# Differentials of a State Reading Assessment: Item Functioning, Distractor Functioning, and Omission Frequency for Disability Categories

Kentaro Kato, Ross E. Moen, and Martha L. Thurlow,
*University of Minnesota*

*Large data sets from a state reading assessment for third and fifth graders were analyzed to examine differential item functioning (DIF), differential distractor functioning (DDF), and differential omission frequency (DOF) between students with particular categories of disabilities (speech/language impairments, learning disabilities, and emotional behavior disorders) and students without disabilities. Multinomial logistic regression was employed to compare response characteristic curves (RCCs) of individual test items. Although no evidence for serious test bias was found for the state assessment examined in this study, the results indicated that students in different disability categories showed different patterns of DIF, DDF, and DOF, and that the use of RCCs helps clarify the implications of DIF and DDF.*

**Keywords:** differential distractor functioning, differential item functioning, disability, reading assessment

States need to include students with disabilities in their assessment and accountability systems. The No Child Left Behind Act of 2001 (NCLB) requires it; equitable treatment of these students demands it. Excluding these students from assessment reporting would leave them outside the accountability systems that are intended to identify places and ways that education needs to be improved.

The participation of students with disabilities in state assessments has increased greatly during the past 2 decades. A 1993 National Center on Educational Outcomes (NCEO) survey showed that in the early 1990s, most states included fewer than 10% of their students with disabilities in state assessments (Shriner, Spande, & Thurlow, 1994). Changes in states' participation policies dramatically increased these rates. Recent analyses of state annual performance reports showed nearly all states reporting at least 95% of their students with disabilities participating in state assessments (Thurlow, Moen, & Wiley, 2005); this percentage is consistent with NCLB requirements.

Ensuring that students with disabilities are included in assessment and accountability systems is an important step, but it is just a beginning step. Studies such as those by Abedi, Leon, and Mirocha (2003), Klein, Wiley, and Thurlow (2006), and Ysseldyke et al. (1998) reported that when the performance levels of students with disabilities were compared with performance levels of students without disabilities, the performance levels of students with disabilities were lower than those of students without disabilities. On the one hand, if test performance of students with disabilities is low because they cannot in fact do what is required, then the tests are doing exactly the job they are supposed to do. By drawing attention to low performance, tests reveal areas that may need additional effort to improve student learning. On the other hand, if test performance of students with disabilities is low because features of the assessment prevent them from showing what they can do, then the assessment needs to change.

Various attempts have been made to change assessment practices to remove inappropriate barriers to performing well. Elliott, Thurlow, Ysseldyke, and Erickson (1997) described a variety of accommodations that are used to overcome assessment barriers. Thompson, Johnstone, and Thurlow (2002) described how principles of universal design can be employed to help develop assessments from the outset that reduce barriers without requiring accommodations or that make it easier to apply needed accommodations (see also Thompson, Thurlow, & Malouf, 2004).

*Kentaro Kato, Research Assistant, National Center on Educational Outcomes, University of Minnesota, 207 Pattee Hall, 150 Pillsbury Drive SE, Minneapolis, MN 55455; kato0027@umn.edu. Ross E. Moen, Research Associate, National Center on Educational Outcomes, University of Minnesota, 207 Pattee Hall, 150 Pillsbury Drive SE, Minneapolis, MN 55455. Martha L. Thurlow, Director, National Center on Educational Outcomes, University of Minnesota, 207 Pattee Hall, 150 Pillsbury Drive SE, Minneapolis, MN 55455..*

 Copyright © 2009 by the National Council on Measurement in Education 

A particular challenge for assessing reading is that many accommodations and principles for universal design of assessment (UDA) used in mathematics and other content areas rely on reducing inappropriate test difficulty caused by unnecessary reading demands. These same practices cannot be directly applied to reading assessments. Consequently, assessing reading requires looking at different accommodations from those used in other content areas or looking differently at ones that are used in other areas. The Partnership for Accessible Reading Assessment (PARA) is one part of a national effort to find ways of making reading assessment accessible for students with disabilities (see www.readingassessment.info).

As part of the PARA project, Abedi, Leon, and Kao (2007a, 2007b) examined test item characteristics that could cause or signal inappropriate barriers to successful test performance for students with disabilities. They compared the test performance of students with and without disabilities in terms of item bias. Item bias is said to occur if examinees in one group (e.g., students with specific disabilities) are more or less likely to answer a test item correctly than examinees in another group (e.g., students without disabilities) because of some characteristic of the test item or testing situation that is not relevant to the testing purpose (Zumbo, 1999). If a substantial number of items in a test show item bias, then the test is also likely to be biased (test bias), lacking equity for all groups of students.

One of the commonly used methods to detect potential item bias is the analysis of differential item functioning (DIF). DIF means that examinees in different groups show differing probabilities of answering the item correctly after matching on ability level (Zumbo, 1999). The focus of DIF is on the patterns of correct responses, correct responses being directly related to test outcomes (test scores) and test characteristics (reliability, validity, etc.). Green, Crone, and Folk (1989) extended the concept of DIF to distractors, that is, incorrect response options in multiple-choice items. The purpose of differential distractor functioning (DDF) analysis is to flag test items in which distractors are chosen differently by different groups of examinees. As long as individual items are scored dichotomously (i.e., correct vs. incorrect), which is the case for most tests, the behavior of distractors and their group differences does not always affect resulting scores and their interpretations. When we observe DDF, however, it suggests that those items probably mean something different for different groups of examinees, and the test scores cannot be interpreted in the same manner for all groups (Green et al., 1989). More importantly, if examinees in one group tend to choose a certain distractor instead of a correct response more often than those in the other group, it does affect resulting scores. Thus, DIF and DDF analyses are both important to identify potentially biased items.

In usual DIF analysis, the proportion of correct responses is compared between the reference and target groups conditional on some kind of overall test score (number-correct scores, scale scores, ability scores in the item response theory, etc.) that represents the ability that the test is intended to measure (hereafter it is called the ability proxy). This is because a test is usually constructed so that correct responses to individual items have high correlation with the ability proxy, and overall difference in the proportions of correct responses between the reference and target groups of examinees may be simply due to different average ability levels. Thus, factors that may lead to item bias should be examined only after controlling for the effect of ability proxy on item responses. The same principle may be applied to DDF analysis as well. As Green et al. (1989) stated, item analysis often reveals that "different distractors are chosen by persons of different ability levels" (p. 148). Thus, group comparisons should be made conditionally on ability in DDF analysis as well.

While the literature of DDF analysis is not very abundant (e.g., Abedi et al., 2007b; Banks, 2006; Green et al., 1989; Marshall, 1983; Middleton & Laitusis, 2007), there is some methodological variation such as the log-linear approach (Green et al., 1989), the logistic regression approach (Abedi et al., 2007b), and the standardization approach (Dorans, Schmit, & Bleinstein, 1992). The log-linear and logistic regression approaches specifically focused on DDF using incorrect responses only (e.g., Abedi et al., 2007b; Green et al., 1989). These approaches are appealing, because they enable simple interpretations, and including the correct response in the model sometimes overwhelms subtle DDF (Green et al., 1989). However, DIF and DDF are not exclusive to each other. An item exhibiting DIF is likely to show DDF for one or more other response options, because response rates are dependent on each other (i.e., a larger proportion for a response option necessarily implies a smaller proportion for at least one other response option). Moreover, DDF is more serious if it triggers DIF, that is, examinees in one group are attracted by a distractor more easily than those in another group and, as a result, they are less likely to choose the correct response.

The standardization approach is more advantageous in this respect; it examines behaviors of all response options simultaneously by calculating a weighted average of the difference of conditional probabilities of each response option between groups (the standardized DIF index), with the average taken with respect to ability levels (Dorans et al., 1992). A possible drawback is that it is a purely descriptive approach, and does not offer formal statistical criteria for the determination of DDF. Also, the standardized DIF index is based on signed differences of response choice probabilities, so the method may fail to detect DDF if positive differences in some ability levels can be canceled out by negative differences in other ability levels.

Given the previously mentioned considerations, this study employed multinomial logistic regression as an extension of Abedi et al.'s (2007b) binary logistic regression approach. By multinomial logistic regression, we model the probabilities of all response options in each test item as functions of the ability proxy at the same time. This lets us draw a whole picture of the behavior of all response options, whether correct or incorrect, for each item; that is, we can examine DIF and DDF simultaneously. In contrast, the binary logistic regression only focuses on one particular distractor for each item. Although it may be more sensitive to DDF of that particular distractor, it may overlook DDF for other distractors and DIF for the correct response. Also, the multinomial regression approach can be regarded as a parametric version of the standardization approach. Unlike the standardization approach, however, multinomial logistic regression enables us to test DIF/DDF statistically, quantify the amount of DIF/DDF both overall and for individual response options for each

item, and visualize item characteristics for detailed examination.

In addition to DIF and DDF, the frequency of omitted responses is another indicator of item bias. Although several explanations are possible, such as fatigue and limited testing time, a general explanation is that cognitive demands of test items overwhelm examinees' motivation and ability to complete a test (e.g., Stone, Stone, & Gueutal, 1990, for general cognitive ability testing). Omitted responses are usually taken as incorrect responses, or at least not counted as correct, leading to underestimation of total test scores and lower precision of those scores. Thus, if examinees in one group tend to omit an item more often than examinees in other groups (differential omission frequency; DOF), resulting test scores may not be equivalent across groups and the cause should be investigated.

It is generally difficult to speculate what factors are responsible for DIF, DDF, and DOF only from item response data, but locations of items that exhibit them could provide a clue. Abedi et al. (2007b) found that items located in the second half of the test showed more DDF than in the first half, and that response patterns of students with disabilities appeared to be more random than those of students without disabilities. A possible interpretation could be that students with disabilities required more time to complete the test or they became more fatigued or frustrated as the test progressed.

This study also differs from earlier test item analysis studies in that data in this study permit analyses by discrete disability categories to examine DIF, DDF, and DOF with respect to disability status. Although several DIF studies found that patterns of DIF differed by disability type (e.g., Bennett, Rock, & Kaplan, 1987; Johnstone, Thompson, Moen, Bolt, & Kato, 2005), previous DDF studies have been limited to findings regarding the entire group of students with disabilities (Abedi et al., 2007b) or students with one particular type of disability (learning disability; Middleton & Laitusis, 2007). The data in this study let us see whether different findings are obtained for students with different kinds of disabilities with respect to DIF, DDF, and DOF.

The purpose of this study is to flag potentially biased items for students with various kinds of disabilities, using data from state reading assessments. Although DIF, DDF, and DOF do not immediately imply item bias, results provide clues to review characteristics of extant items that could cause biases and to design test items that are equally accessible to all students. According to these considerations, this study investigates the following research questions:

1. Do items in the reading tests exhibit DIF/DDF for students with disabilities?
2. Are the amounts of DIF/DDF affected by item locations? Is there any systematic difference in terms of the pattern of DIF/DDF across different disability groups?
3. Does occurrence of omitted responses differ for students with disabilities? Is it affected by item locations?
4. Are the above differences, if any, different for third and fifth graders?

## Method

### *Participants*

The Minnesota Department of Education provided two data sets that were analyzed in this study. Both were part of Minnesota's spring 2003 administration of statewide reading assessments. One data set is for third graders, and the other data set for fifth graders. These data sets were analyzed separately.

The grade 3 data contained demographic information and item responses from 57,071 students who had valid test scores on the 2003 state reading test. The upper portion of Table 1 shows the makeup of the third grade data by disability category, gender, and ethnicity. The number of students without disabilities was 50,290. Speech/language impairment (SLI; $N = 2{,}436$), learning disabilities (LD; $N = 2{,}242$), and emotional behavior disorders (EBD; $N = 768$) constituted the three largest disability groups, while each of the other disability categories had less than 1% of all students. The population percentages of disabilities reported for this state for the 26th Annual Report to Congress on the Implementation of the Individuals with Disabilities Education Act (the IDEA report; U.S. Department of Education, 2006) showed that 11% of all students of age 6–17 in the state in the fall of 2002 had disabilities and that 1.8%, 4.2%, and 1.9% of all students had SLI, LD, and EBD, respectively. We did not request explanations for the discrepancies between the population percentages and the proportions of students with SLI, LD, and EBD who took this test.

The grade 5 data contained demographic information and item responses from 60,364 students who had valid test scores on the 2003 state reading test.

**Table 1. Demographic Information of the Subjects**

| Disability Category[a] | Total | Gender | | Ethnicity[b] | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Male | Female | AI | AS | HP | AA | CC |
| Grade 3 ($N = 57{,}071$) | | | | | | | | |
| No disability | 88.1 | 43.0 | 45.1 | 1.6 | 5.3 | 4.4 | 6.6 | 70.2 |
| SLI | 4.3 | 2.8 | 1.5 | .1 | .2 | .1 | .3 | 3.6 |
| LD | 3.9 | 2.6 | 1.3 | .1 | .2 | .2 | .4 | 3.0 |
| EBD | 1.3 | 1.0 | .3 | .1 | — | — | .3 | 1.0 |
| Other | 2.4 | 1.7 | .7 | .1 | — | .2 | .2 | 1.9 |
| Total | 100.0 | 51.1 | 48.9 | 2.0 | 5.7 | 4.9 | 7.8 | 79.7 |
| Grade 5 ($N = 60{,}364$) | | | | | | | | |
| No disability | 86.4 | 42.0 | 44.5 | 1.7 | 5.1 | 3.5 | 6.2 | 69.9 |
| SLI | 3.6 | 2.3 | 1.4 | .1 | .2 | .1 | .3 | 3.0 |
| LD | 4.8 | 3.2 | 1.6 | .2 | .2 | .3 | .6 | 3.6 |
| EBD | 2.0 | 1.6 | .4 | .1 | — | .1 | .4 | 1.4 |
| Other | 3.2 | 2.0 | 1.0 | .1 | — | .1 | .2 | 2.6 |
| Total | 100.0 | 51.1 | 48.9 | 2.2 | 5.5 | 4.1 | 7.7 | 80.5 |

Note. Each cell shows the percent to the entire sample for each grade level. "—" indicates that the percent is less than .1%.

[a]SLI = speech/language impairment; LD = learning disability; EBD = emotional/behavioral disorder.

[b]AI = American Indian; AS = Asian; HP = Hispanic; AA = African American; CC = Caucasian.

The lower portion of Table 1 shows the make up of the fifth grade data by disability category, gender, and ethnicity. The number of students without disabilities was 52,177. SLI ($N = 2{,}201$), LD ($N = 2{,}889$), and EBD ($N = 1{,}192$) constituted the three largest disability groups. As with the grade 3 test data, these disability percentages were somewhat different from the population percentages reported in the 2006 IDEA report. That report showed that 11% of all students of age 6–17 in the state in the fall of 2002 had disabilities and that 1.8%, 4.2%, and 1.9% of all students had SLI, LD, and EBD, respectively.

*Reading Assessments*

The reading assessments were intended to measure two broad aspects of reading ability: (a) literal comprehension and (b) interpretation and evaluation. Literal comprehension consisted of the following skills: (a) identify main ideas and some supporting details, (b) retell main events or ideas in sequence, (c) pronounce new words using phonic skills, (d) read aloud fluently with appropriate expression, (e) demonstrate appropriate techniques for learning new vocabulary, and (f) interpret presentations of data. Interpretation and evaluation included the following skills: (a) understand ideas not explicitly stated; (b) make predictions based on information in the text; (c) draw conclusions based on information in the text; (d) compare and contrast elements of the text; (e) distinguish facts from opinions; and (f) summarize ideas and identify tone in persuasive, fictional, and documentary presentations.

Both grade 3 and grade 5 reading assessments consisted of 49 items, out of which there were 46 multiple-choice items and 3 constructed-response items. All multiple-choice items have four response options, denoted by A, B, C, and D. Only the multiple-choice items were analyzed in this study. There were additional field-testing items in each test booklet. They were neither scored nor examined in this study.

Both grade 3 and grade 5 assessments were administered in four sessions. The fourth session included the additional field-testing items only, so this study focused on the first three sessions in which the 46 multiple-choice items were administered. In each session, two or three passages were presented and students answered five to seven items for each passage. The first three columns in Tables 2 and 3 show how passages and items were organized across sessions in the grade 3 and 5 assessments, respectively. All passages were presented in the same order for all students. Testing started with the first session and progressed by session; students were allowed to go back and forth among items within each session, but not to move on to the next session until they were told to do so. Sessions were administered consecutively as one large testing session or separately with some interval between sessions (e.g., two sessions in one day and other two sessions in the next day), depending on school's (or teacher's) convenience. In both cases, however, students were allowed to take a 3-minute break every 30 minutes of testing. The test was not timed, and each student had enough time to attempt to answer each item.

For the grade 3 assessment, the overall proportion correct score for the 46 multiple-choice items was .78 with standard deviation .19, and Cronbach's alpha was .92. Thus, the assessment was highly internally consistent and somewhat easy. At the passage level, mean proportion correct score ranged from .67 to .85. The grade 5 assessment showed similar characteristics. The overall mean was .80 with standard deviation .18, and the alpha was .91. Passage means ranged from .71 to .87. The fourth column of Tables 2 and 3 shows the mean proportion correct score by session for grades 3 and 5, respectively.

Scale scores were also provided in the data (mean 1,513 and standard deviation 202.4 for the grade 3 assessment; mean 1,567 and standard deviation 226.7 for the grade 5 assessment). They were standardized to have mean 0 and standard deviation 1 for each assessment and used as an ability proxy in the DIF/DDF and DOF analyses.

*Differential Item/Distractor Functioning Analysis*

In this study, DIF and DDF were examined at the same time by a multi-step multinomial logistic regression analysis. The multinomial logistic regression analysis estimates a response characteristic curve (RCC) for each response option in an item. An RCC represents the probability of choosing the corresponding response option as a function of the ability proxy. Suppose that an item has $K$ response options. The RCC for response option $k$ takes the form of the multinomial logistic function

$$p_k(z) = \frac{\exp(a_k + b_k z)}{\sum_{l=1}^{K} \exp(a_l + b_l z)}, \quad k = 1, \ldots, K,$$

where $z$ denotes the ability proxy (for which the standardized scale score is

**Table 2. Composition of the Grade 3 Reading Assessment and the Number of Items That Exhibited DIF/DDF and DOF by Disability Category**

| Passage | Item | Number of Items | Proportion Correct | DIF/DDF | | | DOF | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | SLI | LD | EBD | SLI | LD | EBD |
| | | | Session 1 | | | | | | |
| 1 | 1–5 | 5 | .85 | 0 | 0 | 0 | 0 | 1 | 0 |
| 2 | 6–10 | 5 | .74 | 0 | 1 | 0 | 0 | 0 | 0 |
| 3 | 11–17 | 6 (7) | .73 | 0 | 0 | 0 | 0 | 5 | 0 |
| | | | Session 2 | | | | | | |
| 4 | 18–23 | 6 | .67 | 0 | 2 | 0 | 0 | 6 | 0 |
| 5 | 24–29 | 6 | .78 | 0 | 3 | 0 | 3 | 5 | 0 |
| 6 | 30–35 | 5 (6) | .83 | 0 | 0 | 0 | 3 | 5 | 0 |
| | | | Session 3 | | | | | | |
| 7 | 36–42 | 7 | .81 | 0 | 1 | 0 | 0 | 0 | 3 |
| 8 | 43–49 | 6 (7) | .86 | 0 | 1 | 0 | 1 | 3 | 0 |
| Total | | 46 (49) | .78 | 0 | 8 | 0 | 7 | 24 | 3 |

Note. SLI = speech/language impairment; LD = learning disability; EBD = emotional/behavioral disorder. Items 17, 35, and 49 are constructed-response items and were not used in the analyses. The third column indicates the numbers of items that are used in the analyses; the actual numbers of items in the test are shown in parentheses.

**Table 3. Composition of the Grade 5 Reading Assessment and the Number of Items That Exhibited DIF/DDF and DOF by Disability Category**

| Passage | Item | Number of Items | Proportion Correct | DIF/DDF | | | DOF | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | SLI | LD | EBD | SLI | LD | EBD |
| | | | Session 1 | | | | | | |
| 1 | 1–7 | 7 | .75 | 0 | 0 | 0 | 0 | 7 | 0 |
| 2 | 8–13 | 6 | .87 | 0 | 1 | 0 | 0 | 5 | 0 |
| 3 | 14–19 | 5 (6) | .82 | 0 | 0 | 0 | 0 | 5 | 2 |
| | | | Session 2 | | | | | | |
| 4 | 21–26 | 6 | .78 | 0 | 2 | 0 | 0 | 6 | 0 |
| 5 | 27–33 | 7 | .83 | 0 | 0 | 0 | 1 | 7 | 0 |
| 6 | 34–38 | 4 (5) | .78 | 0 | 1 | 0 | 0 | 3 | 2 |
| | | | Session 3 | | | | | | |
| 7 | 39–44 | 6 | .82 | 0 | 0 | 0 | 0 | 6 | 3 |
| 8 | 45–50 | 5 (6) | .71 | 0 | 1 | 0 | 0 | 4 | 4 |
| Total | | 46 (49) | .80 | 0 | 5 | 0 | 1 | 40 | 9 |

Note. SLI = speech/language impairment; LD = learning disability; EBD = emotional/behavioral disorder. Items 19, 38, and 50 are constructed-response items and item 20 is not a test question. These items were not used in the analyses. The third column indicates the numbers of items that are used in the analyses; the actual numbers of items in the test are shown in parentheses.

used in this study), and $a_k$s and $b_k$s are regression coefficients that determine the shape of the RCC. Relative sizes of $b_k$s determine the order of RCCs; the order of $b_k$s corresponds to where the corresponding RCCs have their peaks. The correct response option usually has the largest value. Relative sizes of $a_k$s represent the popularity of the response options; larger $a_k$s mean the corresponding response options are chosen more frequently (see Thissen, Steinberg, & Fitzpatrick, 1989).

The multi-step analysis was conducted with two models: ability (Model 1) and ability plus group (Model 2). Model 1 assumed no group differences with respect to RCCs (or equivalently, regression coefficients); the same RCCs were fitted for all examinee groups. In Model 2, RCCs were allowed to vary between students without disabilities (the reference group) and students with disabilities (the target group). One particular disability group (SLI, LD, or EBD) was compared to the reference group at a time, though it was also possible to include all four groups in one model. Thus, Model 2 is equivalent to estimating two sets of regression coefficients, one for the reference group and the other for the target group. Then, pseudo $R^2$ (Nagelkerke's $R^2$; Nagelkerke, 1991), which approximates the variance explained by the ability proxy and is an analogue to $R^2$ in the normal linear regression, was calculated for each of Models 1 and 2. Models 1 and 2 were compared by the likelihood ratio test and also by the difference between the pseudo $R^2$s from the two models.

This procedure is similar to the multi-step procedure described by Zumbo (1999; also see Swaminathan & Rogers, 1990). The common feature is to fit a model with common RCCs for all groups and another model with RCCs differing by group, and then compare the two models by the pseudo $R^2$ difference. The main difference is that Zumbo's (1999) procedure was based on item response theory models (i.e., the latent ability score is used as an ability proxy and must be estimated as a part of the model) and originally intended for DIF. Item response theory requires fitting all items at the same time, possibly allowing RCCs of one of the items to differ by group at one time to detect DIF for that particular item. This implies rescaling all items at each time because regression coefficients of all items are likely affected by constraints imposed on coefficients of one item. In contrast, the procedure of the current study uses the scaled score as the ability proxy, which is not affected by models, and it also allows separate analysis by item.

Items for which the likelihood ratio test was significant at $\alpha = .01$ and the pseudo $R^2$ difference was no smaller than .003 were flagged for DIF/DDF. The pseudo $R^2$ difference is a measure of effect size, and was used in addition to the significance test to avoid picking up trivially significant cases due to the large sample sizes. For continuity with prior research, we adopted the threshold value of .003 following Abedi et al. (2007b). In order to examine the effect of item location on DIF/DDF, pseudo $R^2$ differences were plotted against item locations for each of the disability categories (SLI, LD, and EBD). The regression coefficients and the pseudo $R^2$ in the multinomial logistic regression model were estimated by the NOMREG command in SPSS.

Flagged items were then subjected to further analysis to examine what contributed to the observed differential functioning—DIF, DDF, or both. For this purpose, the mean absolute difference (MAD) was computed between the corresponding RCCs for the target and reference groups. Technically, MAD for response option $k$ in an item is defined as

$$MAD_k = \frac{1}{N}\sum_{i=1}^{N}\left|p_{k,0}(z_i) - p_{k,1}(z_i)\right|,$$

where $z_i$ is the ability proxy value of examinee $i$, $N$ is the total sample size, and $p_{k,0}(z)$ and $p_{k,1}(z)$ denote estimated RCCs for response option $k$ for the reference and target groups, respectively. MAD represents the average difference between two RCCs. Response options with larger MADs were considered to contribute to the observed differential functioning. If the correct response option exhibits the largest MAD, then the differential functioning is mainly due to DIF, while distractors yielding larger MADs contribute to DDF.

As a final check, we plotted RCCs of flagged items that exhibited large MADs. Visual inspection of RCCs allows for finer interpretation of observed DIF/DDF. Especially, we can observe the direction of DIF/DDF and determine which response option favors which group of students at a particular ability level.

