



Research Report
ETS RR-11-34

**Recommendations for Conducting
Differential Item Functioning (DIF)
Analyses for Students With Disabilities
Based on Previous DIF Studies**

Heather Buzick

Elizabeth Stone

August 2011

**Recommendations for Conducting Differential Item Functioning (DIF) Analyses
for Students With Disabilities Based on Previous DIF Studies**

Heather Buzick and Elizabeth Stone
ETS, Princeton, New Jersey

August 2011

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

<http://www.ets.org/research/contact.html>

Technical Review Editor: Dan Eignor

Technical Reviewers: Neil Dorans and John Young

Copyright © 2011 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, and, LISTENING. LEARNING. LEADING. are registered trademarks of Educational Testing Service (ETS).



Abstract

The purpose of this study is to help ensure that strategies for differential item functioning (DIF) detection for students with disabilities are appropriate and lead to meaningful results. We surveyed existing DIF studies for students with disabilities and describe them in terms of study design, statistical approach, sample characteristics, and DIF results. Based on descriptive and graphical summaries of previous DIF studies, we make recommendations for future studies of DIF for students with disabilities.

Key words: differential item functioning, students with disabilities

Acknowledgments

The authors wish to thank Neil Dorans for his contribution to the DIF design framework.

Table of Contents

	Page
Overview.....	1
Standard Practice for Studying Differential Item Functioning (DIF).....	2
DIF and Students With Disabilities	4
Design Frameworks for Differential Item Functioning (DIF) Studies	6
Design 1	6
Design 2.....	6
Design 3	7
Method	8
Summary of Differential Item Functioning (DIF) Studies for Students with Disabilities	8
Recommendations for Studying Differential Item Functioning (DIF) for Students With Disabilities	14
References.....	18
Notes	23
Appendix.....	24

List of Tables

	Page
Table 1. Design 1	6
Table 2. Design 2	7
Table 3. Design 3	8

List of Figures

	Page
Figure 1. Percentage of low-, medium-, and high-difficulty items flagged for differential item functioning (DIF) when groups were tested under different conditions.	11
Figure 2. Percentage of low-, medium-, and high-difficulty items flagged for differential item functioning (DIF) when groups were tested under the same conditions.	11
Figure 3. Percentage of low-, medium-, and high-difficulty items flagged for differential item functioning (DIF) in comparisons involving a nonaccommodated, nondisabled group and accommodated students with disabilities.	12
Figure 4. Percentage of low-, medium-, and high-difficulty items flagged for differential item functioning (DIF) in comparisons involving nonaccommodated students with disabilities and accommodated students with disabilities.	12
Figure 5. Total percentage of items flagged for differential item functioning (DIF) plotted against the standardized mean difference between the empirical score distributions of the groups.	13
Figure 6. Total percentage of items flagged for differential item functioning (DIF) plotted against the number of items on the test.	14

Overview

Differential item functioning (DIF) refers to group differences in performance on a test item that cannot be explained by group differences in the construct targeted by the item (Crocker & Algina, 1986; Clauser & Mazor, 1998). Test items are identified as exhibiting DIF when, after matching examinee groups by a measure of ability, the performance of one group is significantly higher than the other group, on average. When DIF is found to occur, it means that a test item is measuring traits or abilities that are secondary to the targeted ability. For students with disabilities, such secondary traits could be a test taker's ability to access the math content in a word problem or the ability to respond to a computer-delivered constructed response item with a keyboard, for example. For such students, opportunity to learn the content may also be considered a secondary trait.

Secondary traits measured by items showing DIF may be relevant or irrelevant to the targeted ability. When test items measure secondary traits or abilities that are irrelevant to the intended measure for some groups, such items are considered biased. Item bias is one aspect of fairness in testing and test use (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999)). To ensure test fairness, DIF statistical methodology is used to empirically identify items that are performing differently across focal and reference groups after matching examinees based on ability, and human judgment is used to decide whether an item showing DIF is biased based on its characteristics (Zieky, 1993; Zumbo, 1999). When an item shows moderate to high levels of DIF, the item is typically reviewed by content experts. In the test development stage, an item showing DIF may either remain as is, be revised, or be deleted from the item pool. In an operational setting, an item showing DIF may be removed from the calculated test score depending on the results of the item review.

Over the last 5 years, there has been a substantial increase in the number of studies using DIF methods to compare students with and without disabilities. Since students with disabilities are not a homogeneous subpopulation (Johnstone, Thompson, Bottsford-Miller, & Thurlow, 2008), comparison groups must often be disaggregated based on specific disability subtypes. While small sample sizes had limited the number of DIF studies for students with disabilities historically, recent changes have provided opportunities to conduct item-level analyses and to make judgments about fairness for more specific disability subgroups. Such changes include

increased participation of students with disabilities in large-scale statewide assessments (Thurlow, Quenemoen, Altman, & Cuthbert, 2008) and postsecondary education (U.S. Department of Commerce, Bureau of the Census, 2004, 2006), federal requirements for ensuring sound technical quality of alternate assessments taken by some students with disabilities (United States Government Accountability Office, 2009), and novel approaches to item analysis for low incidence disability subtypes (described below).

Given that multiple statistical approaches are available to study DIF and that multiple decisions are made once DIF items are detected (Clauser & Mazor, 1998), we surveyed existing DIF studies for students with disabilities to inform recommendations for future studies on DIF for students with disabilities. The aim of the current synthesis is to help ensure that DIF studies for students with disabilities are appropriate and lead to meaningful results. The following sections provide an overview of standard practice for studying DIF, a description of the characteristics of students with disabilities that may impact DIF analyses, a summary of previous research, and recommendations for studying DIF for students with disabilities.

Standard Practice for Studying Differential Item Functioning (DIF)

Examining DIF is not simply algorithmic; rather, judgments need to be made at various steps in the process. Such decisions include (a) identifying the comparison groups, (b) choosing a matching criterion, (c) choosing a statistical approach, and (d) interpreting DIF results, including what to do with items showing DIF (Clauser & Mazor, 1998). However, some standard practices in DIF analyses may lead to more valid inferences. These include ensuring that scores on the matching criterion are reliable and valid (Clauser & Mazor, 1998), using sufficient sample sizes in the reference and focal groups (Zieky, 1993), obtaining a matching criterion from a standardized administration across comparison groups (Dorans & Holland, 1993), and examining focal and reference groups with similar ability distributions, particularly when methods of detecting DIF that are not based on item response theory (IRT) are used (e.g., Klockars & Lee, 2008; Mazor, Clauser, & Hambleton, 1992; Narayanan & Swaminathan, 1994, 1996).

