CRESST REPORT 743

Jamal Abedi
Seth Leon
Jenny C. Kao

EXAMINING DIFFERENTIAL DISTRACTOR FUNCTIONING IN READING ASSESSMENTS FOR STUDENTS WITH DISABILITIES

SEPTEMBER, 2008

National Center for Research on Evaluation, Standards, and Student Testing

Graduate School of Education & Information Studies
UCLA | University of California, Los Angeles

Examining Differential Distractor Functioning in Reading Assessments for Students with Disabilities

CRESST Report 743

Jamal Abedi
National Center for Research on Evaluation, Standards, & Student Testing
University of California, Davis

Seth Leon & Jenny C. Kao
National Center for Research on Evaluation, Standards, & Student Testing
University of California, Los Angeles

September 2008

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under grant number H324F040002 from the U.S. Department of Education, Office of Special Education Programs.

The findings and opinions expressed in this report do not necessarily reflect the positions or policies of the Office of Special Education Programs or the U.S. Department of Education.

EXAMINING DIFFERENTIAL DISTRACTOR FUNCTIONING IN READING ASSESSMENTS FOR STUDENTS WITH DISABILITIES [1,2]

Jamal Abedi
National Center for Research on Evaluation, Standards, & Student Testing
University of California, Davis

Seth Leon & Jenny C. Kao
National Center for Research on Evaluation, Standards, & Student Testing
University of California, Los Angeles

Abstract

This study examines the incorrect response choices, or distractors, by students with disabilities in standardized reading assessments. Differential distractor functioning (DDF) analysis differs from differential item functioning (DIF) analysis, which treats all answers alike and examines all wrong answers against the correct answer. DDF analysis in contrast examines only the incorrect answers. If different groups, such as students with disabilities and students without disabilities, selected different incorrect responses to an item, then the item could mean something different to the different groups. Our study results found items showing DDF for students with disabilities in Grade 9, but not for Grade 3. Results also suggest that items showing DDF were more likely to be located in the second half of the assessments rather than the first half of the assessments. Additionally, results suggest that in items showing DDF, students with disabilities were less likely to choose the most common distractor than students without disabilities. Results of this study can shed light on potential factors affecting the accessibility of reading assessments for students with disabilities, in an ultimate effort to provide assessment tools that are conceptually and psychometrically sound for all students. A companion report is available examining differential item functioning for students with disabilities.

Introduction

The reauthorization of the Individuals with Disabilities Education Act (IDEA) has heightened demands for equity and accountability in education for the approximately 6.5 million children and youth with disabilities in the United States (U.S. Department of

[1] The authors acknowledge the valuable contribution of colleagues in this study. The authors are thankful to Martha Thurlow, Ross Moen, Christopher Johnstone, and other staff at the National Center on Educational Outcomes, and other members of the Partnership for Accessible Reading Assessment for their helpful comments and suggestions. The authors are also grateful to Eva Baker for her support of this work, and to Joan Herman for her extensive involvement, advice, and support of this work.

[2] An earlier version of this report was published by the University of Minnesota, Partnership for Accessible Reading Assessment.

Education, 2004). In the 2003–2004 school year, almost half of all students with disabilities were in regular classrooms for 80% or more of the school day (U.S. Department of Education, 2005). Furthermore, since the inception of the No Child Left Behind Act of 2001 (NCLB, 2002), more students with disabilities participate in assessments than in the past, and states are required to report the achievement of students with disabilities as a separate subgroup. In a review of state practices, Klein, Wiley, and Thurlow (2006) found that 44 states reported participation and performance for students with disabilities on all of their NCLB assessments. Nearly 84% of middle school students with an Individualized Education Plan (IEP) participated in general reading assessments, as reported by states in the 2002–2003 Annual Performance Reports (Thurlow, Moen, & Wiley, 2005).

Students with disabilities traditionally perform at substantially lower levels than students with no apparent disabilities (Abedi, Leon, & Mirocha, 2003; Ysseldyke et al., 1998). While their lower performance may be partially attributed to their specific disability, other factors may potentially interfere with their performance, such as the lack of opportunity to learn or lack of appropriate testing accommodations. Also, specific characteristics of the test itself may be a reason. Variables unrelated to the construct of an assessment may affect its reliability and validity for students with disabilities. Haladyna and Downing (2004) created a taxonomy of what they found to be construct-irrelevant variance in high-stakes testing. The taxonomy comprised of 21 potential sources of systematic errors associated with construct-irrelevant variance, which included factors relating to test development, such as item quality and test item format. Given the high participation of students with disabilities in standardized assessments, it is necessary to have valid and reliable measures of their knowledge and skills with minimal construct-irrelevant variance.

Classical theory of measurement is based on the assumption that measurement error has a similar random distribution for all students and no differential subgroup trend is assumed (see, for example, Allen & Yen, 1979). However, test bias can occur when performance on a test requires sources of knowledge different than those intended to be measured, causing the test scores to be less valid for a particular group (Penfield & Lam, 2000). Our current study seeks to identify a potential source of construct-irrelevant variance by examining item bias with the eventual hope of creating tests that more accurately reflect the knowledge of students with disabilities.

Educational measurement researchers and theorists have examined many different forms of detecting item bias (Matlock-Hetzel, 1997; O'Neal, 1991), including different forms of analyses for examining distractors, such as linear polytomous scoring (Crehan & Haladyna, 1994), point-biserial discrimination index (Attali & Fraenkel, 2000), factorial

modeling (Wang, 2000), standardization approach (Dorans & Holland, 1992; Dorans, Schmitt, & Bleistein, 1992), and log-linear modeling (Green, Crone, & Folk, 1989; Marshall, 1983). More recent methods of detecting item bias use the framework of differential item functioning (DIF; Penfield & Lam, 2000). DIF analyses have traditionally been used to examine item bias for members of different demographic groups, such as for determining cultural bias in a test item. For example, if a certain group performs lower on average on a specific item, then one could say that that item is biased against that particular group. DIF analyses compares the performance of two groups of the same level of ability in order to disentangle the effects of unfairness and ability level. Consistent differences between the two groups would suggest that DIF is present. However, it must be noted that considering an item as biased would also require determining the non-target constructs that lead to the between-group differences in performance (Penfield & Lam, 2000).

Green et al. (1989) extended the concept of DIF to what they termed differential distractor functioning (DDF). DDF analysis differs from DIF, which treats all answers alike and examines all wrong answers against the correct answer. DDF analysis in contrast examines only the incorrect answers. Green et al. argued that if different groups preferred different incorrect responses to an item, often called foils or *distractors*, then the item could mean something different to the different groups. Although group differences in distractor choices do not affect test scores (because all distractors are wrong), group differences might suggest differential functioning for different subgroups. In their DDF study, Green et al. used log-linear models to examine subgroups while holding ability constant, to ensure that any group differences detected were not due to differences in ability. DDF analysis acknowledges that people of different abilities are expected to pick different wrong answers, but people of different backgrounds may prefer different distractors. Green et al. argued that when a test shows substantial DDF, it is not blind to a particular group, and therefore test scores cannot be interpreted in the same way for the different groups.

While there is research examining measurement issues related to distractors in multiple-choice items, research on the role of distractors specifically on the assessment outcomes of students with disabilities is scarce. Assessment outcome studies that examine distractors tend to focus on other subgroups, such as gender. For instance, Marshall (1983) found a significant interaction between gender and choice of distractor in a large majority of items in a Grade 6 assessment. Other past research explored differences between students' ability and incorrect option choices (Levine & Drasgow, 1983; Huntley & Welch, 1993). However, as aforementioned, DDF analysis already acknowledges that different ability groups will likely pick different incorrect answers.

Given the paucity of studies examining distractor choices for students with disabilities, we were interested in exploring whether there is a differential trend of selecting distractors among students with disabilities and students without disabilities, while controlling for ability, using existing data. It is imperative to provide assessment tools that are conceptually and psychometrically sound for all students, particularly those with special needs. Results of this study can provide insight on factors that may affect the performance of students with disabilities, which may open avenues for future studies as part of an ultimate effort to ameliorate assessments for all students.

Research Questions

The following research questions guided the analyses and reporting of this study:

1. Do items on standardized Reading Comprehension (RC) and Word Analysis (WA) subscales exhibit differential distractor functioning (DDF) for students with disabilities?
2. Does item location have an impact on DDF for students with disabilities? Specifically, are more items that exhibit DDF for students with disabilities located in the second half of RC and WA subscales rather than in the first half?
3. Do the results of DDF vary by grade (from Grade 3 to Grade 9)?

Methodology

Data Source

Data from a single state provided the impetus for answering the above research questions. The data were obtained for the 1997–1998 academic year from a small state with an average number of students with disabilities, and included item-level information on students' responses in the Stanford Achievement Test, Ninth Edition (Stanford 9). Published by Harcourt Brace Educational Measurement in 1996, the Stanford 9 is a standardized, norm-referenced test in several subject areas, including reading. According to the Harcourt Assessment website, the Stanford 9 uses an "easy-hard-easy format" in which "difficult questions are surrounded by easy questions to encourage students to complete the test" (HarcourtAssessment.com, n.d.). The reading portion of the test is characterized by three different types of reading selections: recreational, textual, and functional, and includes items that assess initial understanding, interpretation, critical analysis, and reading strategy (HarcourtAssessment.com, n.d.).

The present study examines two subscales of the Stanford 9, Reading Comprehension (RC) and Word Analysis (WA) (more commonly known as "phonics" or "decoding"), from the above mentioned state (which is not named to preserve anonymity). Data from public

school students in Grades 3 and 9 were analyzed to present data over a wider age range. Students with valid scores were included in our analyses. Students with limited English proficient (LEP) classifications (including LEP students with disabilities) were excluded from the analyses to reduce the possible confounding of language proficiency issues. Of the 6,611 third-grade students included in the present analyses, 448 (6.8%) were considered to be students with disabilities. Of the 5,287 ninth-grade students, 522 (9.9%) were considered to be students with disabilities.

Procedure & Statistical Design

Multiple-choice items in two reading subscales (RC & WA) were selected for this study. Each multiple-choice item had four response options consisting of a correct response and three distractors. Analyses were conducted for Grade 3 and Grade 9 students. We selected two grades apart enough from each other to examine the possible differences between the grade and age of students.

To examine the possibility of differential distractor functioning (DDF) across the categories of students with disabilities and students without disabilities, we used a multi-step logistic regression procedure. Item distractors are often designed to draw the attention of students with partial knowledge of the question. Therefore it is important to control for student ability in the construct when attempting to test for DDF. This is especially true in subgroups such as students with disabilities, who are known to have performance gaps relative to students without disabilities.

Only incorrect responses were considered in this analysis, and responses were grouped into two categories. One category represented students who selected the most common distractor, and the other category consisted of students who selected one of the two less common distractors. This indicator of distractor selection was used as the criterion variable. A total score on the applicable subscale (RC or WA) was computed as a proxy for ability on the construct. This score was standardized to provide easier interpretation of odds ratios. In Step 1, the ability proxy was entered into the model and a measure of the explained variance (Nagelkerke *R*-square) was obtained. In Step 2, the students with disabilities grouping variable and an interaction between the grouping variable and the ability proxy were entered into the model. Again the *R*-square estimate was obtained. The change in *R*-square between Step 1 and Step 2 was calculated and tested for significance. Analyses were performed for each item separately. Since there were large numbers of items in each content area the adjustment of type I error rate (α) for multiple analyses was not practical. Therefore results emphasize trends in the significance levels rather than focusing on findings for any particular

item. Items were identified for closer inspection as having differential distractor functioning if the *R*-square change was at least 0.003 and was significant at $p < 0.01$.

A similar approach was used to determine if item location influences DDF for students with disabilities. Rather than using the total score as a proxy for ability, only the total score on items from the first half of the assessment was used as an ability proxy (first 27 out of 54 items for RC; first 15 out of 30 items for WA). Using this second type of proxy enables us to examine potential influences of item location. Items that exhibited DDF were examined more closely by looking at the odds ratios of the variables in the final model. If systemic differences in the DDF findings arose between the two approaches they could then be compared. For example, if items showed larger DDF effects on the items from the latter portion of the assessment when the second proxy was used, and if the odds ratios on those items were in a consistent direction, then it would be apparent that item location was influencing DDF. There could be multiple reasons why item location might influence DDF, for students with disabilities including but not limited to time pressures, fatigue, frustration and motivation. However, determining specific reasons was beyond the scope of this study.

The RC and WA subscales from a single state for both Grade 3 and Grade 9 were used for these analyses. The next section describes findings from our analyses. More detailed results of the DDF analyses are available in the Appendix.

Results

The analyses examine the following research questions:

1. Do items on standardized Reading Comprehension (RC) and Word Analysis (WA) subscales exhibit differential distractor functioning (DDF) for students with disabilities?
2. Does item location have an impact on DDF for students with disabilities? Specifically, are more items that exhibit DDF for students with disabilities located in the second half of RC and WA subscales rather than in the first half?
3. Do the results of DDF vary by grade (from Grade 3 to Grade 9)?

Separate models were used to answer the above research questions. To answer Question 1, student ability was measured as the total score on the applicable subscale (RC or WA). To answer Question 2, student ability was measured as the score on items from the first half of the applicable assessment. In each model the additional variance explained by the introduction of disability status and the interaction between disability status and student ability was examined to determine if the item in question met the specified thresholds to be considered as showing DDF. To answer research Question 3, the pattern of DDF across

Grade 3 and Grade 9 was compared. Results are described in the following pages by subscale and by grade.

Reading Comprehension

Grade 3. Table 1 presents a summary of the results of the Grade 3 reading comprehension items. With Model 1, in which the total score on the 54-item RC assessment was used as an ability proxy, four items showed DDF. Three of the four items that showed DDF were from the second half of the assessment. In Model 2 only the score on the first 27 RC items was used to measure student reading ability. If systemic DDF is present for items on the second half of the assessment we would expect to see more items show DDF using this approach. However, only three items showed DDF using the Model 2. These results suggest that DDF is minimal on the RC assessment in Grade 3, and there is little evidence of item location influencing DDF.

Table 1

Grade 3 Item-level Reading Comprehension

Ability proxy	Total number of items	Number of items showing DDF		
		Items 1–27	Items 28–54	All items
Model 1	54	1	3	4
Model 2	54	1	2	3

Note. In Model 1, the total score was used as an ability proxy. In Model 2, the score on the first 27 items was used as an ability proxy, DDF = differential distractor functioning.

Grade 9. Table 2 presents results for the Grade 9 RC subscale which are very different from what was seen Grade 3. With Model 1, in which the total score on the 54-item RC assessment was used as an ability proxy, 10 items showed DDF. Six of the ten items that showed DDF were from the second half of the assessment. With Model 2, in which the first half of the RC assessment was used as the measure of student reading ability, 13 items exhibited DDF. Using the Model 2 approach, 10 of the 13 items that showed DDF originated from the second half of the RC assessment. There was substantially more DDF present in Grade 9 than in Grade 3.

Table 2

Grade 9 Item-level Reading Comprehension

Ability proxy	Total number of items	Number of items showing DDF		
		Items 1–27	Items 28–54	All items
Model 1	54	4	6	10
Model 2	54	3	10	13

Note. In Model 1, the total score was used as an ability proxy. In Model 2, the score on the first 27 items was used as an ability proxy, DDF = differential distractor functioning.

Table 3 presents the results for items in Model 2 that showed DDF. Results are presented for each step in the logistic regression. Odds ratios are also reported for each variable to determine if the DDF is operating similarly for items from the second half of the RC assessment. Of the 10 items that showed DDF from the second half of the assessment, 7 had a significant main effect for the disability status variable. For each of these seven items the odds ratio for students with disabilities was less than 1.0, indicating that students with disabilities were less likely to choose the most commonly chosen distractor when compared to students without disabilities. For example, on Item 33, while controlling for reading ability on the first 27 items, students with disabilities were about one third as likely (0.35) to select the most commonly selected distractor when compared to students without disabilities.

Table 3

Grade 9 Item-level Reading Comprehension Logistic Regression Results for Items Showing DDF with Ability Proxy Based On First 27 Items Score

	R-square results at each step in the sequential logistic regression			Odds ratios – Final model		
Item no.	Step 1 Ability proxy	Step 2 Ability proxy and disability status (Uniform)	Step 3 Ability Proxy, disability status and interaction (Non-uniform)	Ability proxy	Disability status	Interaction
17	0.006**	0.010*	0.14*	1.05	1.12	1.42*
22	0.218**	0.236**	0.240**	2.67**	0.26**	0.63**
23	0.021**	0.025**	0.038**	0.66*	1.46*	2.01**
30	0.169**	0.169	0.179**	2.59*	0.43**	0.54**
33	0.104**	0.120**	0.124**	1.85**	0.35**	0.68**
38	0.065**	0.070**	0.070	1.52**	0.69*	1.04
39	0.030**	0.036**	0.040*	1.38**	0.43**	0.69*
41	0.013**	0.021**	0.025**	0.72**	0.83	1.44**
42	0.097**	0.104**	0.105	1.77**	0.54**	0.91
44	0.115**	0.119**	0.121	1.90**	0.54**	0.76*
46	0.000	0.010**	0.016*	0.88	1.09	1.53*
52	0.102**	0.105**	0.107	2.03**	1.13	0.78*
54	0.016**	0.025**	0.026	1.23**	0.55**	0.86

Note. DDF = differential distractor functioning.
* denotes significance at $p < .05$. ** denotes significance at $p < .01$

Word Analysis

Grade 3. Table 4 presents a summary of the results of the Grade 3 Word Analysis items. With Model 1, in which the total score on the 30-item WA assessment was used as an ability proxy, three items showed DDF. All three items that showed DDF were from the second half of the assessment. In Model 2 only the score on the first 15 WC items were used to measure student reading ability. If systemic DDF is present for items on the second half of the assessment we would expect to see more items showing DDF. Only one item showed DDF using the Model 2 approach. These results are similar to the results on the RC assessment in Grade 3 and suggest that DDF is minimal on the WC assessment with little evidence of item location influencing DDF.

Table 4

Grade 3 Item-level Word Analysis

Ability proxy	Total number of items	Number of items showing DDF		
		Items 1–15	Items 16–30	All items
Model 1	30	0	3	3
Model 2	30	0	1	1

Note. In Model 1, the total score was used as an ability proxy. In Model 2, the score on the first 15 items was used as an ability proxy, DDF = differential distractor functioning.

Grade 9. Table 5 presents results for the Grade 9 WA subscale which are again quite different than what was seen Grade 3. With Model 1, in which the total score on the 30-item WA assessment was used as an ability proxy, 12 items showed DDF. Eight of the 12 items that showed DDF were from the second half of the assessment. With Model 2, in which the first half of the WC assessment was used as a measure of student reading ability, 11 items showed DDF. Seven of the 11 items that showed DDF were from the second half of the RC assessment. There was substantially more DDF present in Grade 9 than in Grade 3.

Table 5

Grade 9 Item-level Word Analysis

Ability proxy	Total number of items	Number of items showing DDF		
		Items 1–15	Items 16–30	All items
Model 1	30	4	8	12
Model 2	30	4	7	11

Note. In Model 1, the total score was used as an ability proxy. In Model 2, the score on the first 15 items was used as an ability proxy, DDF = differential distractor functioning..

Table 6 presents the results for items in Model 2 that showed DDF. Results are presented for each step in the logistic regression. Odds ratios are also reported for each variable to determine if DDF is operating similarly for items from the second half of the RC assessment. Of the seven items that showed DDF from the second half of the assessment, six had a significant main effect for the disability status variable. For each of these six items the odds ratio for students with disabilities was less than 1.0, indicating that students with disabilities were less likely to choose the most commonly chosen distractor when compared to students without disabilities. For example, on Item 20, while controlling for reading ability on the first 15 items, students with disabilities were less than one third as likely (0.31) to select the most commonly selected distractor when compared to students without disabilities.

Table 6

Grade 9 Item-level Word Analysis Logistic Regression Results for Items Showing DDF with Ability Proxy Based On First 15 Items Score

	R-square results at each step in the sequential logistic regression			Odds ratios – Final model		
		Step 2	Step 3			
Item no.	Step 1 Ability proxy	Ability proxy and disability status (Uniform)	Ability proxy, disability status and interaction (Non-uniform)	Ability proxy	Disability status	Interaction
1	0.007	0.025**	0.029	1.03	0.87	1.42
5	0.033**	0.041**	0.044*	1.32**	0.80	1.41**
6	0.128**	0.135**	0.135	0.48**	1.85**	1.16
8	0.019**	0.029**	0.030	1.22**	0.66**	1.20
18	0.022**	0.036**	0.036	1.23**	0.59**	1.05
20	0.127**	0.148**	0.152**	2.27**	0.31**	0.69*
22	0.050**	0.067**	0.067	1.52**	0.46**	0.93
24	0.000	0.007**	0.009	0.93	0.80	1.28
25	0.010**	0.033**	0.035*	1.17**	0.33**	0.74
26	0.027**	0.041**	0.041	1.32**	0.48**	0.90
27	0.002	0.009**	0.009	1.04	0.65**	0.97

Note. DDF = differential distractor functioning.
* denotes significance at $p < .05$, ** denotes significance at $p < .01$

Discussion

The national achievement trend shows that students with disabilities perform considerably lower than students with no apparent disabilities. While these achievement gaps can be partly explained by the interference of students' specific disabilities, other factors may also contribute to the performance gaps. Factors related to both the instruction and assessment of these students play a large role in their achievement. While we believe factors related to instruction and assessment are intricately intertwined, this study focuses on the factors that influence the assessment of students with disabilities, particularly related to test format. Specifically, this study focuses on the test items themselves, and whether the items have a potential bias against students with disabilities. The present study therefore explored distractor choices amongst students with disabilities using an existing data set. Results of this study can shed light on potential factors affecting the accessibility of reading assessments for

students with disabilities, in an ultimate effort to provide assessment tools that are conceptually and psychometrically sound for all students.

The following research questions guided this study:

1. Do items on standardized Reading Comprehension (RC) and Word Analysis (WA) subscales exhibit differential Distractor Functioning (DDF) for students with disabilities?
2. Does item location have an impact on DDF for students with disabilities? Specifically, are more items that exhibit DDF for students with disabilities located in the second half of RC and WA subscales rather than in the first half?
3. Do the results of DDF vary by grade (from Grade 3 to Grade 9)?

To answer these research questions, student responses on multiple-choice items were compared across the disability status categories in two reading subscales of the Stanford 9, Reading Comprehension and Word Analysis, in two grade levels (3 and 9) from public schools in an entire state. Each multiple-choice item consisted of four response options (one correct and three distractors). Item distractors are often designed to draw the attention of students with partial knowledge of the question. Therefore it is important to control for student ability in the construct when attempting to examine differential distractor functioning (DDF).

DDF assumes that people of different ability already naturally choose different wrong answers (Green et al., 1989), when they are not sure about the correct response to the item. It is when substantial DDF is shown for a particular group that there is cause for concern. Our present study results suggest that a substantial number of items exhibit DDF for students with disabilities in Grade 9. Results also suggest that items showing DDF were more likely to be located in the second half of the assessments rather than the first half of the assessments. Results also suggest that DDF was present for Grade 9 test items, but not for Grade 3 items. Even when controlling for ability using only the items in the first half of the assessments, more Grade 9 items exhibited DDF than Grade 3 items.