### *Differential Omission Frequency Analysis*

Similar steps to those in the DIF/DDF analysis were taken for the DOF analysis. The dependent variable was a binary indicator: whether the response was omitted or not. Accordingly, the

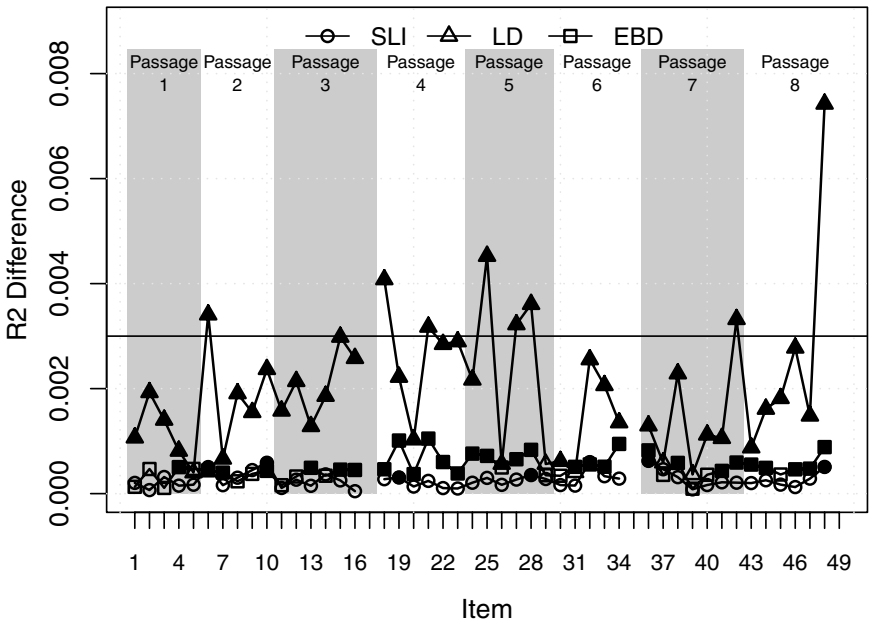


FIGURE 1. DIF/DDF $R^2$ difference between students with and without disabilities in grade 3. No $R^2$ is available for items 17, 35, and 49 because these are constructed-response items and were not analyzed.

multinomial logistic regression in the DIF/DDF analysis was simplified to the ordinary binary logistic regression, by which an omitted response curve (ORC) was estimated for each item for each group of examinees. Since ability may be responsible for occurrence of omitted responses (for example, examinees of lower ability may skip items more easily; Stone et al., 1990), the ability proxy was entered in the model (Model 1). Then, another model, in which a main effect of the grouping variable and its interaction with the ability proxy are added, was fit (Model 2). Pseudo $R^2$ differences were plotted against item locations. Items were flagged if they were significant at the .01 level on the likelihood ratio test between Models 1 and 2 and yielded pseudo $R^2$ difference no smaller than .003.

## Results

*Grade 3 DIF/DDF Analysis*

$R^2$ differences were plotted against item location for SLI, LD, and EBD (see Figure 1). In the plot, filled marks indicate that the difference between students without disabilities and those with disabilities was significant. Thus, items for which the corresponding marks were filled and above the threshold of .003 (also shown in the plot as a horizontal line) are those flagged for DIF/DDF. Although seven items were significant for SLI and 28 items for EBD, none of the corresponding $R^2$ differences exceeded .003, indicating that there was no substantive DIF/DDF for these groups. In contrast, almost all items were significant for LD, and eight items yielded $R^2$ differences greater than .003 (items 6, 18, 21, 25, 27, 28, 42, and 48). Overall, the LD group indicated much larger DIF/DDF than the SLI and EBD groups for all items.

Columns 5 through 7 in Table 2 summarize the number of items that exhibited DIF/DDF by session, passage, and disability category. Most of the DIF/DDF items were found in the second session (passages 4 and 5) for students with LD, while the largest DIF/DDF was found for item 48, which was in the third session.

MADs were computed for the flagged items for the LD group (see Table 4). For two items (items 6 and 42), the largest MAD was found for one of the distractors, indicating DDF. For the other six items, the largest MAD was found for the correct response option. For two of these items (items 25 and 48), however, there were distractors for which MADs were only slightly smaller than those of the correct response option. RCCs for these four items (items 6, 25, 42, and 48) are examined here (see Figure 2).

Figure 2(a) depicts RCCs for item 6, where distractor C had the largest MAD. For both groups of students, a fair number of students chose distractor C regardless of their ability levels. However, students without disabilities tended to choose distractor C more often than students with LD, resulting in the lower correct response rate overall. In other words, students without disabilities were distracted more easily by response option C than students with LD. In fact, over all ability levels, response C, was the most popular distractor among students without disabilities (27%), while it was least popular among students with LD (16%); students with LD chose response A most (27%) among the distractors. However, when looking at lower ability students, students without disabilities tended to choose distractor A more often than students with LD.

Figure 2(b) shows RCCs for item 25, for which the correct response (B) had the largest MAD; distractor D also had a large MAD. Figure 2(b) indicates that

**Table 4. Mean Absolute Difference for the Flagged Items for Students with LD in Grade 3**

| Item | Response Option A | B | C | D | $R^2$ Difference |
|---|---|---|---|---|---|
| 6 | .052 | .068* | *.100* | .005 | .0034 |
| 18 | *.104** | .071 | .024 | .019 | .0041 |
| 21 | *.072** | .038 | .037 | .039 | .0032 |
| 25 | .017 | *.127** | .011 | .102 | .0045 |
| 27 | .010 | .024 | .035 | *.067** | .0032 |
| 28 | *.150** | .047 | .062 | .044 | .0036 |
| 42 | .037 | .041* | .012 | *.051* | .0033 |
| 48 | .065 | .008 | *.073** | .009 | .0074 |

Note. The correct response options are indicated by an asterisk (*). The largest MAD for each item is shown in italic.

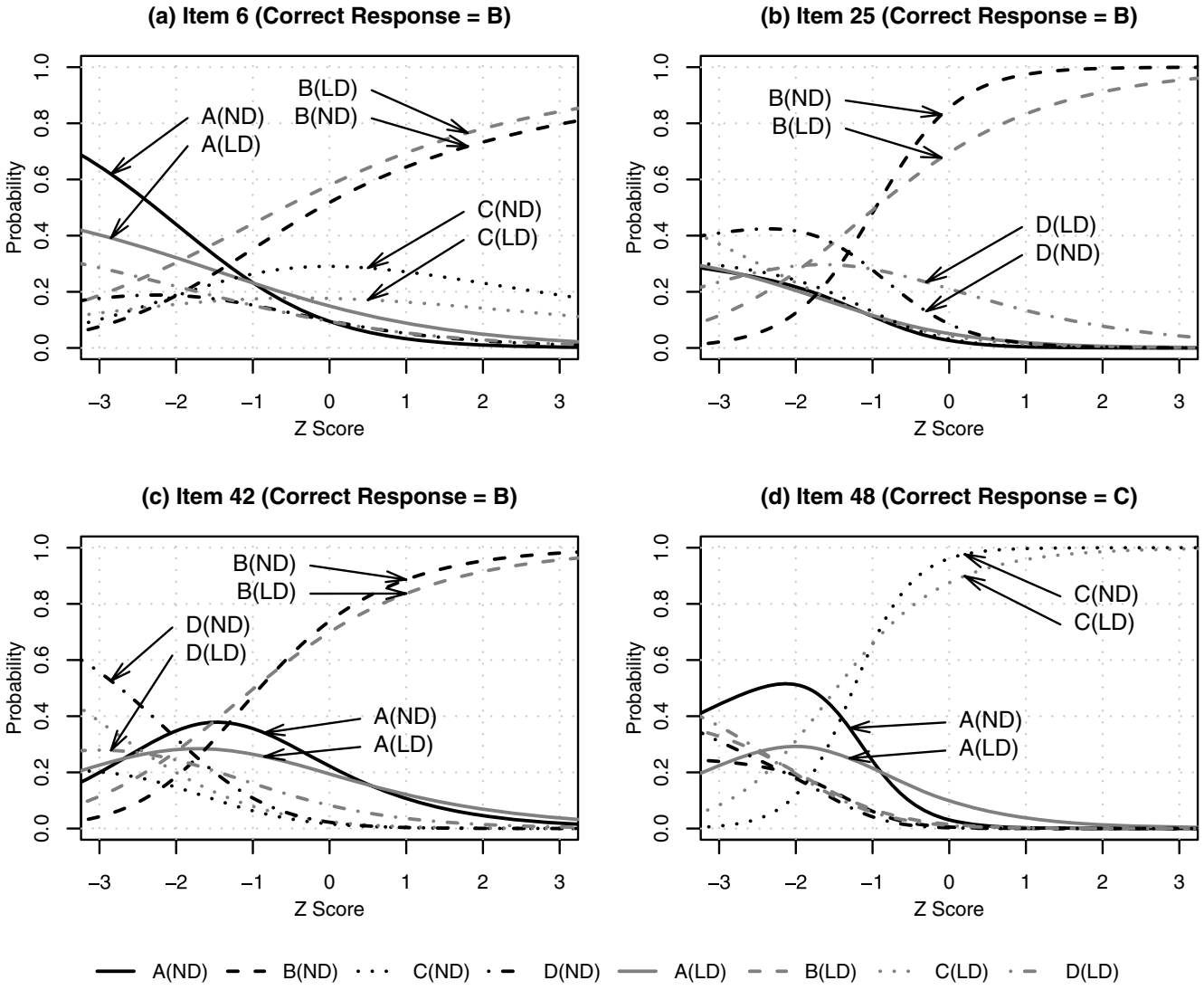


FIGURE 2. Comparison of RCCs between students without disabilities and students with LD in grade 3. RCCs for students without disabilities are shown in black, and those for students with LD in gray. ND = students without disabilities; LD = students with LD.

students with LD of average to high ability chose distractor D instead of the correct response more often than comparable ability students without disabilities. This tendency is reversed in the low ability range, where students with LD were more likely to choose the correct response and less likely to choose distractor D than students without disabilities. As a result, discrimination of this item was lower for students with LD than for students without disabilities.

For item 42, distractor D had the largest MAD, followed by the correct response (B) and distractor A as shown in Figure 2(c). Overall, choice of distractors by students with LD was less reflective of ability than for students without disabilities; this case is indicated by the flatter RCCs. Students without disabilities of very low ability ($z < -2$) tended to choose distractor D, and those of slightly higher ability ($-2 < z < -1.5$) preferred distractor A. In contrast, it was less clear what specific response students with LD in the same range of ability tended to choose.

Figure 2(d) depicts RCCs for item 48. The $R^2$ difference for this item for LD was very high relative to the others (see Figure 1). This large difference reflects substantially different RCCs for response options A and C. Although the correct response (C) produced the largest MAD, distractor A had a comparable MAD, as is indicated by its RCCs, which behaved very differently for the two groups of students. For the reference group, students of low ability ($z < -1.5$) tended to choose distractor A most often (with probability greater than .4), while for students with LD, the probabilities of choosing the distractors were more similar to each other. In the moderate to high ability range, students with LD were slightly more likely to choose distractor A instead of the correct response than students without disabilities.

### *Grade 3 DOF Analysis*

Overall, omitted response rates were very small for all items, ranging from .2% to 1.2% of the entire sample. Figure 3 indicates omitted response rates for the entire sample. Clearly, the omitted response rate increases as examinees go forward within each passage. Whenever examinees moved to the next passage, however, the rate seemed to be "reset." We employed the DOF analysis to determine whether this pattern was

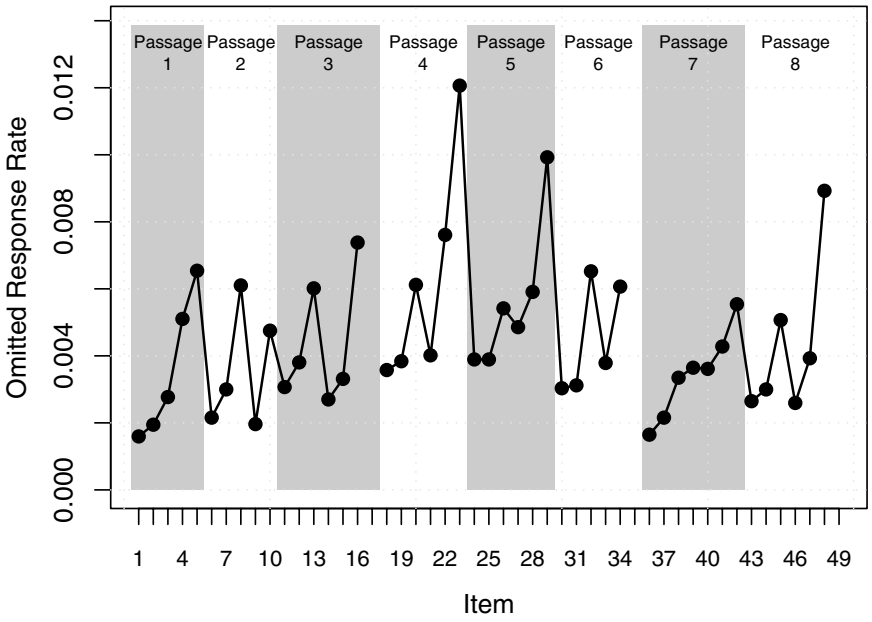


FIGURE 3. Omitted response rates by item for grade 3.

different for different disability groups if it was compared conditional on the ability proxy. In the logistic regression analyses, coefficients for the ability proxy were negative and significant for all items. Thus, omitted responses become less frequent as ability increased.

$R^2$ differences were plotted against item location for SLI, LD, and EBD (see Figure 4). Seven items exhibited DOFs for SLI, twenty-five items for LD, and three items for EBD. Students in the three disability categories showed different omitted response patterns in terms of item locations. For SLI and LD, most of such items were located in the second session, while for EBD, all such items were found in the third session (detailed counts are shown in the last three columns in Table 2). As seen in the DIF/DDF analysis, the LD group showed much larger discrepancies.

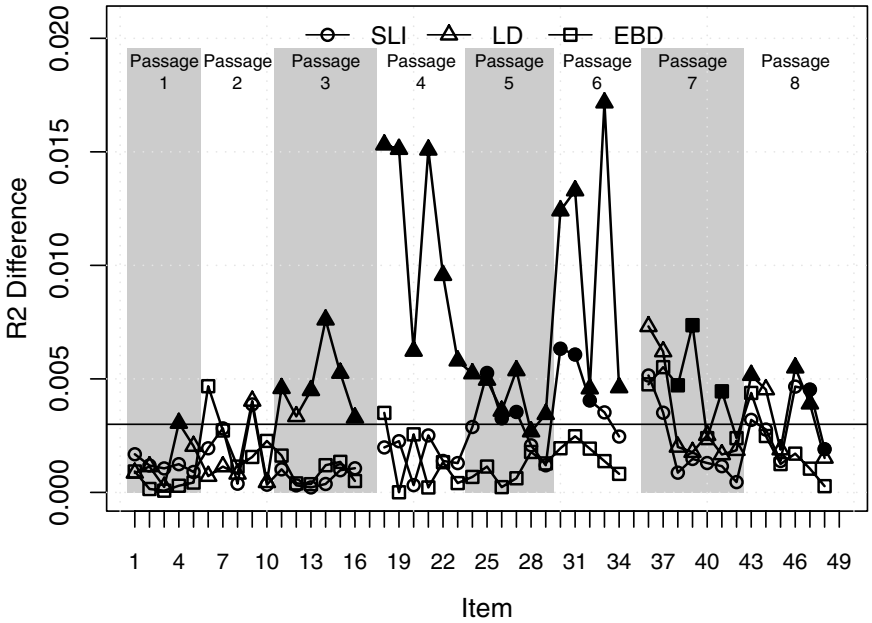


FIGURE 4. DOF $R^2$ difference between students with and without disabilities in grade 3. No $R^2$ is available for items 17, 35, and 49 because these are constructed-response items and were not analyzed.

We examined coefficients (log odds ratios; LORs) for the main effect of disability status and its interaction with the ability proxy obtained in Model 2. Most of the main effect and interaction LORs were negative irrespective of disability categories. This indicates that occurrence of an omitted response is more sensitive to ability for students with disabilities than for students without disabilities. In other words, omission rates for students with disabilities increased more abruptly than they did for students without disabilities as ability became lower. However, these results should be taken with caution, because omitted responses were quite rare in the entire sample; as a result, estimated LORs may not be reliable.

### *Grade 5 DIF/DDF Analysis*

$R^2$ differences were plotted against item location for SLI, LD, and EBD (see Figure 5). Five items were significant for SLI and 29 items for EBD, but none of the corresponding $R^2$ differences exceeded .003, indicating that there was no substantive DIF/DDF for these groups. In contrast, 43 out of 46 items were significant for LD, and five items yielded $R^2$ differences greater than .003 (items 9, 22, 24, 34, and 49). Overall, the LD group indicated much larger DIF/DDF than SLI and EBD for all items as seen for the grade 3 data.

Columns 5 through 7 in Table 3 summarize the number of items that exhibited DIF/DDF by session, passage, and disability category. Most of the items that indicated DIF/DDF were located in the second session for students with LD. This is the same tendency as found for the grade 3 data.

MADs were computed for the flagged items for the LD group (see Table 5). For all of the five items, the largest MADs were found for the correct response options. As seen in the grade 3 analysis, however, there were several items for which distractors yielded MADs comparable to those of the correct responses (items 9, 22, 24, and 49). RCCs for these items are displayed in Figure 6, in which response options that exhibited discrepancies are indicated by arrows. Similar interpretations to those of the grade 3 results also apply here.

### *Grade 5 DOF Analysis*

Omission rates were very small for all items, ranging from .1% to .6% of the

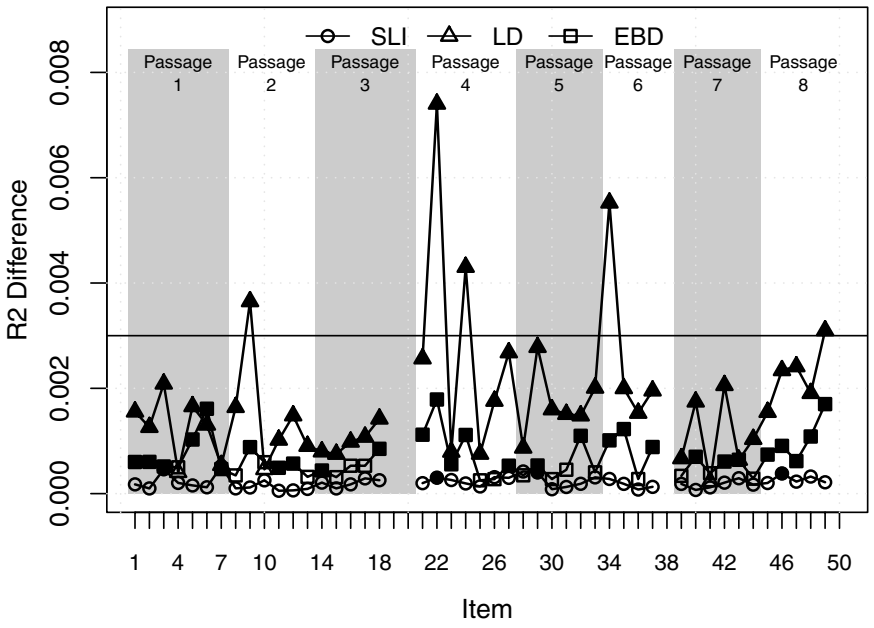


FIGURE 5. DIF/DDF $R^2$ difference between students with and without disabilities in grade 5. No $R^2$ is available for items 19, 38, and 50 because these are constructed-response items and were not analyzed. Item 20 is not a test item and was also excluded.

entire sample. Figure 7 shows that within each session, the omission rate gradually increased and suddenly jumped up for the last two or three items. This is a pattern quite different from the grade 3 result. We used the DOF analysis to determine whether this pattern was different for different disability groups if it was compared conditional on the ability proxy. In the logistic regression analyses, coefficients for the ability proxy were negative and significant for all items. Thus, omitted responses became less frequent as ability increased.

Figure 8 indicates very contrasting patterns of DOF for the SLI, LD, and EBD groups, respectively. For students with SLI, only one item (item 32) was significant. For students with LD, most items were significant and differential behavior of items tended to decrease within each session. For students with EBD, most of the significant items were found in the last session. Distribution of items that exhibited DOF is shown in the last three columns in Table 3. Overall, the LD group indicated much larger differences than the other disability groups. For students with SLI, item 32 had an extremely small omission rate, resulting in the large $R^2$ difference as can be seen in Figure 8. For students with LD, item 22 showed the largest $R^2$ difference; regression coefficients showed that occurrence of an omitted response was more frequent and more dependent on ability for students with LD than for those without disabilities. This item also showed the largest amount of DIF/DDF for students with LD (see Figure 5). The same kind of difference was found for item 1, which produced the second largest $R^2$ difference.

We examined LORs for the main effect of disability status and its interaction with the ability proxy, and found a similar tendency to the grade 3 results. Most of the main effect and interaction LORs were negative, indicating that omission rates for students with disabilities increased more abruptly as ability became lower than for the students without disabilities. Again, these results should be taken with caution because of the small number of omissions.

**Table 5. Mean Absolute Difference for the Flagged Items for Students with LD in Grade 5**

| Item | Response Option A | B | C | D | $R^2$ Difference |
|---|---|---|---|---|---|
| 9 | .022 | .012 | .006 | *.030** | .0037 |
| 22 | .096 | *.109** | .016 | .016 | .0074 |
| 24 | .053 | .018 | *.059** | .024 | .0043 |
| 34 | .008 | *.063** | .025 | .031 | .0055 |
| 49 | .041 | .021 | *.055** | .009 | .0031 |

Note. The correct response options are indicated by an asterisk (*). The largest MAD for each item is shown in italic.

## Summary and Conclusions

This study examined DIF, DDF, and DOF for items on third and fifth grade statewide reading tests for three disability groupings: students with SLI, LD, and EBD. Although the percentages of students with SLI, LD, and EBD in our sample did not match the federally reported population percentages, we did not assume that this represented a challenge to the study findings. There could be any number of reasons for differences between population percentages and the proportions of students participating in a state's regular assessment. For example, higher proportions of students with SLI participating in the assessment in this study might be the result of IEP team decisions that more students from other categories (e.g., LD and EBD) would participate in alternate assessments. Nevertheless, future studies would do well to examine the extent to which the population is represented in the regular assessment and whether any discrepancies seem to be related to findings of DIF, DDF, or DOF.

Due to the large number of records analyzed, many items showed statistically significant DIF/DDF results for all three disability groups. When the $R^2$ difference criterion used in previous such studies was applied, only a small number of items were judged to exhibit meaningful DIF/DDF. The traditional "DIF only" approach would likely detect DIF as the multinomial logistic regression did in this study, but the latter provided additional information about the differential performance of distractors.

Items that exhibited DIF and DDF were found only for those students with LD. In contrast to previous research

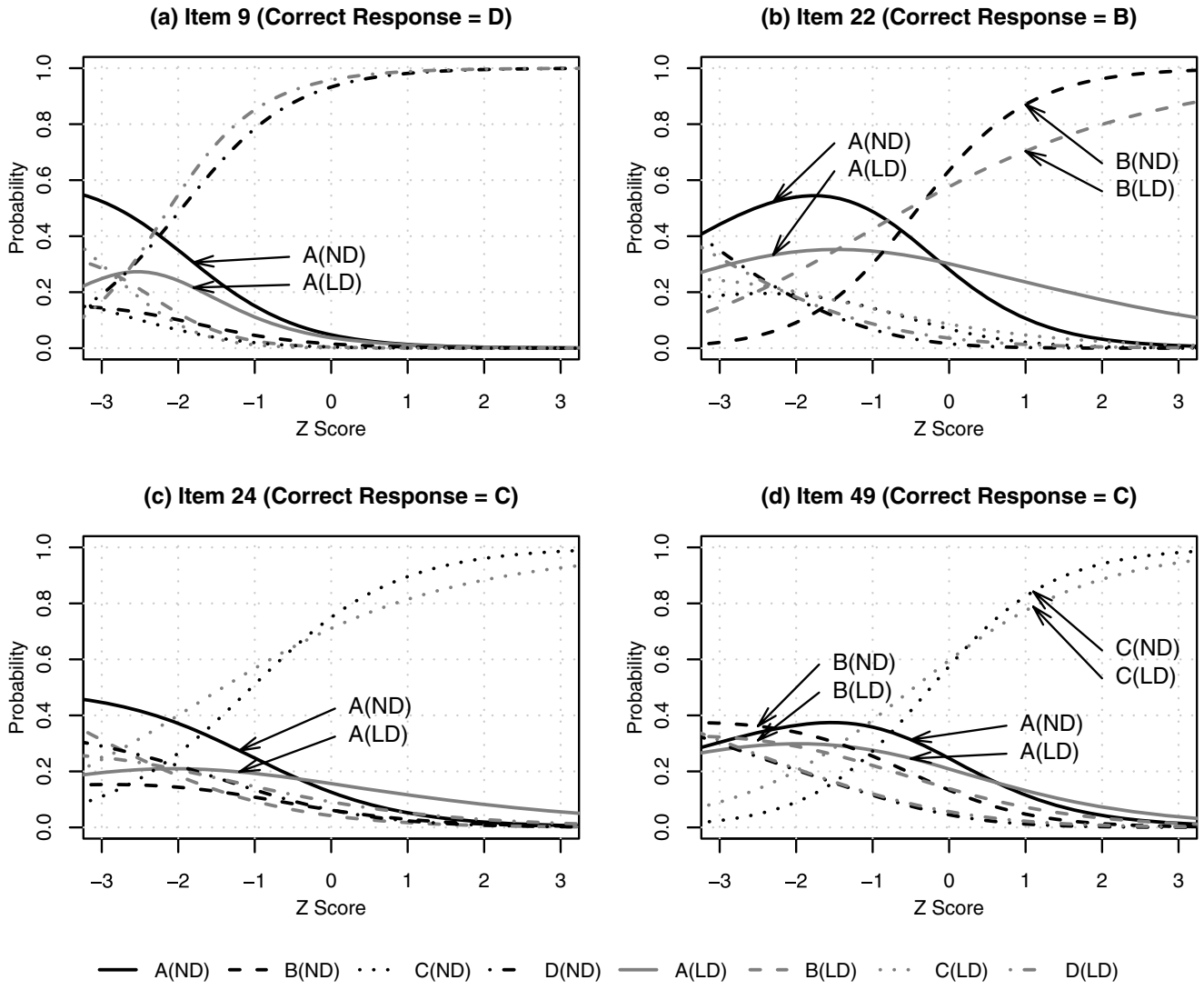


FIGURE 6. Comparison of RCCs between students without disabilities and students with LD in grade 5. RCCs for students without disabilities are shown in black, and those for students with LD in gray. ND = students without disabilities; LD = students with LD.

with undifferentiated students with disabilities (Abedi, Leon, & Kao, 2007a, 2007b), our observations of DIF/DDF for students with LD did not suggest a general trend for DIF/DDF to increase as items were located later in the test. This study found that most of the items that exhibited DIF/DDF (and also DOF) for students with LD were located in the second session for both grade 3 and 5 tests, but we speculate that this was caused by characteristics specific to items rather than their locations, because, by the test administration procedure, the second session could be located in any place in one testing occasion (i.e., it can be the last session on day 1 or the first session on day 2 if the test is administered over 2 days, or it comes in the middle if the test is administered in 1 day).

The group sizes were sharply unequal between students with and without disabilities, and this might affect the values of pseudo $R^2$ differences. Nagelkerke's $R^2$ is based on the likelihood function for the entire sample, to which each student contributes equally. In Model 1, the estimated RCCs common to all student groups were almost the same as those for students without disabilities in Model 2 because of the disproportionately large size of the no disability group. In other words, the contribution of students without disabilities to the likelihood was almost the same in both models. Model 2 provided larger likelihood (and thus larger $R^2$) than Model 1 by allowing different RCCs for students without disabilities, but such improvement was likely overwhelmed by the almost unchanged contribution of students without disabilities. As a result, pseudo $R^2$ differences tended to be smaller in this study than they would have been if group sizes had been approximately equal. This speculation suggests that care should be taken when setting a threshold on the $R^2$ difference. In this study, however, the threshold value of .003 worked just as well for the given samples, because the items selected by this criterion showed substantive DIF/DDF as confirmed by MADs and visual inspection of RCCs. We recommend using multiple measures to judge DIF/DDF.

As with DIF/DDF, many items showed statistically significant DOF for all three disability categories. Applying the $R^2$ difference criterion used with DIF/DDF made little difference in these findings. Nevertheless, the large

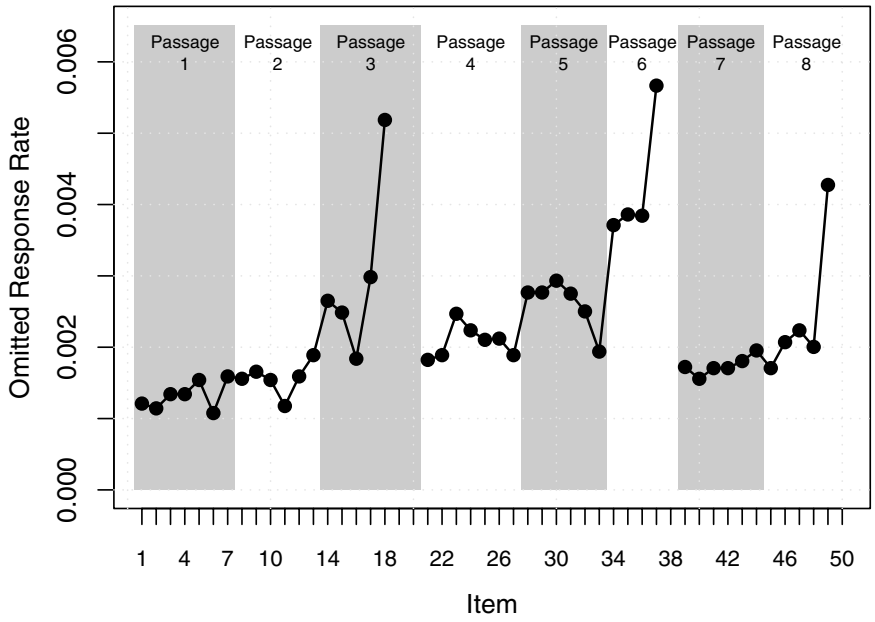


FIGURE 7. Omitted response rates by item for grade 5.

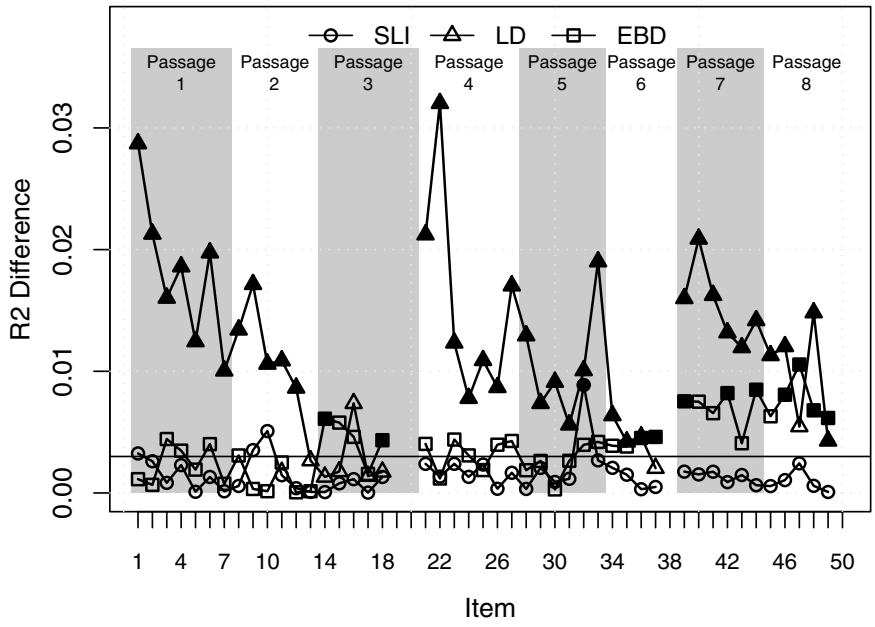


FIGURE 8. DOF $R^2$ difference between students with and without disabilities in grade 5. No $R^2$ is available for items 19, 38, and 50 because these are constructed-response items and were not analyzed. Item 20 is not a test item and was also excluded.

number of records and the small magnitude of omissions dictate caution in generalizing. As with DIF/DDF, students with LD exhibited more DOFs than the other disability groups in both grades 3 and 5. The tendency found among all groups for lower performing students to have higher omissions rates was stronger for students with disabilities than for students without disabilities. Item location (i.e., in which session or passage the items are located and where they are located within passage) seemed to affect omission rates for all students, both those with and without disabilities. There was less indication that item location affected DOF consistently across the three groups of students with disabilities.

*Discussion and Implications*

A first point to note is that different disability groups showed different results. Items showed DIF and DDF results that merited closer examination only for students with LD. No such items were identified for students with SLI or EBD given our criteria. The reason for the distinct results for students with LD is not explained by these analyses. Presumably, there is something about the characteristics of students with LD that interacts with the characteristics of some of these items. Perhaps, some of the items were particularly challenging for students with significant reading or processing challenges. More research is needed, especially examinations of the characteristics of those items showing DIF and DDF for students with LD.

The finding of different results by disability category underscores the importance of recognizing the limitations of treating all students with disabilities as a single homogeneous group and suggests that the behavior of students with different kinds of disabilities needs to be examined separately whenever possible. We speculate that for this particular study no items might show DIF/DDF if all three disability groups had been treated in one group as students with disabilities, because the differences exhibited by students with LD could be attenuated by students with SLI or EBD who responded to test items in a manner relatively similar to students without disabilities. This concern about the heterogeneity of students with disabilities led the Partnership for Academic Reading Assessment (PARA) to prepare a short literature review on the challenges of instructing and assessing reading for students with various disabilities (PARA, 2006a, 2006b, 2006c, 2006d, 2006e, 2006f, 2007) and to undertake an examination of the different ways that students' disabilities may affect their performance on reading tests (Moen, Thurlow, & Liu, 2007).

A second observation is that examining RCCs helps clarify the implications of DIF and DDF. The RCC pattern that would raise the strongest concerns about distractor bias would be if higher performing students with disabilities, those plotted toward the right side of the charts, proportionally selected a particular distractor more than other students. This would suggest that there might be something about the distractor that was a particular problem for students with disabilities.

However, this pattern was rarely observed. More often, the charts showed differential selection of distractors by the low performing students. In most cases the pattern showed that low performing students *without* disabilities were disproportionately more likely to select a particular distractor. For low performing students *with* disabilities, this often meant that all of the choices, including the correct choice, had roughly equal chances of being selected. One plausible interpretation of this pattern is that while low performing students without disabilities were led to making a wrong choice, low performing students with disabilities were making random choices. Following up on this observation would likely entail examining other aspects of the test performance of students with LD.

Third, although several items showed DIF and DDF that merited examination, no evidence of serious test bias was found for the state reading assessment examined in this study. The effect sizes ($R^2$ differences) of flagged items were not very large, even though they exceeded the criterion that was more stringent than mere statistical significance. Those items that had the largest effect sizes showed patterns where the DDF was attributable to low performing students without disabilities being attracted by a false distractor more than comparable performing students with disabilities. This pattern is not indicative of items that are biased against students with disabilities.

Finally, it is worth noting what steps to be taken in general when DIF, DDF, or DOF is found, even though the particular assessments analyzed in this study did not show serious bias as noted previously. In practice, items that exhibited DIF should be eliminated if possible because removal of these items increases validity of the test in the sense that the resulting test score is equally indicative of the underlying ability for all groups of students. If DIF is found in a post hoc analysis as in this study, then future development of similar items would be improved by examining these items for what caused the differential performance. DDF and DOF have less direct impact on item or test scores than DIF, but they would help in better understanding observed DIF; the results of this study indicated that DIF might result from DDF, that is, students at a particular ability level in one group tended to choose a particular distractor more often than their counterparts in the other group. This fact suggests that item bias can be avoided by considering not only the behavior of correct responses but also that of distractors.

As a general guideline, O'Neil and McPeek (1993) examined item and test characteristics, such as subject areas, item contents, and item formats, that might interact with characteristics of a particular group of students to produce DIF. Although their analysis did not include grouping by disability status, a similar approach can be applied to the study of students with disabilities as well as interpretation of DDF and DOF. In order to understand DDF or DOF more fully, we need to examine response processes specific to students with a particular type of disabilities, possibly by think-aloud or other qualitative procedures (Haladyna, 2004, chapter 10). Content review of DDF/DOF items can be informed by considering how the response tendencies for the problematic distractor (or omitted responses) differ, and how they relate to group-specific response processes. More research in this direction is needed.

## References


Abedi, J., Leon, S., & Kao, J. (2007a). *Examining differential item functioning in reading assessments for students with disabilities*. Minneapolis, MN: University of Minnesota, Partnership for Accessible Reading Assessment (http://www.readingassessment.info/resources/publications/examiningDIFreport.pdf, accessed April 19, 2007).

Abedi, J., Leon, S., & Kao, J. (2007b). *Examining differential distractor functioning in reading assessments for students with disabilities*. Minneapolis, MN: University of Minnesota, Partnership for Accessible Reading Assessment (http://www.readingassessment.info/resources/publications/examiningDDFreport.pdf, accessed April 19, 2007).

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of student language background on content-based performance: Analyses of extant data* (CSE Technical Report No. 603). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Banks, K. (2006). A comprehensive framework for evaluating hypotheses about cultural bias in educational testing. *Applied Measurement in Education, 19*, 115–132.

Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1987). SAT differential item performance for nine handicapped groups. *Journal of Educational Measurement, 24*, 41–55.

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

Elliott, J., Thurlow, M., Ysseldyke, J., & Erickson, R. (1997). *Providing assessment accommodations for students with disabilities in state and district assessments* (Policy Directions 7). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes (http://education.umn.edu/NCEO/OnlinePubs/Policy7.html, accessed August 28, 2007).

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

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Johnstone, C. J., Thompson, S. J., Moen, R. E., Bolt, S., & Kato, K. (2005). *Analyzing results of large-scale assessments to ensure universal design* (Technical Report 41). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Klein, J. A., Wiley, H. I., & Thurlow, M. L. (2006). *Uneven transparency: NCLB tests take precedence in public assessment reporting for students with disabilities* (Technical Report 43). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Marshall, S. P. (1983). Sex differences in mathematical errors: An analysis of distractor choices. *Journal for Research in Mathematics Education, 14*, 325–336.

Middleton, K., & Laitusis, C. C. (2007). *Examining test items for differential distractor functioning among students with learning disabilities* (Research Report RR-07–43). Princeton, NJ: Educational Testing Service.

Moen, R. E., Thurlow, M. L., & Liu, K. L. (2007). *Less accurately measured students*. Presentation at the Council of Chief State School Officers Large Scale Assessment Conference, Nashville, TN.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*, 691–692.

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002).

O'Neil, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–276). Hillsdale, NJ: Lawrence Erlbaum.

Partnership for Accessible Reading Assessment (PARA) (2006a). *Reading and students with autism* (http://www.readingassessment.info/resources/publications/readingandautism.htm, accessed August 14, 2007).

Partnership for Accessible Reading Assessment (PARA) (2006b). *Reading and students with emotional or behavioral disabilities* (http://www.reading

assessment.info/resources/publications/readingandemotbehav.htm, accessed August 14, 2007).

Partnership for Accessible Reading Assessment (PARA) (2006c). *Reading and students with specific learning disabilities* ( http : / / www . readingassessment . info/resources/publications/readingandld.htm, accessed August 14, 2007).

Partnership for Accessible Reading Assessment (PARA) (2006d). *Reading and students with speech or language impairments* (http://www.readingassessment.info/resources/publications/readingandspeech.htm, accessed August 14, 2007).

Partnership for Accessible Reading Assessment (PARA) (2006e). *Reading and students with mental retardation* (http://www.readingassessment.info/resources/publications/mentalretardation.htm, accessed August 14, 2007).

Partnership for Accessible Reading Assessment (PARA) (2006f). *Reading and students with visual impairments or blindness* (http://www.readingassessment.info/resources/publications/visualimpairment.htm, accessed August 14, 2007).

Partnership for Accessible Reading Assessment (PARA) (2007). *Reading and students who are deaf or hard of hearing* (http://www.readingassessment.info/resources/publications/deafOrhardofhearning.html, accessed August 14, 2007).

Shriner, J. G., Spande, G., & Thurlow, M. L. (1994). *State special education outcomes 1993*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Stone, E. F., Stone, D. L., & Gueutal, H. G. (1990). Influence of cognitive ability on responses to questionnaire measures: Measurement precision and missing response problems. *Journal of Applied Psychology, 75*, 418–427.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 360–370.

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

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S. J., Thurlow, M. L., & Malouf, D. (2004, May). Creating better tests for everyone through universally designed assessments. *Journal of Applied Testing Technology* (http://www.testpublishers.org/Documents/Creating_Better_Tests%20Final%20Revision%205.15.04.pdf, accessed October 6, 2008).

Thurlow, M. L., Moen, R., & Wiley, H. I. (2005). *Annual performance reports: 2002–2003 state assessment data*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes (www.nceo.info/OnlinePubs/APRsummary2006.pdf, accessed April 19, 2007).

U.S. Department of Education (2006). *26th Annual (2004) Report to Congress on the Implementation of the Individuals with Disabilities Education Act, vol. 2*. Washington, DC: U.S. Department of Education, Office of Special Education and Rehabilitative Services, Office of Special Education Programs.

Ysseldyke, J., Thurlow, M., Langenfeld, K., Nelson, J. R., Teelucksingh, E., & Seyfarth, A. (1998). *Educational results for students with disabilities: What do the data tell us?* (Technical Report 23). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense (http://educ.ubc.ca/faculty/zumbo/DIF/handbook.pdf; January 8, 2007).