Reliability and validity of matching criterion. In DIF analyses, the matching criterion should be a valid measure of the target ability measured by the items. Several decisions need to be made regarding the choice of matching criterion that may impact validity (Clauser & Mazor, 1998). When ability differences exist between the reference and focal groups, an unreliable criterion will lead to the most discriminating items being flagged for DIF (Clauser & Mazor,

1998). When an internal matching criterion is used, such as the total test score, the matching criterion should also be minimally impacted by DIF items (Clauser & Mazor, 1998). Purification of the matching criterion can be accomplished with statistical procedures (e.g., using an iterative Mantel-Haenszel [MH] procedure [Holland & Thayer, 1988], or by selecting the option using the SIBTEST software [Shealy & Stout, 1993]). Administering test items to the reference and focal groups under standardized testing conditions can also help ensure that the matching criterion is essentially free from DIF.

Standardized administration. Studies for DIF typically compare students under the same testing conditions (Zieky, 1993). This comparison helps to ensure that examinees from the reference and focal groups are appropriately matched when using an internal matching criterion, which is essential for obtaining valid DIF results (Clauser & Mazor, 1998). When the matching variable is based on incomparable measures, such as when the measurement conditions have been altered for one group or for some members of one or both groups, DIF techniques are not appropriate unless the measures are shown to be comparable.

Sample size. Sufficient sample sizes in both focal and reference groups are necessary in order to have enough power to detect differences in performance across groups matched on ability. Based on research by Narayanan and Swaminathan (1994) and Rogers and Swaminathan (1993), sample sizes of 200 to 250 per group will likely have enough power to detect DIF using non-IRT methods including MH (Holland & Thayer, 1988), logistic regression (Swaminathan & Rogers, 1990), and SIBTEST (Shealy & Stout, 1993). At one time, the practice at ETS when using MH was for the smaller group to comprise at least 100 examinees, with the total number of examinees equal to 500 or more for test development; for postadministration and prescore reporting, examinees in the smaller group must total at least 200 with a minimum of 600 combined examinees; and at least 500 people must be included in the smaller group postscore report when analyses are conducted on a group that has never been studied (Zieky, 1993). However, sample size requirements have varied over time (e.g., recent guidelines included a 300/700 rule) and across testing programs and purposes. IRT-based methods for detecting DIF generally require larger sample sizes in order to estimate model parameters for both the reference and focal groups (Clauser & Mazor, 1998). Alternative approaches, described below, have been used to study DIF for smaller groups, but the success of these methods has yet to be studied.

Differences in ability distributions. When the ability distributions of the reference and focal groups differ, the efficacy of the matching criterion and the results from DIF analyses can be impacted. Previous methodological work has evaluated the impact of differences in ability distributions on DIF results. Studies have found that large mean differences in ability distributions across groups are associated with decreased power to detect DIF using non-IRT methods (e.g., Klockars & Lee, 2008; Mazor, Clauser, & Hambleton, 1992; Narayanan & Swaminathan, 1994, 1996). In addition, the MH statistic and its modifications have been shown to have higher Type I error rates as ability distributions become more discrepant and discrimination parameters differ across groups (Fidalgo & Madeira, 2008).

DIF and Students With Disabilities

The feasibility of studying DIF for students with disabilities has improved as more students with disabilities are being assessed (e.g., on state criterion-referenced assessments, Center on Education Policy, 2009). However, complexities still exist. For example, some DIF studies for students with disabilities have had reference and focal groups tested under different conditions and different proficiency distributions that can influence the proportion and type of items that are flagged for DIF (Dorans & Holland, 1993; Fidalgo & Madeira, 2008; Sireci, 2009).

Some students with disabilities take assessments with accommodations. Testing accommodations are intended to remove barriers to accessing test items due to students' disabilities. That is, "the psychometric function of accommodations is to increase the validity of inferences about students with [disabilities] by offsetting specific disability-related, construct-irrelevant impediments to performance" (Koretz & Hamilton, 2006, p. 562). For example, a braille format test allows students who are blind to access items that would be unreadable in paper or computer format. In testing, the term *accommodations* typically refers to changes in the test that are not intended to alter the construct being measured. The term *modifications* refers to changes in the test that do alter the construct being measured. The designation of a test condition alteration as an accommodation or a modification may be based on policy or research and may change based on the purposes of the test. For example, state departments of education differ in their designation of the read-aloud accommodation on a reading assessment as an accommodation or a modification (Laitusis, 2008).

DIF analyses are routinely performed on groups with different measurement conditions because of accommodation use. However, using reference and focal groups that differ in whether they received testing accommodations may impact the validity of the matching criterion. When accommodation use itself alters most or all of the items on a test such that they become a measure of the students' ability to use the accommodation as well the targeted ability, then accommodation use will introduce error in the DIF results because an appropriate internal matching criterion will not be available.

Small sample sizes and nonoverlapping proficiency distributions are two common characteristics of data from students with disabilities that may influence the results from statistical analyses including DIF (Sireci, 2009). Historically, students with specific types of disabilities either have been excluded from DIF studies or have been included in analyses by aggregating students with different types of disabilities under an umbrella classification. This latter option is sometimes chosen because of small sample sizes in groups related to specific disability subtypes since small sample sizes can lead to low power to detect performance differences between groups. In addition, students with disabilities, particularly those with cognitive disabilities, tend to have lower performance levels than students without disabilities (e.g., Klein, Wiley, & Thurlow, 2006; Ysseldyke et al., 1998) due, in part, to disability classification criteria that include low achievement levels or abilities. As such, a focal group comprising students with disabilities may have a test score distribution that is positively skewed with a mean far below the reference group, which could lead to too many items, particularly easy items, being flagged for DIF (Sireci, 2009). Low test reliability for students with disabilities is also a concern, particularly for tests with a broader proficiency range, such as state accountability assessments (Koretz & Hamilton, 2006).

Several other characteristics of students with disabilities and assessments taken by some students with disabilities can impact DIF results. Such characteristics include (a) performance assessments with too few items or insufficient statistical properties; (b) complications from accommodation use such as discrepancies between assignment and use, database errors, or no information on accommodations available to the researcher; (c) low performance, which may lead to low precision of estimates if the test is linear and is targeted to an ability level far above the subgroup ability level; and (d) factors that may differentially impact the underlying ability distribution of subgroups (e.g., students with physical vs. cognitive disabilities, opportunity to

learn impacted by disability). In addition, classification of disability status is often inconsistent (Koretz & Hamilton, 2006) and there are numerous stages where errors in coding student characteristics in a database can occur, consequently contributing to inaccurate DIF results.

Design Frameworks for Differential Item Functioning (DIF) Studies

The following describes three different design options that have been used to study DIF. The subscript s denotes scores obtained from a standard administration. The subscript a denotes scores obtained under an accommodated administration.

Design 1

Design 1 (see Table 1) is a standard DIF design. This design can be used to determine whether DIF exists for students with disabilities relative to students without disabilities (i.e., Group 1 is students without disabilities and Group 2 is students with disabilities). No students receive accommodations under this design.¹ Studies we surveyed that included comparisons with Design 1 were Bolt and Ysseldyke (2006), Bennett, Rock, and Kaplan (1985), Cline, Stone, and Cook (2008), Engelhard (2009), Kato, Moen, and Thurlow (2009), and Steinberg, Cline, Ling, Cook, and Tognatta (2008).

Table 1

Design 1

	Item X_s	Matching variable (Y_s)
Group 1	@	@
Group 2	@	@

Note. Groups 1 and 2 receive item X_s and have the matching variable Y_s under standard conditions.

Design 2

Design 2 (see Table 2) is routinely used with standard DIF methods to study the impact of accommodation use (e.g., Bolt & Ysseldyke, 2006; Laitusis, Cook, & Aicher, 2004; Cohen, Gregg, & Deng, 2005; Finch, Barton, & Meyer, 2009; Ling & Stone, 2008; Stone, Cook, Laitusis, & Cline, 2010). In some situations, this design is the only feasible option (e.g., studying DIF for blind students tested with items delivered in Braille relative to sighted students). Design

2 is also used when DIF is performed on existing datasets in which students in one or both groups use accommodations but either they are different in the different groups (e.g., Bennett, Rock, & Kaplan, 1985; Laitusis, Maneckshana, & Monfils, 2009) or it is unknown whether students in either group received accommodations (e.g., Abedi, Leon, & Kao, 2008).

Table 2

Design 2

	Item X_s	Matching variable (Y_s)	Item X_a	Matching variable (Y_a)
Group 1	@	@		
Group 2			@	@

Note. Group 1 receives item X_s and matching variable Y_s under standard conditions; Group 2 receives item X_a and matching variable Y_a under accommodated conditions.

In most situations, Design 2 violates the assumptions of standard DIF analysis because there is no common matching variable (i.e., $Y_s \neq Y_a$) and evidence of DIF would likely be a function of differences between Y_s and Y_a . Were standard DIF procedures to be used with such a design, evidence that Y_s and Y_a measure the same thing should be provided.² Design 3 studies rely on the assumptions that accommodations are appropriately administered to those who need them and that they do not alter the construct being measured.³ Such assumptions should also be verified to provide evidence that the matching criterion, and consequently the DIF results, is not impacted by accommodation use.

Design 3

Design 3 (see Table 3) is ideal for examining the effects of an accommodation (or bundle of accommodations) that can be administered to both groups (e.g., a read-aloud accommodation). An example of this data collection design can be found in Laitusis (2010). Among the studies surveyed, Engelhard (2009) and Ling and Stone (2008) used this design. In this design, DIF procedures could be used separately for both the standard administration and the accommodated administration, and results compared to see if the change in measurement conditions associated with accommodation use alters the results of the DIF analysis.

Table 3***Design 3***

	Item X_s	Matching variable (Y_s)	Item X_a	Matching variable (Y_a)
Group 1	@	@	@	@
Group 2	@	@	@	@

Note. Group 1 receives both item X_s and matching variable Y_s under standard conditions and item X_a and matching variable Y_a under accommodated conditions; Group 2 receives both item X_s and matching variable Y_s under standard conditions and item X_a and matching variable Y_a under accommodated conditions.

Method

Existing DIF studies on students with disabilities were surveyed from among research on ETS testing programs and external research published in peer-reviewed journals. We collected information from the studies including choice of reference and focal groups, how accommodation use is treated, and the statistical method(s) used to conduct item-level analysis and address test fairness. In addition, we recorded item-level information including the number of DIF items, the magnitude of DIF, and the difficulty of the DIF items, when available. The number of items, the sample sizes for reference and focal groups, and the observed score distribution (i.e., mean and variance) were also recorded. We chose to focus on comparisons that resulted in evidence supporting DIF in order to obtain information on factors associated with finding DIF items for students with disabilities.

We summarized the studies by looking for trends in proportion of items being flagged for DIF based on choice of reference and focal groups, treatment of accommodation use, and type of disability (i.e., cognitive vs. physical).⁴ The studies are summarized below both descriptively and graphically. The graphical summary focuses on the relationship between the percentage of items flagged for DIF and item difficulty.

Summary of Differential Item Functioning (DIF) Studies for Students with Disabilities

We collected 17 unique studies on DIF for students with disabilities published between 1986 and 2010; among those, 9 were conducted by ETS researchers. The appendix contains information about the number of DIF comparisons, type of assessment, and studied disability

groups. The 17 studies comprised 123 separate DIF comparisons that resulted in finding items that exhibited DIF. Of the 123 comparisons, 72% used students without disabilities as the reference group; in the remaining comparisons, both reference and focal groups comprised students with disabilities. Slightly less than half of the comparisons involved studying the impact of accommodations, 28% did not involve accommodation use, and for the remaining 24% of the comparisons, the authors did not know whether or not students received accommodations. The ability of the focal and reference groups, based on disability type and accommodation status, was similar for 54% of the comparisons and different for 23%. For the remaining comparisons, information on the types of disabilities the students had was unavailable so it was unclear whether cognitive ability was similar or different across reference and focal groups.

Based on the observed score distribution, the mean score of the focal group was lower than the reference group for 66% of the comparisons, the means were similar for 18% of the comparisons, and the focal group had a higher mean than the reference group in 6% of the comparisons. All comparisons in which the observed score distributions were similar involved students with disabilities in both the reference and the focal groups. Seventy-four percent of the comparisons that used students without disabilities in the reference group had lower observed score means for the focal group, and among those comparisons, 48% comparisons involved studying the impact of accommodations. Of the comparisons in which students with disabilities were in both focal and reference groups, 32% had a lower observed score mean for the focal group. Among these comparisons, all but one involved studying the impact of accommodations.

The average sample size for the reference group across comparisons was 53,620 with a minimum of 92 and the median equal to 5,949.⁵ The focal group average was 5,419 with the median equal to 485 and a minimum of 74. For comparisons in which students with disabilities were in both reference and focal groups, the average sample size of the reference group was 2,495 and the average sample size for the focal groups was 665. Across all comparisons, the mean number of items was 53 with a minimum⁶ of 8 and a maximum of 75. Thirty-eight percent of the comparisons were carried out with the MH statistic. Other methods used were SIBTEST (10%), logistic regression (13%), and IRTLRDIF (Thissen, 2001; 13%). We found several novel approaches to studying DIF for students with disabilities. Johnstone, Thompson, Moen, Bolt, and Kato (2005) used a combination of item analyses including item ranks, item total correlation, and DIF with contingency tables and IRT-based methods. Engelhard (2009) framed DIF analyses in

terms of model-data fit and residual analyses. Sixty-two percent of the comparisons considered only uniform DIF, whereas the remaining studied both uniform and nonuniform DIF.

Trends in DIF flagging for the surveyed studies are shown in the following graphs.⁷ The graphs in Figures 1 to 4 display percentages of low-, medium-, and high-difficulty items flagged for DIF. In each graph, each comparison is represented by up to three points depending on whether it had a percentage greater than zero of items flagged for DIF in the respective difficulty strata. Note that most studies included more than one DIF comparison. Because the tests we collected are quite different (e.g., different grade levels, different content, and different item difficulty statistics used by the authors), we categorized items by relative difficulty within each test. This categorization was done using proportion correct in the reference group for a majority of the comparisons and a proxy for proportion correct (e.g., based on the IRT difficulty parameter) for the remaining comparisons. Items were sorted by relative difficulty and then divided into three categories based on where their difficulty statistic fell compared to other items in the test.

Figures 1 and 2 include DIF comparisons categorized by whether the reference group and focal group were tested under the same conditions. Groups that were tested under the same conditions were either both administered the assessment under standard conditions or both administered the assessment with the same accommodation(s). For a majority of the comparisons, testing conditions were clearly defined by the authors; however, there were a few comparisons for which testing conditions were unknown, or accommodated and nonaccommodated groups were combined. We excluded cases in which the accommodations were unknown or when accommodated and nonaccommodated groups were combined.

As shown in Figures 1 and 2, comparisons involving different administration conditions generally had higher percentages of DIF items than comparisons involving the same administration conditions. In addition, there appears to be a slight trend of higher percentages of easy items flagged for DIF when groups were tested under different conditions.

As discussed above, when tests are administered under different conditions due to offering accommodations to students with disabilities who need them, the influence on the quality of DIF results is unclear. Figures 3 and 4 display percentages of items flagged for DIF for comparisons that are distinguished by disability status and accommodation use. In Figure 3, comparisons involve a nondisabled group and a group of students with disabilities who received

accommodations are displayed. Figure 4 shows comparisons involving a nonaccommodated group and an accommodated group, with both groups comprised of students with disabilities.

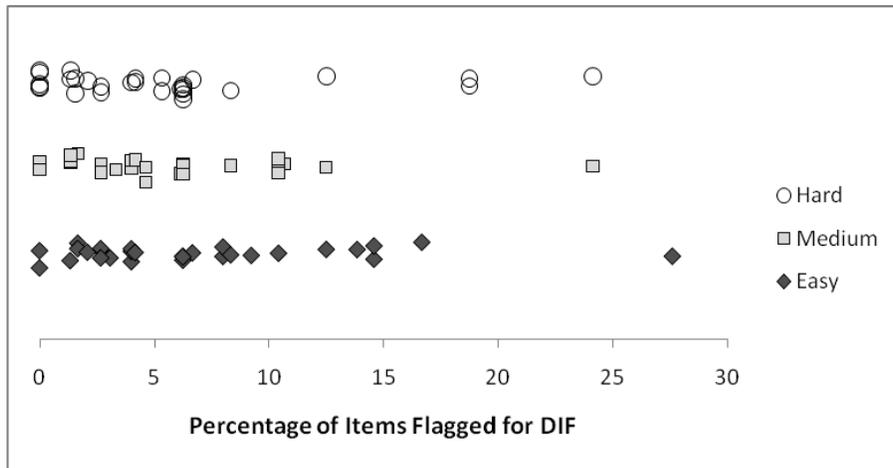


Figure 1. Percentage of low-, medium-, and high-difficulty items flagged for differential item functioning (DIF) when groups were tested under different conditions.

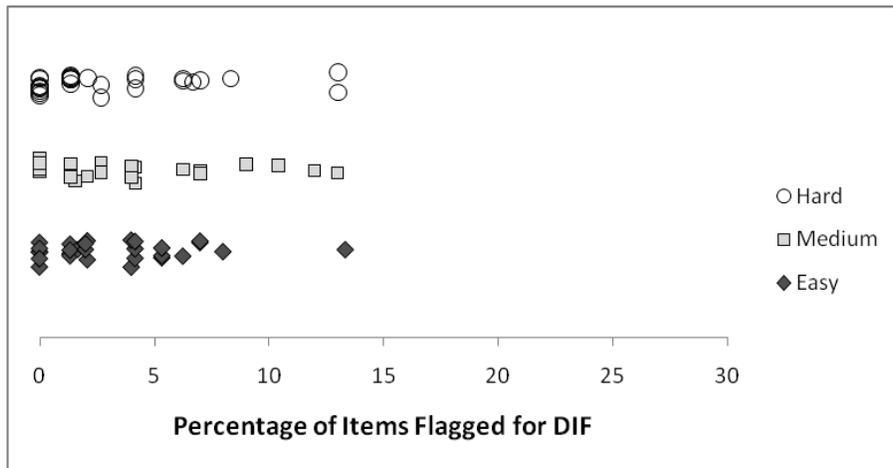


Figure 2. Percentage of low-, medium-, and high-difficulty items flagged for differential item functioning (DIF) when groups were tested under the same conditions.

Both Figures 3 and 4 show that there were higher percentages of easy and medium difficulty items flagged for DIF. The majority of comparisons in Figure 4 had relatively low percentages of items flagged for DIF. Relative to Figure 1, in Figure 3 there are fewer comparisons with higher percentages of item flagged for DIF and fewer comparisons with a high percentage of easy items flagged for DIF. This result is expected since the comparisons in

Figure 3 involve groups that should be most similar in terms of ability—nondisabled students and students with disabilities receiving accommodations. The comparisons in Figure 4 involve students with disabilities in both groups in an attempt to improve matching of students in the reference and focal groups by comparing groups with similar disabilities. The percentages of items flagged for DIF for these studies, which evaluate the impact of accommodations, are similar to those in Figure 1, which shows all comparisons under different testing conditions.

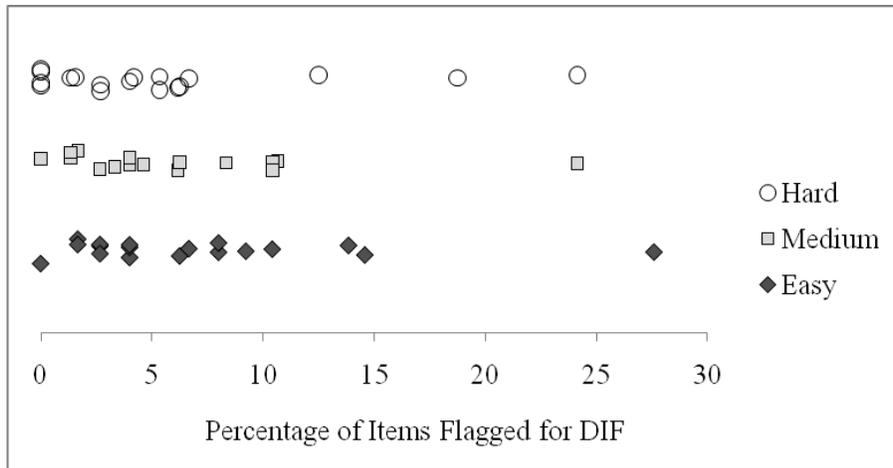


Figure 3. Percentage of low-, medium-, and high-difficulty items flagged for differential item functioning (DIF) in comparisons involving a nonaccommodated, nondisabled group and accommodated students with disabilities.

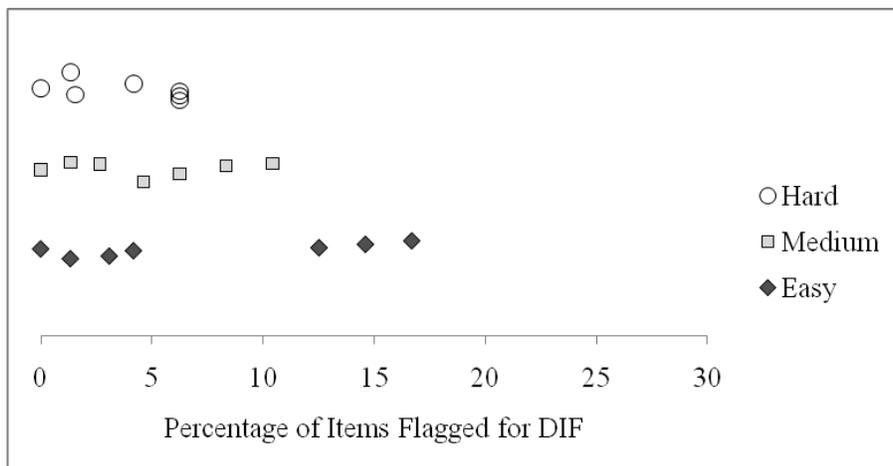


Figure 4. Percentage of low-, medium-, and high-difficulty items flagged for differential item functioning (DIF) in comparisons involving nonaccommodated students with disabilities and accommodated students with disabilities.

To further explore DIF results when substantial ability (or estimated ability) differences are present between groups, we plotted the percentage of items flagged for DIF versus a measure of the effect size for the score distribution difference between groups for a subset of the surveyed studies for which summary statistics were available. We used the standardized mean difference (SMD) to calculate Cohen's D, where SMD is computed as

$$SMD = \frac{\bar{X}_{Ref} - \bar{X}_{Foc}}{\sqrt{\frac{(n_{Ref}-1) \cdot Var_{Ref} + (n_{Foc}-1) \cdot Var_{Foc}}{n_{Ref} + n_{Foc} - 2}}}$$

Using the typical significance cut-off values of 0.2 (negligible), 0.5 (moderate), and 0.8 (important), Figure 5 shows that the majority of comparisons have at least a moderate difference in scores, with many comparisons involving an important or large score difference. Among comparisons in which the reference group has higher ability than the focal group (i.e., comparisons with positive Cohen's D values), there is a slight positive relationship ($r = .19$) between the size of the group ability difference and the percentage of DIF items flagged in the studies for which we were able to obtain summary statistics.

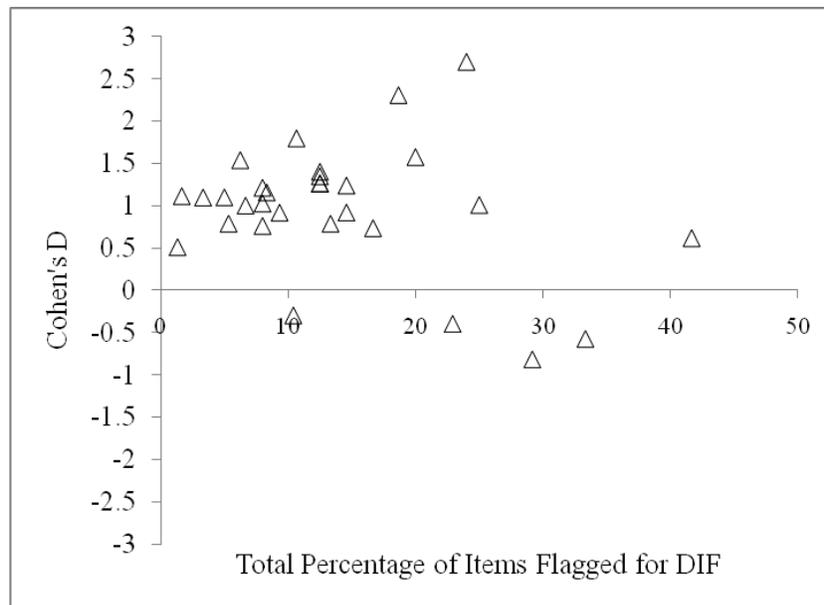


Figure 5. Total percentage of items flagged for differential item functioning (DIF) plotted against the standardized mean difference between the empirical score distributions of the groups.

Another concern discussed above involves the reliability of the matching criterion, which is often the total test score. Although it is not the only factor, the number of items strongly influences the reliability of the test (i.e., generally, adding more items to a test increases the reliability). Figure 6 shows the relationship between the number of items on the test and the percentages of items flagged for DIF for the reviewed studies. The correlation between the number of items and the percentage that were flagged for DIF was -0.34 in this group of comparisons. Because many of the comparisons involved tests with the same number of items, the numbers of items are jittered to allow for a clearer display in the graph. There is a general trend of a higher percentage of items flagged for DIF when fewer items are included on the test.

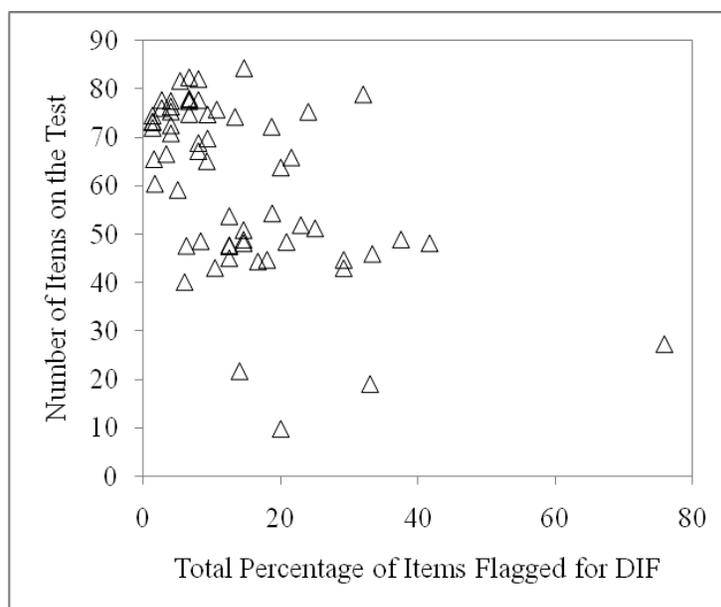


Figure 6. Total percentage of items flagged for differential item functioning (DIF) plotted against the number of items on the test.

Recommendations for Studying Differential Item Functioning (DIF) for Students With Disabilities

The information we obtained from the DIF studies surveyed provided some insight into our hypotheses. We found that, in general, comparisons involving focal and reference groups that were tested under the same conditions resulted in lower percentages of items being flagged for DIF relative to comparisons involving groups that were tested under different conditions. The summary also suggested that comparing students without disabilities to students with disabilities

who receive accommodation as well as comparing groups with similar disabilities that differ on accommodation use may result in meaningful comparisons. Prior research suggests that accommodations differ in their effectiveness; as such, some accommodations may be more appropriate to study with DIF methods than others. We recommend that DIF studies continue to compare students without disabilities to students with disabilities who receive accommodations that they need and that have been shown to be effective. However, we urge that caution be taken when using DIF as a tool to study the impact of accommodations using a non-experimentally designed study. If DIF analyses are to be used to study the impact of accommodations, we recommend that an external matching criterion be considered and that decisions based on the statistical results be supplemented by expert opinion or existing research on the efficacy of the specific accommodation, the impact of accommodation use on the appropriateness of the matching criterion, and the amount of construct-irrelevant variance expected to be introduced from the interaction between the item characteristics and the accommodation.

While we were unable to calculate the Type 1 error rate for the comparisons in this study since we did not know the truth, many of the comparisons involved observed score mean differences or groups that differed in their cognitive ability based on disability subtype and accommodation use. We found a slight trend of higher percentages of items being flagged for DIF as the ability distributions between the groups became more discrepant. Future DIF studies on students with disabilities should take into account the methodological literature and employ methods that are most robust to discrepant ability distributions (e.g., Fidalgo & Madeira, 2008; Klockars & Lee, 2008; Mazor, Clauser, & Hambleton, 1992; Narayanan & Swaminathan, 1994, 1996) when the groups of interest are not expected to perform similarly on the total test due to their disability. In addition, we caution against combining students with different disability subtypes, particularly those who differ in their impact on cognitive ability, and instead support the creation of separable, well-defined focal and reference groups that are defined by theoretically important research questions rather than sampling convenience.

Finally, when conducting DIF studies for students with disabilities on assessments with few items, we suggest ensuring that the test is sufficiently reliable and exhibits robust psychometric properties. Items found to exhibit DIF on such assessments should be carefully evaluated by content experts to ensure that results are due to legitimate causes of DIF rather than spurious statistical findings. If the technical quality of the assessment is in question due to the

small number of items or test format, then we recommend evaluating fairness with methods other than DIF analyses and content review, such as cognitive interviews, to understand whether or not items are functioning as intended.

Through our survey of DIF studies for students with disabilities, we found that studies varied widely in their design and analysis approach and in how the assumptions of the analysis procedures were treated. Consequently, DIF studies for this subpopulation are difficult to summarize, making it challenging to synthesize results in order to generalize inferences. The lack of uniformity in DIF studies for this subpopulation—and in general—highlights the importance of content experts in identifying the practical importance of DIF results and in determining whether items and tests are biased. Without the interpretation of results from content experts, the interpretation of results comes into question since they are supported by statistical analyses that are based on decisions at many phases of the research study that are not backed by methodological research.

The lack of uniformity also highlights the need for guidelines for conducting DIF studies in general and for specific subgroups. As noted in the section on standard DIF procedures, there are many aspects of DIF for which some rules of thumb or standard practices have been reported: reliability and validity of matching criterion, sample size, administration condition, and ability distribution. In many studies one or two of these aspects are of particular concern. However, all of these aspects are potentially called into question when evaluating DIF in comparisons involving students with disabilities. Until these guidelines are created, it is of the utmost importance that publications make clear the characteristics of their samples and the analysis so that research can be used to accumulate knowledge rather than exist in isolation. In some of the studies that we evaluated, this was not the case, and it was difficult to place the study results in context. For example, when groups with different accommodations are combined, or the testing conditions are unknown, the inability to separate out differential effects may be problematic. Similarly, grouping all students with physical disabilities into one focal group ignores the heterogeneous nature of the population of students with disabilities. In making inferences based on that focal group, for example, one would have to determine whether it makes sense to assume that students with visual, hearing, and motor skills difficulties experience similar obstacles during testing. This concern can be even greater when the heterogeneous focal group is loosely based on cognitive disabilities, which can manifest in very different response profiles.

Consideration must also be given to the methods used to undertake DIF analysis and whether the selected method is robust to violations of assumptions.

This study explored existing DIF studies on students with disabilities; while no statistical tests were performed, we summarized the existing studies in terms of technical characteristics and described trends in the percentage of items flagged for DIF. As such, this work provides a foundation for further evaluations of DIF studies for students with disabilities. Future research can build upon this work by directly evaluating the interaction between item characteristics, accommodation use, and students' disabilities, with the aim of understanding the most appropriate ways to evaluate and ensure fairness in testing students with disabilities.

References

- Abedi, J., Leon, S., & Kao, J. C. (2008). *Examining differential item functioning in reading assessments for students with disabilities* (CRESST Report No. 744). Los Angeles, CA: UCLA.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1985). *The psychometric characteristics of the SAT for nine handicapped groups* (ETS Research Report No. RR-85-49). Princeton, NJ: ETS.
- Bolt, S. E., & Ysseldyke, J. E. (2006). Comparing DIF across math and reading/language arts tests for students receiving a read-aloud accommodation. *Applied Measurement in Education, 19*, 329–355.
- Center on Education Policy. (November, 2009). *State test score trends through 2007-08: Has progress been made in raising achievement for students with disabilities?* Washington, DC: Author.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items: An NCME instructional module. *Educational Measurement: Issues & Practice, 17*, 31–44.
- Cline, F., Stone, E., & Cook, L. (2008, March). *An examination of differential item functioning on grade 5 math and science assessments for students with disabilities*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.
- Cohen, A. S., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities Research & Practice, 20*, 225–233.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Group.
- Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum.

- Engelhard, G., Jr. (2009). Using item response theory and model-data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement, 69*, 585–602.
- Fidalgo, A. M., & Madeira, J. M. (2008). Generalized Mantel-Haenszel methods for differential item functioning detection. *Educational and Psychological Measurement, 68*, 940–958.
- Finch, H., Barton, K., & Meyer, P. (2009). Differential item functioning analysis for accommodated versus nonaccommodated students. *Educational Assessment, 14*, 38–56.
- Holland, W. P., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Johnstone, C. J., Thompson, S. J., Bottsford-Miller, N. A., & Thurlow, M. (2008). Universal design and multimethod approaches to item review. *Educational Measurement: Issues and Practice, 27*, 25–36.
- 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). Retrieved from University of Minnesota, National Center on Educational Outcomes website: <http://education.umn.edu/NCEO/OnlinePubs/Technical41.htm>
- Kato, K., Moen, R., & Thurlow, M. (2009). Differential of a state reading assessment: Item functioning, distractor functioning, and omission frequency for disability categories. *Educational Measurement: Issues and Practice, 28*, 28–40.
- 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: University of Minnesota, National Center on Educational Outcomes.
- Klockars, A. J., & Lee, Y. (2008). Simulated tests of differential item functioning using SIBTEST with and without impact. *Journal of Educational Measurement, 45*, 271–285.
- Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: Praeger.
- Laitusis, C. C. (2008). State reading assessments and inclusion of students with dyslexia. *Perspectives on Language and Literacy, 34*, 31–33

- Laitusis, C. C. (2010). Examining the impact of audio presentation on tests of reading comprehension. *Applied Measurement in Education, 23*, 153–167.
- Laitusis, C. C., Cook, L. L., & Aicher, C. (2004, April). *Examining test items for students with disabilities by testing accommodation on assessments of English language arts*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.
- Laitusis, C. C., Maneckshana, B., & Monfils, L. (2009). Differential item functioning comparisons on a performance-based alternate assessment for students with severe cognitive impairments, autism, and orthopedic impairments. *Journal of Applied Testing Technology, (10)2*. Retrieved from http://data.memberclicks.com/site/atpu/Differential_Item_article_6.pdf
- Ling, G., & Stone, E. (2008, October). *DIF analysis for students with and without reading disabilities: Evaluating the impact of matching criterion*. Paper presented at the annual meeting of the Northeastern Educational Research Association, Rocky Hill, CT.
- Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*, 443–451.
- Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*, 315–328.
- Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*, 257–274.
- Rogers, H. J., & Swaminathan, H. (1993). A comparison of the logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105-116.
- Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/dif. *Psychometrika, 58*, 159–194.
- Sireci, S. G. (2009). No more excuses: New research on assessing students with disabilities. *Journal of Applied Testing Technology (special issue), 10*, 1–18. Retrieved from <http://data.memberclicks.com/site/atpu/Special%20issue%20article%205.pdf>

- Steinberg, J., Cline, F., Ling, G., Cook, L. L., & Tognatta, N. (2008). Examining the validity and fairness of a state standards-based assessment of English-language arts for deaf or hard of hearing students. *Journal of Applied Testing Technology*, *10*(2), Retrieved from <http://data.memberclicks.com/site/atpu/Special%20issue%20article%205.pdf>
- Stone, E., Cook, L. L., Laitusis, C. C., & Cline, F. (2010). Using differential item functioning to investigate the impact of testing accommodations on an English-language arts assessment for students who are blind or visually impaired. *Applied Measurement in Education*, *23*(2), 132–152.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370.
- Thissen, D. (2001). IRTLRDIF v. 2.0b: Software for the computation of the statistics involved in item-response theory likelihood-ratio tests for differential item functioning [Computer software]. Chapel Hill: University of North Carolina, L. L. Thurstone Psychometric Laboratory.
- Thurlow, M. L., Quenemoen, R., Altman, J. R., & Cuthbert, M. (2008). *Trends in the participation and performance of students with disabilities* (Technical Report 50). Minneapolis: University of Minnesota, National Center on Educational Outcomes.
- United States Department of Commerce, Bureau of the Census. (2004). *2004 American community survey*. Retrieved from <http://www.census.gov/acs/>
- United States Department of Commerce, Bureau of the Census. (2006). *2006 American community survey*. Retrieved from <http://www.census.gov/acs/>
- United States Government Accountability Office. (2009, September). *No child left behind act: Enhancements in the department of education's review process could improve state academic highlights of GAO-09-911, a report to the assessments*. Retrieved from <http://www.gao.gov/new.items/d09911.pdf>
- 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: University of Minnesota, National Center on Educational Outcomes.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–348). Hillsdale, NJ: Erlbaum.

