large LOR Differences by DSF Magnitude for Items with Significant DSF Effects

	Magnitude of DSF Effect	$\lvert CU\text{-}AC\rvert \leq 0.25$	$(CU - AC) > 0.25$	$(AC - CU) > 0.25$	Total
Grade 5	Small	8 (53%)	4 (27%)	3 (20%)	15 (100%)
	Moderate	1 (50%)	0 (0%)	1 (50%)	2 (100%)
	Large	6 (67%)	3 (33%)	0 (0%)	9 (100%)
	Total	15 (58%)	7 (27%)	4 (15%)	26 (100%)
Grade 8	Small	11 (79%)	2 (14%)	1 (7%)	14 (100%)
	Moderate	2 (29%)	1 (14%)	4 (57%)	7 (100%)
	Large	3 (20%)	6 (40%)	6 (40%)	15 (100%)
	Total	16 (44%)	9 (25%)	11 (31%)	36 (100%)

Table 7

Classification of DSF Effects for Items with Flagged Steps

			CU-LOR DSF Effect			
			Small	Moderate	Large	Total
AC-LOR DSF Effect	Grade 5	Small	15	1	0	16
		Moderate	1	0	0	1
		Large	3	5	1	9
		Total	19	6	1	26
	Grade 8	Small	14	4	2	20
		Moderate	2	1	0	3
		Large	2	2	9	13
		Total	18	7	11	36

References

ACT (2008). *Fairness Report for the ACT Tests*. Iowa City, IA: ACT.

Alvarez, K., & Penfield, R. D. (November, 2007). Differentiating between step-level and item-level invariance using WINSTEPS. Paper presented at the 2007 meeting of the Florida Educational Research Association, Tampa, FL.

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), (1999). *Standards for educational and psychological testing*. Washington, DC: APA.

Bishop, N. S., & Omar, M. H. (2002, Month). *Comparing vertical scales derived from dichotomous and polytomous IRT models for a test composed of testlets*. Paper presented at the annual meeting of the National Council on measurement in Education, New Orleans, LA.

Brennan, R. L. (Ed.). (2006). *Educational Measurement (4th ed.).* American Council on Education and Praeger, Westport, CT.

Camilli, G. (2006). Test Fairness. In Brennan, R. L. (Ed.), *Educational Measurement: Fourth Edition*. (pp. 221-256). American Council on Education and Praeger, Westport, CT.

Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics*, *24*, 323-341.

Camilli, G., & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. Newbury Park, CA: Sage.

Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, *33*, 333-353.

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement*, *17*, 335-350.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational Measurement (3rd ed.)*. (pp. 201-220). American Council on Education and Oryx Press, Pheoniz, AZ.

Cox, D. R., (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society B*, *20*, 215-232.

Dorans, N. J., & Kulick, E. M. (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. J. (1991). *Constructed item response and differential item functioning: A pragmatic approach* (Research Rep. No. 91-47). Princeton, NJ: Educational Testing Service.

Educational Testing Service (2002). *ETS Standards for Quality and Fairness*. Princeton, NJ: Educational Testing Service.

Educational Testing Service (2008). *ETS Fairness Review Guidelines*. Princeton, NJ: Educational Testing Service.

Educational Testing Service (2009). *Guidelines for the Assessment of English Language Learners*. Princeton, NJ: Educational Testing Service.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, *33*, 315-332.

Gattamorta, K. A., & Penfield, R. D. (2008). *Incorporating a taxonomy of differential step functioning to guide the interpretation of DIF in polytomous items.* Paper presented at the 2008 meeting of the National Council on Measurement in Education, New York, NY.

Glas, C. A. W., & Dagohoy, A.V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika*, *72*, 159-180.

Herman, J. L., Klein, D. C. D., & Abedi, J. (2000). Assessing students' opportunity to learn: teacher and student perspectives. *Educational Measurement: Issues and Practice*, *19(4)*, 16-24.

Hidalgo, M. D., & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*, 903-915.

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

Kim, S.-H., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education*, *8*, 291-312.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Lawrence Erlbaum, Hillsdale, NJ.

Mantel, N. (1963). Chi-square tests with one degree of freedom. Extension of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, *58*, 690-700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *58*, 690-700.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, *19*, 91-100.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement*, *30*, 107-122.

Millsap, R. E., & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement*, *16*, 389-402.

Muraki, R. (1992). A generalized partial credit model: Applications of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Penfield, R. D. (2005). DIFAS: Differential item functioning analysis system. *Applied Psychological Measurement*, *29*, 150-151.

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement*, *44*, 187-210.

Penfield, R. D. (2008). Three classes of nonparametric differential step functioning effect estimators. *Applied Psychological Measurement*, *32*, 480-501.

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement*, *40*, 353-370.

Penfield, R. D., Alvarez, K., & Lee, O. (2009). Using a taxonomy of differential step functioning to improve the interpretation of DIF in polytomous items: An illustration. *Applied Measurement in Education*, *22*, 61-78.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C. R. Rao (Eds.), *Handbook of Statistics: Vol. 26. Psychometrics* (pp. 125-167). New York: Elsevier.

Penfield, R. D., Gattamorta, K. A. & Childs, R. A. (2009). An NCME instructional module on using differential step functioning to refine the analysis of DIF in polytomous items. *Educational Measurement: Issues and Practice*, *28(1)*, 38-49.

Penfield. R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice*, *19(3)*, 5-15.

Penfield, R. D., Myers, N. D., & Wolfe, E. W. (2008). Methods for assessing, item, step, and threshold invariance in polytomous items following the partial credit model. *Educational and Psychological Measurement*, *68*, 717-733.

Phillips, S. E., & Camara, W. J. (2006). Legal and Ethical Issues. In Brennan, R. L. (Ed.), *Educational Measurement: Fourth Edition*. (pp. 733-755). American Council on Education and Praeger, Westport, CT.

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

Rudner, L. M., Geston, P. R., & Knight, D. L. (1980). Biased item detection techniques. *Journal of Educational Statistics*, *5*, 213-233.

Samejima, F. (1997). Graded response model. In van der Linden, W. V. & Hambleton, R. K. (Eds.), *Handbook of Modern Item Response Theory* (pp. 85-100). New York: Springer.

Somes, G. W. (1986). The generalized Mantel-Haenszel statistic. *The American Statistician*, *40*, 106-108.

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

Wang, J. (1998). Opportunity to learn: the impacts and policy implications. *Educational Evaluation and Policy Analysis*, *20*, 137-156.

Wang, S., & Wang, T. (2002, Month). *Relative precision of ability estimation in polytomous CAT: A comparison under the generalized partial credit model an graded* response model. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum.