Check for updates

Hammill Institute on Disabilities

The Journal of Special Education
45(1) 54–62
© Hammill Institute on Disabilities 2011
Reprints and permission: http://www.sagepub.com/journalsPermissions.nav
DOI: 10.1177/0022466909350224
http://journalofspecialeducation.sagepub.com
SAGE

Accommodations and Item-Level Analyses Using Mixture Differential Item Functioning Models

Stanley E. Scarpati[1], Craig S. Wells[1], Christine Lewis[1], and Stephen Jirka[2]

Abstract

The purpose of this study was to use differential item functioning (DIF) and latent mixture model analyses to explore factors that explain performance differences on a large-scale mathematics assessment between examinees allowed to use a calculator or who were afforded item presentation accommodations versus those who did not receive the same accommodations. Data from a state accountability assessment of mathematics for students in Grade 8 were analyzed. More than 73,000 students participated, of which 12,268 were students with disabilities (SWD) receiving test accommodations. DIF analyses detected performance differences between examinees without accommodations and those who used a calculator or those where the item presentation was altered. Latent performance class analyses revealed that performance differences were associated with item difficulty and abilty in addition to accommodation status. Results support validity studies that use mixture models that can consider context variables related to item type, academic skills, and accommodations.

Keywords

accommodations, differential item functioning, test item analysis

Accountability testing among all students, including students with disabilities (SWD), used for high-stakes policy and instructional decisions require that the test scores yield accurate indications of how much students have achieved. For SWD, test accommodations have become routine when on-demand tests are administered. Some SWD may be given more time to complete the test, some have portions of the test directions or test items read to them, or some may have the test translated into another form such as a Braille version. The fundamental basis for the accommodations is argued from a social validity perspective of equal access (e.g., Phillips, 1994) where the accommodation removes the barrier the disability may contribute to irrelevant variance of the construct being measure. Regardless of how the accommodation alters the way test items are presented or varies the way students are allowed to respond that go beyond the standardized procedures, accommodations are not intended to interfere with the measurement construct and the validity of the inferences made about what students know and what they can do. This practice is designed to level the playing field so that the test format or test administration conditions do not unduly prevent SWD from demonstrating their true knowledge, skills, and abilities.

The past decade has seen a significant body of research emerge that has attempted to tease apart the issues of construct irrelevant variance, accommodation type, and disability (see Sireci, Scarpati, & Li, 2005; Thompson, Blount, & Thurlow, 2002, for reviews). However, no clear markers have been established that delineate when an accommodation begins to alter the measurement construct. It may be argued that although an accommodation is used to remove barriers to equitable assessment for SWD they may also produce a new and unintended source of irrelevant variance that invalidates the inferences drawn from the test score (Sireci, 2005). An accommodation can only be judged effective and useful when making accountability policy and education program adjustments if the resulting test scores are accurate, valid indicators of achievement for the population of interest.

Test-Level Accommodations and the Interaction Hypothesis

The prevailing hypothesis is that an interaction exists between the accommodation provided to a SWD and the score that does not exist for test takers not receiving the

[1]University of Massachusetts at Amherst, MA, USA
[2] Pearson Publishing, San Antonio, TX, USA

Corresponding Author:
Stanley E. Scarpati, University of Massachusetts, Hills South, Amherst, MA 01002, USA
E-mail: scarpati@educ.umass.edu

accommodation. At the test level an *interaction hypothesis* or a *differential boost* has been proposed to justify the use of test accommodations and states that test accommodations will lead to improved test scores for students who need the accommodation but not for students who do not need the accommodation (Fuchs et al., 2000; Shepard, Taylor, & Betebenner, 1998; Zuriff, 2000). If, however, the interaction between accommodation condition (accommodated vs. standard test administration) and type of student (e.g., SWD vs. students without disabilities) with respect to test performance reveals that *both* types of students benefit from an accommodation, it is possible that the score for SWD might be invalidly inflated.

Reviews of empirical research, most of which focused on students with learning disabilities getting extra time or a read-aloud accommodation, that either directly or implicitly tested the interaction hypothesis indicate that the interaction effect has not been supported and that no unequivocal conclusions can be drawn regarding the effects of accommodations on test scores (see Johnstone, Altman, Thurlow, & Thompson, 2006; Sireci et al., 2005; Tindal & Fuchs, 2000; Zenitsky & Sireci, 2007, for reviews). Other studies point to the relative comparable performance between students with and without learning disabilities (no differential boost) on a state test where an oral administration did not alter the internal factor structure of the test (Chui & Pearson, 1999; Huynh & Barton, 2006). This would imply that the forms were comparable and valid comparisons and score interpretations could be made. This may be explained in part by the significant diversity of characteristics and abilities within groups of SWD, limited number of experimental studies where both SWD and students without disabilities receive accommodations, and the inconsistencies in which accommodations are designed and implemented (Sireci et al., 2005). Another explanation is that test-level performance differences (i.e., SWD vs. students without disabilities) fail to take into account the extraordinary influence item-level characteristics, item complexity and form (e.g., constructed response item; select-a-response item), item difficulty, and individual student ability may have on test performance as a function of the accommodation. For example, many students who have a learning disability when decoding text score more poorly on "story"-type math items than on routine calculation items (Fuchs & Fuchs, 2002). Similarly, Fletcher et al. (2006) found that when an oral administration (i.e., reading words and word parts) was matched to specific disabilities such as poor decoding skills of dyslexics on a reading test, performance improved substantially, although not for all poor readers. Although the interaction effect is appealing from a test-level analysis it is limited as to how it helps us understand how accommodations benefit SWD, support their use, improve test validity, and improve academic instruction. It is from this perspective that the present study was conducted by analyzing item differences between eighth-grade students with and without disabilities on a large-sample standardized statewide mathematics test.

The purpose of the present study was to compare item difficulty between SWDs who received a specific accommodation versus non-SWDs who did not receive an accommodation and to determine whether accommodation status was primarily responsible for any observed differences. Each item on a mathematical assessment was examined to determine whether it exhibited differential item functioning (DIF) between the groups. A latent mixture DIF model analysis (Cohen & Bolt, 2005; Rost, 1990) was used to explore possible factors that explain the presence of DIF between examinees allowed to use a calculator or who were afforded item presentation accommodations versus those who did not receive the same accommodations. The advantage of using the latent mixture model is that it allowed us to create subgroups of examinees of comparable ability based on item performance rather than accommodation status alone.

Item-Level Analyses of Accommodations in Mathematics

Research studies that consider the unique ways SWD perform on test items when accommodations are provided might produce more valid interpretations of achievement. Item-level analyses create an opportunity to better understand the role ability plays when accommodations are used to improve test performance for SWD. Comparability studies at the test level based on the interaction hypothesis assume that once the barrier presented by a test procedure is removed, comparable test performance would occur, but the research methodology typically leaves the ability parameter out of the model. Individual student characteristics (e.g., language skills, cognitive ability, math ability) vary sufficiently within groups of SWD to influence the interaction with a test item and ultimately interfere with accurate measurement of the construct. For example, linguistic complexity is often a concern in mathematics for SWD in the form of decoding skills, processing speed, and comprehension and can mask true math ability when calculations alone are the construct of interest in a math performance item. A read-aloud accommodation is a common adjustment to word-dense mathematics problems (e.g., more than 20 words) and has been shown to improve performance as compared to similar application items (Helwig, Rozek-Tedesco, & Tindal, 2002) but is more likely to benefit better readers with low math skills than less skilled readers with low math skills (Helwig, Rozek-Tedesco, Tindal, Heath, & Almond, 1999). Reading math problems to students with literacy disabilities apparently benefits their achievement on basic problem solving tasks but yields little benefit when more complex math items require computations that arise from defining and applying math concepts (e.g., Fuchs et al., 2000). Interpreting the variation of the effect of a read-aloud accommodation on math performance is difficult due in part to the lack of consistency in which the accommodation is implemented. In a similar fashion, the effect calculator use had on math test performance for SWD in Grades 4, 7, and 12 was equivocal, and effect sizes

did not reveal any clear patterns based on condition and accommodation (Engelhard, Fincher, & Domaleski, 2006). However, it has been noted that using a calculator to solve math problems that differ in complexity may be confounded by ability, and students may continue to apply simple keystrokes and basic algorithms associated with less complex items when attempting to solve items that require higher ability and advanced procedural knowledge. In this case, students' reliance on a calculator to help them solve items that are beyond their knowledge or ability level or inappropriate calculator use may obscure an accurate assessment of mathematics ability (Cohen & Kim, 1992).

Test- and item-level analyses of the effects of accommodations on the mathematics performance of SWD remain equivocal, at least to the degree that inconsistencies remain in implementation of the accommodation, who does and does not receive the accommodation, item type, measurement construct, and how the data are analyzed. Researchers argue that these inconsistencies threaten the validity of what we know about accommodations and their effect on the measurement construct and that improving the methodology at the risk of forsaking the construct will produce little to inform policy and practice (e.g., Ketterin-Geller, Yovanoff, & Tindal, 2007).

Test-level DIF studies with students with a variety of disabilities (e.g., learning disabilities, visual impairments) completing math or literacy tasks point to the inconsistency of the performance differences across items that may or may not be related to the accommodation. For example, these data suggest that certain accommodations might significantly affect performance on certain items for certain types of disabilities but not for other construct-related items for different disabilities, calling into question the appropriateness of the accommodation (e.g., Bolt, 2004; Bolt & Ysseldyke, 2006).

This study departed from current research that considers accommodation status (e.g., receiving or not receiving an accommodation) as the major source of variance between students on a high-stakes mathematics test. A differential boost (or not) was not of interest per se, as we focused on item-level performance differences between students, and data were produced that provided more depth to better understand the complexities of calculator use and test item presentation accommodations, item types, and math ability. The DIF procedures used in this study are similar to those used by Cohen, Gregg, and Deng (2005) in their study of an extended time accommodation used by 10th grade students taking a high stakes test in mathematics.

DIF

DIF occurs when the probability of answering an item correctly differs between examinees from separate groups with the same ability level. Differences in how groups of interest perform on test items can be detected using a DIF method that reveals a potential bias in how one group performs in comparison to another. A simple distinction here is that DIF is statistical and not social or evaluative where issues of fairness come into play. For our purposes, DIF is the simple observation that an item displays different statistical properties for different groups after controlling for differences in ability (mathematic skills) in the groups (e.g., matching on total score or proficiency). One group in this study is composed of students taking a high-stakes assessment in mathematics with accommodations (focal group) and the other group is composed of students taking the same test without accommodations (reference group). Practically, if DIF exists, it prompts an investigation into the nature of the construct for the focal group and a reestimation of respondents' measures.

There are several readily available statistics for detecting DIF in dichotomous items. DIF statistics may be classified as either model based or observed score (nonmodel) based. One of the more popular model-based methods is Lord's (1980) χ^2 DIF statistic. Lord presented a test statistic, distributed as χ^2, in which item parameter estimates calibrated for a reference (R) and focal (F) group are compared. Lord's calculation for the Rasch model is as follows:

$$\chi^2 = \frac{(b_{iR} - b_{iF})^2}{Var(b_{iR}) + Var(b_{iF})} \quad (1)$$

where b_{iR} and b_{iF} are the difficulty parameter estimates for item i for the reference and focal groups, respectively, and $Var(b_{iR})$ and $Var(b_{iF})$ represent the squared standard errors for b_{iR} and b_{iF}, respectively. The degrees of freedom are equal to the number of parameters being compared (e.g., df=1 for the Rasch model). In this study, a DIF analysis was performed between the accommodated and nonaccommodated examinees.

Mixture DIF Models

Whereas the previously described DIF analyses compared item performance between *manifest* groups (e.g., accommodated vs. nonaccommodated), it is possible to use mixture DIF models to identify *latent* groups of examinees for which an item (or set of items) is behaving differentially. In the standard DIF model, little is known about the participants who performed differently on various items because the manifest groups establish the analytic model. In other words, DIF is revealed but no information is provided about why it occurred. When using a mixture DIF model, we first define a model in which a specific set of items exhibit DIF with the assumption that there are latent groups of examinees for which the specified items behave differentially. The model is subsequently used to classify examinees based on the consistency with which their item responses conform to the DIF model. The advantage of mixture DIF models is that the final result helps researchers understand the causes of the DIF observed.

Method

In this study, we first examined items for DIF between students who received a specific type of accommodation versus those who did not receive an accommodation while taking a Grade 8 statewide mathematics assessment. Once the items that were functioning differentially were obtained, we then fit a mixture DIF model to define groups of students whose item responses were consistent with the pattern of accommodated-related DIF.

Data

The DIF and mixture DIF analyses were completed using data from a recent (2007) state accountability assessment of mathematics ability for students in Grade 8. The mathematics assessment consisted of 39 items, 34 of which were scored dichotomously. The 34 dichotomously scored items were used in the present study. More than 73,000 students participated in the assessment, of which 12,268 were SWD receiving various test accommodations. Item types were sorted into multiple choice, short answer, and open response formats. A multipe choice item, for example, would require a student to calculate the diagonal length (in meters) of a theater lobby rope where the length and width of the lobby are provided. Students select from four options, each specifiyng a range of lengths, of which only one contains the correct response. An example of an open response item would ask a student to convert the wingspan of a large jet airline from yards to feet. A diagram of the airliner including the wingspan dimension would be provided. An open response item would present an intial scenario and then ask a series of related follow-up questions based on the various skills and knowledge required. Students are also required to include an explanation of how they arrived at the answers. The accoutability assessment is designed as a criterion referenced test where items are arranged from less to more difficult. Allowable accommodations are specified according to a student's individual education plan and are not specified as such so that they are pegged to any one item type or level of diifculty or complexity.

We were interested in DIF for examinees who used a calculator and for examinees in which the test item presentation was altered from the standard protocol (see Table 1 for a list of allowable accommodations on how items can be presented). Therefore, we compared examinees without accommodations to those who used a calculator and to those where the item presentation was altered. For each DIF comparison, we randomly selected examinees from the nonaccommodated group to have equal sample sizes between the groups of interest.

Table 1. Accommodations on Test Item Presentation

Accommodation	Description
Familiar test administrator	Test administrator is familiar to the student.
Noise buffers	The student wears noise buffers, after test administration instructions have been read.
Magnification or overlays	Student uses magnifying equipment, enlargement devices, colored visual overlays, or specially tinted lenses.
Test directions	Test administrator clarifies general administration instructions. No portion of the test items or reading selections (including the introduction to reading selections) may be read or signed.
Large print	The student uses a large-print (18-point font) version of the test. All answers must be transcribed verbatim from the large-print answer booklet to the student's standard answer booklet.
Braille	The student uses a Braille version of the test. All answers must be transcribed verbatim from the Braille test to the student's standard answer booklet.
Place marker	Student uses a place marker.
Track test items	Test administrator assists the student in tracking test items (e.g., moving from one test question to the next) or by redirecting the student's attention to the test.
Amplification	Student uses sound amplification equipment.
Test administrator reads test aloud (except reading comprehension test)	Test administrator reads the composition writing prompt or the mathematics test items to the student.
Test administrator signs test (except reading comprehension test)	Test administrator signs the composition writing prompt or the mathematics test items to the student.
Electronic text reader (except reading comprehension test)	Student uses an electronic text reader for the mathematics test.

DIF Analyses

The Rasch item difficulty parameters were estimated for each group separately using the computer program BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003). The mean-sigma method (Marco, 1977) was used to place the item difficulty parameter estimates from the focal group (accommodation group) onto the scale of the reference group (nonaccommodated group). The Lord's χ^2 DIF statistic was implemented to test for a meaningful magnitude of DIF between the accommodated and nonaccommodated groups using a range-null hypothesis (Wells, Cohen, & Patton, 2008). Meaningful DIF was defined as the difference in the difficulty parameter values between the reference and focal groups that was greater than 0.2. This assured that the items flagged for DIF represented substantively interesting differences.

Once the DIF items were identified, the item parameter estimates for both groups were reestimated via a concurrent

calibration. The item difficulty parameter values for the non-DIF items were constrained to be equal between the reference and focal groups whereas the DIF items were freely estimated in both groups. The computer program MULTILOG (Thissen, 2003) was used to perform the concurrent calibration. The resulting item parameter estimates for both groups were used in the subsequent DIF mixture model analysis.

A two-group mixture Rasch model (MRM; Rost, 1990) was constructed using the item parameter estimates from the concurrent calibration. In other words, based on the accommodation-related DIF observed previously, an MRM was specified to identify examinees whose item responses were similar to the DIF related to the accommodation. Therefore, two classes of examinees were produced: Class 1 represented students who responded similarly to the nonaccommodated item parameter estimates, and Class 2 represented students who responded similarly to the accommodated group. If accommodation is the primary factor that defines the observed DIF, the classes should be defined by their accommodation status (i.e., Class 1 would be composed of nonaccommodated students whereas Class 2 would contain students who used an accommodation). The computer program WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003) was used to estimate the parameters for the two-group MRM.

Results

Calculator Versus No-Calculator Group

Table 2 reports the item difficulty parameter estimates for items that exhibited a meaningful magnitude of DIF between examinees who used a calculator versus those who did not have any accommodation.

Fourteen of the 34 items exhibited a meaningful magnitude of DIF. Of the 14 DIF items, 8 were easier for the accommodated group (smaller difficulty parameter estimates indicate that the item was easier for the respective group) and 6 were easier for the nonaccommodated group after controlling for ability differences. DIF items that favored the accommodated group measured number sense and operation (3 items); geometry (1 item); patterns, relations, and algebra (2 items); measurement (1 item); and data analysis, statistics, and probability (1 item). DIF items that favored the nonaccommodated group measured number sense and operation (4 items); data analysis, statistics, and probability (1 item); and patterns, relations, and algebra (1 item).

The number of examinees in the accommodated and nonaccommodated groups that were classified into Class 1 or Class 2 is reported in Table 3.

Because there were an equal number of examinees in the accommodated and nonaccommodated groups (i.e., $n = 6{,}140$ for each group), it was expected that the proportion of examinees in each class would be roughly 0.5 if the classes, as defined by the DIF items, corresponded to accommodation status. The mixture model resulted in proportions of roughly 0.47 in Class 1 and 0.53 in Class 2. Therefore, approximately 47% of the examinees exhibited item responses consistent with the nonaccommodated group, and roughly 53% exhibited item responses consistent with the accommodated group. Furthermore, 81% (4,949 of 6,140) of the accommodation group actually responded in a manner consistent with the students using the calculator, and 19% responded in a consistent manner as the nonaccommodated group.

Although the majority of students who used a calculator produced item responses that were consistent with the accommodation-related DIF (i.e., 81%), a nontrivial number of those students responded more like those who did not receive an accommodation. On further inspection of the two classes of students who used a calculator, it was apparent that there was a large difference in ability between the two groups. The average ability parameter estimate for the accommodated group whose responses were consistent with the nonaccommodated group was nearly 1 *SD* larger than their counterparts that were placed into Class 2 ($\bar{\theta} = 0.56$) versus ($\bar{\theta} = -0.64$). This may be due to an interaction between ability level and accommodation status. More specifically, although the accommodation appeared to differentially

Table 2. Accommodated Item Difficulty and Nonaccommodated Item Difficulty Calculator and Standard Administration