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, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Notes

- ¹An equivalent standard design would have both groups receiving the same accommodation(s) under the same measurement conditions. None of the surveyed studies used such a design.
- ²Possible sources of evidence include multigroup factor analysis, review of item characteristics and their interactions with the accommodation(s), or results from experimentally designed research on the impact of specific accommodations.
- ³In many testing situations, it may be unclear from the data collected whether a test taker used an assigned accommodation for a particular studied item. In such cases, providing evidence on the comparability of the measurement conditions is even more complex.
- ⁴The proportion of items flagged for DIF may be a function of effect sizes used to flag items.
- ⁵The large average sample size was due to one study with a number of comparisons with over 400,000 students in the reference group.
- ⁶This assessment is a rubric-scored, on-demand performance assessment.
- ⁷The DIF comparisons relating to an alternative assessment were excluded from the graphical displays because both the student groups and the test administration differ greatly from the other studies.

Appendix

Differential Item Functioning (DIF) Studies Surveyed

Study	Number of DIF comparisons	Assessment, disability subtypes (if known)
Abedi, J., Leon, S., & Kao, J. C. (2008). <i>Examining differential item functioning in reading assessments for students with disabilities</i> (CRESST Report 744). Los Angeles, CA: UCLA.	16	High stakes reading assessment (Stanford 9 reading comprehension and word analysis), small and large states, grades 3 and 9
Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1985). <i>The psychometric characteristics of the SAT for nine handicapped groups</i> (ETS Research Report No. RR-85-49). Princeton, NJ: ETS.	3	SAT, hearing impairment, learning disabilities, visual impairment
Bennett, R. E., Rock, D. A., & Jirele, T. (1986). <i>The psychometric characteristics of the GRE General test for three handicapped groups</i> (ETS Research Report No. RR-86-6). Princeton, NJ: ETS.	1	GRE General, visual impairment
Bennett, R. E., Rock, D. A., & Novatkoski, I. (1989). Differential item functioning on the SAT-M Braille edition. <i>Journal of Educational Measurement</i> , 26(1), 67–79.	1	SAT Math, visual impairment
Bolt, S. E., & Ysseldyke, J. E. (2006). Comparing DIF across math and reading/language arts tests for students receiving a read-aloud accommodation. <i>Applied Measurement in Education</i> , 19, 329–355.	12	State general assessment, math and reading, elementary, middle, and high school
Bolt, S. E., & Ysseldyke, J. E. (2008). Accommodating students with disabilities in large-scale testing: A comparison of differential item functioning (DIF) identified across disability types. <i>Journal of Psychoeducational Assessment</i> , 26, 121–138.	4	State general assessment, math, grades 4 and 8, cognitive and physical disabilities

Study	Number of DIF comparisons	Assessment, disability subtypes (if known)
Cline, F., Stone, E., & Cook, L. (2008, March). <i>An examination of differential item functioning on grade 5 math and science assessments for students with disabilities</i> . Paper presented at the annual meeting of the American Educational Research Association, New York, NY.	7	Grade 5 math and science assessments, learning disabilities
Cohen, A. S., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. <i>Learning Disabilities Research & Practice</i> , 20, 225–233.	1	FCAT 9th grade math test, learning disabilities
Engelhard, Jr., G. (2009). Using item response theory and model-data fit to conceptualize differential item and person functioning for students with disabilities. <i>Educational and Psychological Measurement</i> , 69, 585–602.	3	Georgia Criterion Referenced Competency Test, geometry, grade 7
Finch, H., Barton, K., & Meyer, P. (2009). Differential item functioning analysis for accommodated versus nonaccommodated students. <i>Educational Assessment</i> , 14, 38–56.	12	Nationally normed achievement test for language and mathematics, grades 3–8
Kato, K., Moen, R. E., & Thurlow, M. L. (2009). Differentials of a state reading assessment: Item functioning, distractor functioning, and omission frequency for disability categories. <i>Educational Measurement: Issues and Practice</i> , 28, 28–40.	6	Minnesota 2003 reading assessment, grades 3 and 5, speech/language impairment, learning disabilities, emotional behavior disorders
Laitusis, C. C., Cook, L. L., & Aicher, C. (2004, April). <i>Examining test items for students with disabilities by testing accommodation on assessments of English language arts</i> . Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.	12	ELA assessments, grades 3 and 7, deaf or hard of hearing, learning disabilities

Study	Number of DIF comparisons	Assessment, disability subtypes (if known)
Laitusis, C. C., Maneckshana, B., & Monfils, L. (2009). Differential item functioning comparisons on a performance-based alternate assessment for students with severe cognitive impairments, autism, and orthopedic impairments. <i>Journal of Applied Testing Technology</i> , (10)2.	14	State alternate assessment, ELA and mathematics, level I-V, autism with severe cognitive impairment, orthopedic impairment with severe cognitive impairment
Ling, G., & Stone, E. (2008, October). <i>DIF analysis for students with and without reading disabilities: Evaluating the impact of matching criterion</i> . Paper presented at the annual meeting of the Northeastern Educational Research Association, Rocky Hill, CT.	16	Gates-MacGinitie Reading Test (comprehension subtest), two forms, grades 4 and 8, reading-based learning disabilities
Steinberg, J., Cline, F., Ling, G., Cook, L., & Tognatta, N. (2008). Examining the validity and fairness of a state standards-based assessment of English-language arts for deaf or hard of hearing students. <i>Journal of Applied Testing Technology</i> , 10(2).	7	ELA assessments, grades 4 and 8, deaf or hard of hearing, deaf, hard of hearing
Stone, E., Cook, L, Laitusis, C.C., & Cline, F. (2010). Using differential item functioning to investigate the impact of testing accommodations on an English-language arts assessment for students who are blind or visually impaired. <i>Applied Measurement in Education</i> , 23(2), 132–152.	4	ELA assessments, grade 4 and 8, visual impairment
Zebehazy, K. T. (2006). <i>Ability or access-ability: Test item functioning and accommodations for students with visual impairments on Pennsylvania's alternate assessment</i> (Doctoral dissertation). Retrieved from http://etd.library.pitt.edu/ETD/available/etd-12072006-104720/unrestricted/Zebehazy_KT2_etdPitt2006.pdf	4	2005 level A PA alternate assessment, math and reading, grades 3/4 and 7/8, visual impairment