For items showing DDF, odds ratios for students with disabilities were less than 1.0, which suggests that students with disabilities were less likely to choose the most common distractor as compared to students without disabilities. This may suggest that students with disabilities might be more randomly selecting one of the four response options rather than making an "educated guess." Our concurrent study, which employs DIF analyses, sheds additional light on differential response patterns for students with disabilities, and is available in a companion report. Findings from our concurrent DIF study are consistent with the DDF analyses, in that students with disabilities were shown to perform more poorly on items

located in the second half of the assessment, even while controlling for their performance on the first half of the assessment (see our companion report—CRESST Tech. Rep. No. 744—for more details: Abedi, Leon, & Kao, 2008).

The findings of this study have multiple implications. First we might speculate why items located in the second half of assessments showed more DDF than items located in the first half. It could be that students with disabilities require more time than is allowed to complete the tests, or that they became fatigued or frustrated by a certain point in the test. Since responses to these items appear to have been selected more randomly, it could be that many students with disabilities did not have the time or energy to have thoroughly read or given the items much thought. Or it could be that they reached a certain cognitive overload. More research would be necessary to determine the actual cause or causes, possibly with in-depth qualitative research. We might also speculate why DDF was present for Grade 9 items but not for Grade 3 items. This might be attributed to the content and construct of the test, or to the students themselves. The content of assessments is likely more complex in the higher grades and thus, more nuisance variables influence assessments in the higher grades.

It is necessary to note that this study has several major limitations. For instance, it does not differentiate between different categories of disabilities. Student performance across different categories of disabilities may be quite different and these factors may affect their performance quite differently. Given the heterogeneity of students with disabilities, it is not ideal to group them together into one category, and additional insight could be gathered from analyzing data by specific disability groups. This study was also limited in terms of scope. We did not have access to the specific types of testing accommodations that these students may have received, such as whether students received extended time to complete the assessments. Also, without access to the actual test items, we were unable to make any conclusive statements regarding the content of the tests, especially with respect to the differences across the two grade levels. Further investigation is required for future studies.

Nevertheless, findings of this study provide evidence that other factors related to the assessments, such as test format, may contribute to the performance gap between students with disabilities and their students without disabilities peers. Controlling for factors that are not related to the content being assessed may help test developers provide more accessible and more valid assessments for students with disabilities.

References

Abedi, J., Leon, S., & Kao, J. C. (2008). *Examining differential item functioning in reading assessments for students with disabilities*. (CRESST Tech. Rep. No. 744). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Attali, Y., & Fraenkel, T. (2000). The point-biserial as a discrimination index for distractors in multiple-choice items: Deficiencies in usage and an alternative. *Journal of Educational Measurement, 37*(1), 77–86.

Crehan, K. D., & Haladyna, T. M. (1994). *A comparison of three linear polytomous scoring methods*. (ERIC Document Reproduction Services No. ED 377246).

Dorans, N. J. & Holland, P. W. (1992). *DIF detection and description: Mantel-Haenszel and standardization*. (Educational Testing Service Report ETS-RR-92-10).

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement, 29*(4), 309–319.

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement, 26*(2), 147–160.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27.

HarcourtAssessment.com. (n.d.). *Stanford Achievement Test series, Ninth edition–Complete battery.* Retrieved May 18, 2006, from http://harcourtassessment.com/hai/ProductLongDesc.aspx?ISBN=E132C&Catalog=TPC-USCatalog&Category=AchievementAccountability

Huntley, R. M., & Welch, C. (1993, April). *Numerical answer options: Logical or random order?* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

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 No. 43). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Levine, M. V., & Drasgow, F. (1983). The relationship between incorrect option choice and estimated ability. *Educational and Psychological Measurement, 43*, 675–685.

Marshall, S. P. (1983). Sex differences in mathematical errors: An analysis of distracter choices. *Journal for Research in Mathematics Education, 14*(5), 325–336.

Matlock-Hetzel, S. (1997, January). *Basic concepts in item and test analysis*. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX.

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

O'Neal, M. R. (1991). A comparison of methods for detecting item bias (Doctoral dissertation, The University of Alabama, 1991). *Dissertation Abstracts International, 52*(5), 1723A.

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

Thurlow, M. L., Moen, R. E., & Wiley, H. I. (2005). *Annual performance reports: 2002–2003 state assessment data*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved May 2, 2006, from http://education.umn.edu/nceo/OnlinePubs/APRsummary2005.pdf

U.S. Department of Education. (2004). *Digest of education statistics, 2003* (NCES 2005-025). Author: National Center for Education Statistics.

U.S. Department of Education. (2005). *The condition of education 2005* (NCES 2005-094). Author: National Center for Education Statistics.

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 No. 23). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Wang, W. (2000). Factorial modeling of differential distractor functioning in multiple-choice items. *Journal of Applied Measurement, 1*(3), 238–256.

Appendix

Detailed DDF Results

Table A1

Grade 3 Item-level Reading Comprehension Ability Proxy Based On All 54 Items

	R-squared values at each step in the sequential hierarchical regression		DDF results	
Item	Step 1: Ability proxy	Step 2 Ability proxy, disability status & interaction	Chi-Sq *P*-value	Change in *R*-Square (Effect size)
1	0.005	0.011	0.139	0.006
2	0.018	0.023	0.535	0.003
3	0.137	0.143	0.161	0.006
4	0.015	0.016	0.349	0.001
5	0.014	0.031	0.018	0.017
6	0.029	0.033	0.054	0.004
7	0.024	0.026	0.309	0.002
8	0.144	0.145	0.175	0.001
9	0.168	0.170	0.168	0.002
10	0.005	0.006	0.238	0.001
11	0.001	0.006	0.175	0.005
12	0.051	0.052	0.479	0.001
13	0.012	0.015	0.208	0.003
14	0.015	0.017	0.152	0.002
15	0.000	0.001	0.450	0.001
16	0.006	0.007	0.264	0.001
17	0.011	0.013	0.442	0.002
18	0.031	0.032	0.075	0.001
19	0.005	0.007	0.520	0.002
20	0.005	0.005	0.648	0.000
21	0.032	0.033	0.448	0.001
22	0.002	0.004	0.194	0.002
23	0.002	0.012	0.000	0.010
24	0.007	0.007	0.974	0.000
25	0.023	0.025	0.402	0.002

(table continues)

Item	*R*-squared values at each step in the sequential hierarchical regression		DDF results	
	Step 1: Ability proxy	Step 2 Ability proxy, disability status & interaction	Chi-Sq *P*-value	Change in *R*-Square (Effect size)
26	0.064	0.065	0.330	0.001
27	0.023	0.026	0.050	0.003
28	0.003	0.004	0.523	0.001
29	0.021	0.022	0.584	0.001
30	0.000	0.005	0.063	0.005
31	0.034	0.038	0.200	0.004
32	0.053	0.053	0.578	0.000
33	0.000	0.001	0.925	0.001
34	0.004	0.004	0.986	0.000
35	0.020	0.020	0.820	0.000
36	0.028	0.028	0.666	0.000
37	0.088	0.090	0.053	0.002
38	0.179	0.180	0.205	0.001
39	0.204	0.208	0.010	0.004
40	0.034	0.038	0.019	0.004
41	0.027	0.035	0.000	0.008
42	0.016	0.018	0.165	0.002
43	0.015	0.015	0.730	0.000
44	0.065	0.065	0.950	0.000
45	0.075	0.075	0.570	0.000
46	0.006	0.009	0.010	0.003
47	0.028	0.028	0.723	0.000
48	0.008	0.009	0.390	0.001
49	0.005	0.006	0.333	0.001
50	0.123	0.124	0.214	0.001
51	0.003	0.007	0.005	0.004
52	0.004	0.004	0.903	0.000
53	0.103	0.103	0.674	0.000
54	0.010	0.011	0.316	0.001

Table A2

Grade 3 Item-level Reading Comprehension Ability Proxy Based On First 27 Items

	R-squared values at each step in the sequential hierarchical regression		DDF results	
Item	Step 1 Ability proxy	Step 2 Ability proxy, disability status, interaction	Chi-Sq *P*-value	Change in *R*-Square (Effect size)
1	0.007	0.014	0.085	0.007
2	0.006	0.016	0.363	0.008
3	0.140	0.145	0.197	0.005
4	0.020	0.021	0.473	0.001
5	0.014	0.029	0.034	0.015
6	0.023	0.027	0.064	0.004
7	0.029	0.031	0.378	0.002
8	0.134	0.135	0.327	0.001
9	0.169	0.170	0.237	0.001
10	0.005	0.007	0.178	0.002
11	0.001	0.006	0.158	0.005
12	0.042	0.044	0.398	0.000
13	0.013	0.014	0.562	0.001
14	0.021	0.024	0.058	0.003
15	0.000	0.000	0.545	0.000
16	0.007	0.009	0.186	0.002
17	0.013	0.014	0.520	0.001
18	0.029	0.030	0.140	0.001
19	0.004	0.005	0.562	0.001
20	0.002	0.003	0.431	0.001
21	0.021	0.022	0.345	0.001
22	0.006	0.007	0.256	0.001
23	0.004	0.013	0.000	0.009
24	0.008	0.008	0.927	0.000
25	0.022	0.023	0.327	0.001
26	0.073	0.074	0.572	0.001
27	0.013	0.017	0.015	0.004
28	0.006	0.007	0.750	0.001
29	0.022	0.022	0.697	0.000

(table continues)

Item	*R*-squared values at each step in the sequential hierarchical regression		DDF results	
	Step 1 Ability proxy	Step 2 Ability proxy, disability status, interaction	Chi-Sq *P*-value	Change in *R*-Square (Effect size)
30	0.000	0.004	0.094	0.004
31	0.021	0.023	0.432	0.002
32	0.042	0.043	0.522	0.001
33	0.000	0.001	0.464	0.001
34	0.007	0.007	0.934	0.000
35	0.021	0.022	0.517	0.001
36	0.022	0.023	0.700	0.001
37	0.075	0.077	0.108	0.002
38	0.132	0.133	0.233	0.001
39	0.152	0.156	0.019	0.004
40	0.024	0.029	0.011	0.005
41	0.010	0.015	0.002	0.005
42	0.011	0.014	0.044	0.003
43	0.012	0.012	0.658	0.000
44	0.043	0.043	0.718	0.000
45	0.052	0.052	0.750	0.000
46	0.008	0.010	0.014	0.002
47	0.019	0.019	0.783	0.000
48	0.005	0.006	0.301	0.001
49	0.002	0.003	0.352	0.001
50	0.081	0.083	0.138	0.002
51	0.002	0.007	0.003	0.005
52	0.003	0.003	0.853	0.000
53	0.068	0.068	0.902	0.000
54	0.007	0.008	0.376	0.001

Table A3

Grade 9 Item level Reading Comprehension Ability Proxy Based on All 54 Items

	R-squared values at each step in the sequential hierarchical regression		DDF results	
		Step 2		
Item	Step 1 Ability proxy	Ability proxy, disability status & interaction	Chi-Sq *P*-value	Change in *R*-Square (Effect size)
1	0.090	0.095	0.214	0.005
2	0.006	0.012	0.019	0.006
3	0.049	0.050	0.580	0.001
4	0.006	0.018	0.090	0.012
5	0.023	0.025	0.345	0.002
6	0.024	0.026	0.126	0.002
7	0.096	0.100	0.023	0.004
8	0.002	0.009	0.083	0.007
9	0.201	0.202	0.658	0.001
10	0.192	0.194	0.316	0.002
11	0.021	0.022	0.960	0.001
12	0.000	0.002	0.270	0.002
13	0.275	0.290	0.008	0.015
14	0.003	0.005	0.327	0.002
15	0.031	0.032	0.557	0.001
16	0.317	0.319	0.201	0.002
17	0.005	0.013	0.004	0.008
18	0.013	0.016	0.202	0.003
19	0.002	0.005	0.143	0.003
20	0.008	0.011	0.060	0.003
21	0.003	0.007	0.194	0.004
22	0.259	0.272	0.000	0.013
23	0.034	0.056	0.000	0.022
24	0.335	0.339	0.052	0.004
25	0.020	0.024	0.050	0.004
26	0.241	0.243	0.313	0.002
27	0.166	0.172	0.014	0.006
28	0.021	0.023	0.119	0.002

(table continues)

Item	*R*-squared values at each step in the sequential hierarchical regression		DDF results	
	Step 1	Step 2		
	Ability proxy	Ability proxy, disability status & interaction	Chi-Sq *P*-value	Change in *R*-Square (Effect size)
29	0.001	0.004	0.107	0.002
30	0.217	0.227	0.002	0.010
31	0.001	0.001	0.764	0.000
32	0.027	0.033	0.040	0.005
33	0.151	0.163	0.000	0.012
34	0.202	0.207	0.063	0.005
35	0.019	0.019	0.887	0.000
36	0.030	0.031	0.19	0.001
37	0.011	0.013	0.341	0.002
38	0.080	0.083	0.029	0.003
39	0.051	0.058	0.020	0.007
40	0.027	0.027	0.604	0.000
41	0.019	0.032	0.000	0.013
42	0.121	0.125	0.005	0.004
43	0.031	0.034	0.123	0.003
44	0.151	0.154	0.033	0.003
45	0.040	0.041	0.537	0.001
46	0.000	0.013	0.006	0.013
47	0.037	0.037	0.752	0.000
48	0.186	0.186	0.999	0.000
49	0.012	0.015	0.047	0.003
50	0.018	0.019	0.286	0.001
51	0.032	0.033	0.552	0.001
52	0.115	0.122	0.000	0.007
53	0.000	0.001	0.320	0.001
54	0.031	0.037	0.026	0.006

Table A4

Grade 9 Item-Level Reading Comprehension Ability Proxy Based On First 27 Items

	R-squared values at each step in the sequential hierarchical regression		DDF results	
Item	Step 1 Ability proxy	Step 2 Ability proxy, disability status & interaction	Chi-Sq *P*-value	Change in *R*-Square (Effect size)
1	0.072	0.079	0.130	0.007
2	0.003	0.007	0.105	0.004
3	0.050	0.051	0.653	0.001
4	0.008	0.014	0.300	0.006
5	0.032	0.034	0.312	0.002
6	0.027	0.028	0.228	0.001
7	0.100	0.104	0.036	0.004
8	0.002	0.008	0.121	0.006
9	0.196	0.197	0.544	0.001
10	0.190	0.193	0.219	0.002
11	0.034	0.035	0.874	0.001
12	0.000	0.002	0.184	0.002
13	0.265	0.275	0.041	0.010
14	0.001	0.003	0.324	0.002
15	0.044	0.044	0.910	0.001
16	0.323	0.327	0.056	0.004
17	0.006	0.014	0.007	0.008
18	0.013	0.016	0.229	0.003
19	0.002	0.006	0.120	0.004
20	0.005	0.008	0.067	0.003
21	0.001	0.005	0.159	0.004
22	0.218	0.240	0.000	0.022
23	0.021	0.038	0.000	0.017
24	0.325	0.330	0.019	0.005
25	0.022	0.024	0.239	0.002
26	0.222	0.226	0.037	0.004
27	0.147	0.153	0.014	0.006
28	0.017	0.019	0.157	0.002
29	0.003	0.004	0.223	0.001

(table continues)

Item	*R*-squared values at each step in the sequential hierarchical regression		DDF results	
	Step 1 Ability proxy	Step 2 Ability proxy, disability status & interaction	Chi-Sq *P*-value	Change in *R*-Square (Effect size)
30	0.169	0.179	0.001	0.010
31	0.001	0.001	0.568	0.000
32	0.025	0.030	0.064	0.005
33	0.104	0.124	0.000	0.020
34	0.158	0.166	0.011	0.008
35	0.016	0.016	0.951	0.000
36	0.022	0.024	0.077	0.002
37	0.005	0.007	0.359	0.002
38	0.065	0.070	0.004	0.005
39	0.030	0.040	0.003	0.010
40	0.013	0.015	0.290	0.002
41	0.013	0.025	0.000	0.012
42	0.097	0.105	0.000	0.008
43	0.027	0.029	0.246	0.002
44	0.115	0.121	0.002	0.006
45	0.030	0.032	0.140	0.002
46	0.000	0.016	0.002	0.016
47	0.030	0.030	0.484	0.000
48	0.119	0.121	0.318	0.002
49	0.010	0.013	0.038	0.003
50	0.017	0.019	0.241	0.002
51	0.021	0.021	0.971	0.000
52	0.102	0.107	0.003	0.005
53	0.000	0.002	0.175	0.002
54	0.016	0.026	0.003	0.010

Table A5

Grade 3 Item-Level Word Analysis Ability Proxy All 30 Items

Item	*R*-squared values at each step in the sequential hierarchical regression		DDF results	
	Step 1 Ability proxy	Step 2 Ability proxy, disability status & interaction	Chi-Sq *P*-value	Change in *R*-Square (Effect size)
1	0.011	0.011	0.943	0.000
2	0.018	0.020	0.087	0.002
3	0.025	0.029	0.296	0.004
4	0.000	0.003	0.056	0.003
5	0.045	0.047	0.062	0.002
6	0.000	0.009	0.356	0.009
7	0.000	0.002	0.564	0.002
8	0.000	0.001	0.368	0.001
9	0.164	0.165	0.599	0.001
10	0.000	0.002	0.424	0.002
11	0.007	0.008	0.583	0.001
12	0.001	0.003	0.582	0.002
13	0.000	0.003	0.036	0.003
14	0.001	0.004	0.442	0.003
15	0.034	0.035	0.610	0.001
16	0.000	0.001	0.731	0.001
17	0.033	0.041	0.000	0.008
18	0.001	0.011	0.313	0.010
19	0.213	0.219	0.003	0.006
20	0.014	0.015	0.299	0.001
21	0.047	0.050	0.065	0.003
22	0.115	0.119	0.054	0.006
23	0.087	0.087	0.759	0.000
24	0.002	0.002	0.765	0.000
25	0.079	0.080	0.583	0.001
26	0.002	0.002	0.443	0.000
27	0.003	0.005	0.113	0.002
28	0.159	0.161	0.147	0.002
29	0.003	0.004	0.238	0.001
30	0.011	0.017	0.009	0.006

Table A6

Grade 3 Item-Level Word Analysis Ability Proxy First 15 Items

Item	*R*-squared values at each step in the sequential hierarchical regression		DDF results	
	Step 1 Ability proxy	Step 2 Ability proxy, disability status & interaction	Chi-Sq *P*-value	Change in *R*-Square (Effect size)
1	0.015	0.016	0.768	0.001
2	0.014	0.016	0.097	0.002
3	0.030	0.033	0.458	0.003
4	0.000	0.003	0.054	0.003
5	0.033	0.037	0.040	0.004
6	0.000	0.007	0.485	0.007
7	0.000	0.001	0.788	0.001
8	0.000	0.002	0.106	0.002
9	0.155	0.156	0.194	0.001
10	0.000	0.003	0.369	0.003
11	0.007	0.007	0.904	0.000
12	0.004	0.005	0.602	0.001
13	0.000	0.002	0.100	0.002
14	0.002	0.004	0.455	0.002
15	0.029	0.029	0.869	0.000
16	0.000	0.001	0.709	0.001
17	0.026	0.037	0.000	0.011
18	0.001	0.003	0.741	0.002
19	0.123	0.127	0.018	0.004
20	0.010	0.011	0.107	0.001
21	0.035	0.039	0.041	0.004
22	0.086	0.091	0.039	0.005
23	0.063	0.063	0.601	0.000
24	0.001	0.001	0.560	0.000
25	0.063	0.064	0.587	0.001
26	0.001	0.002	0.476	0.001
27	0.001	0.002	0.222	0.001
28	0.136	0.138	0.133	0.002
29	0.003	0.004	0.309	0.001
30	0.014	0.020	0.012	0.006

Table A7

Grade 9 Item-Level Word Analysis Ability Proxy All 30 Items

Item	*R*-squared values at each step in the sequential hierarchical regression		DDF results	
	Step 1 Ability proxy	Step 2 Ability proxy, disability status & interaction	Chi-Sq *P*-value	Change in *R*-Square (Effect size)
1	0.017	0.032	0.009	0.015
2	0.005	0.017	0.005	0.012
3	0.003	0.003	0.556	0.000
4	0.002	0.004	0.093	0.002
5	0.024	0.036	0.000	0.012
6	0.181	0.184	0.058	0.003
7	0.026	0.027	0.266	0.001
8	0.028	0.036	0.000	0.008
9	0.056	0.058	0.060	0.002
10	0.018	0.019	0.473	0.001
11	0.009	0.009	0.777	0.000
12	0.011	0.013	0.054	0.002
13	0.005	0.008	0.066	0.003
14	0.021	0.024	0.076	0.003
15	0.001	0.006	0.074	0.005
16	0.001	0.004	0.094	0.003
17	0.030	0.030	0.842	0.000
18	0.042	0.051	0.006	0.009
19	0.001	0.003	0.369	0.002
20	0.225	0.233	0.004	0.008
21	0.009	0.010	0.802	0.001
22	0.094	0.101	0.011	0.007
23	0.013	0.014	0.720	0.001
24	0.000	0.014	0.000	0.014
25	0.018	0.037	0.000	0.019
26	0.037	0.047	0.000	0.010
27	0.003	0.010	0.004	0.007
28	0.055	0.055	0.760	0.000
29	0.005	0.010	0.005	0.005
30	0.008	0.020	0.007	0.012

Table A8

Grade 9 Item-Level Word Analysis Ability Proxy Based on First 15 Items

	R-squared values at each step in the sequential hierarchical regression		DDF results	
Item	Step 1 Ability proxy	Step 2 Ability proxy, disability status & interaction	Chi-Sq *P*-value	Change in *R*-Square (Effect size)
1	0.007	0.029	0.001	0.022
2	0.013	0.023	0.013	0.010
3	0.003	0.003	0.676	0.000
4	0.001	0.002	0.120	0.001
5	0.033	0.044	0.000	0.011
6	0.128	0.135	0.000	0.007
7	0.017	0.019	0.062	0.002
8	0.019	0.030	0.000	0.011
9	0.047	0.048	0.557	0.001
10	0.010	0.013	0.199	0.003
11	0.011	0.011	0.831	0.000
12	0.011	0.015	0.017	0.004
13	0.001	0.003	0.238	0.002
14	0.023	0.028	0.024	0.005
15	0.001	0.004	0.143	0.003
16	0.000	0.004	0.059	0.004
17	0.027	0.027	0.995	0.000
18	0.022	0.036	0.000	0.014
19	0.003	0.007	0.233	0.004
20	0.127	0.152	0.000	0.025
21	0.008	0.009	0.666	0.001
22	0.050	0.067	0.000	0.017
23	0.004	0.005	0.432	0.001
24	0.000	0.009	0.000	0.009
25	0.010	0.035	0.000	0.025
26	0.027	0.041	0.000	0.014
27	0.002	0.009	0.003	0.007
28	0.037	0.039	0.228	0.002
29	0.001	0.002	0.345	0.001
30	0.007	0.016	0.015	0.009