Item	Content	Accommodated item difficulty	Nonaccommodated item difficulty
2	NSO	–1.33	–0.80
4	NSO	–0.57	–0.29
6	G	–0.96	–0.36
10	PRA	–1.46	–1.16
11	NSO	–0.71	–0.38
15	NSO	–0.65	–1.17
17	PRA	–0.92	–0.41
18	NSO	–0.35	–0.63
20	NSO	–1.50	–1.86
26	DASP	–1.01	–1.33
27	PRA	–0.13	–0.59
29	M	–1.24	–0.79
30	DASP	–1.23	–0.91
31	NSO	–0.37	–0.80

Note: NSO = number sense and operations; G = geometry; PRA = patterns, relations, and algebra; DASP = data analysis, statistics, and probability; M = measurement.

Table 3. Mixture Model Group Assignment for Calculator Versus No-Calculator Groups

	Accommodation status		
Latent Class	No calculator	Calculator	Total
Class 1	4,527	1,191	5,718
Class 2	1,613	4,949	6,562
Total	6,140	6,140	12,280

influence item performance for students with lower math skills, it did not necessarily have the same effect on students with higher math skills. Therefore, although the observed DIF shown in Table 2 illustrates the effect of accommodation status on item performance controlling for overall ability level, the actual observed DIF may be due to a subset of the accommodated group defined by their ability level.

Item Presentation Accommodation Versus Nonaccommodated Group

Table 4 reports the item difficulty parameter estimates for items that exhibited a meaningful magnitude of DIF between examinees who had an accommodation related to item presentation versus those who did not have any accommodation.

Nine of the 34 items exhibited a meaningful magnitude of DIF. Of the 9 DIF items, 5 were easier for the accommodated group and 4 were easier for the nonaccommodated group. DIF items that favored the accommodated group measured number sense and operation (1 item); geometry (1 item); patterns, relations, and algebra (2 items); and measurement (1 item). DIF items that favored the nonaccommodated group measured number sense and operation (3 items) and patterns, relations, and algebra (1 item). Interestingly, these same items were a subset of the items from the previous analysis that were detected as DIF.

The number of examinees in the accommodated and nonaccommodated groups that were classified into Class 1 or Class 2 is reported in Table 5.

Because there were an equal number of examinees in the accommodated and nonaccommodated groups (i.e., n = 10,024 for each group), it is expected that the proportion of examinees in each class would be roughly 0.5 if the classes, as defined by the DIF items, corresponded to accommodation status. The mixture model resulted in proportions of roughly 0.56 in Class 1 and 0.44 in Class 2. Therefore, approximately 56% of the examinees exhibited item responses consistent with the nonaccommodated group and roughly 44% exhibited item responses consistent with the accommodated group. Furthermore, 64% (6,386 of 10,024) of the accommodation group actually responded in a manner consistent with the students receiving an accommodation, and 36% responded in a manner consistent with the nonaccommodated group.

Similar to the two-group mixture model solution based on calculator usage, Class 1 and Class 2 for the accommodation group exhibited considerably different math ability. The average ability parameter estimate for the students who received an accommodation was much larger ($\bar{\theta}$ = 0.63) than their counterparts who were placed into Class 2 ($\bar{\theta}$ = –0.60)—a 1 *SD* difference. Therefore, it appears that accommodated-related DIF is being primarily defined based on examinees from the accommodation group with moderate to lower ability. Students who received an accommodation but who had more advanced math skills responded similarly to the nonaccommodated group.

Table 4. Accommodated Item and Nonaccommodated Item Difficulty Presentation and Standard Administration

Item	Content	Accommodated item difficulty	Nonaccommodated item difficulty
2	NSO	–1.13	–0.85
6	G	–0.91	–0.39
10	PRA	–1.52	–1.21
15	NSO	–0.79	–1.19
17	PRA	–0.75	–0.49
20	NSO	–1.62	–1.94
27	PRA	–0.26	–0.65
29	M	–1.24	–0.80
31	NSO	–0.47	–0.86

Note: NSO = number sense and operations; G = geometry; PRA = patterns, relations, and algebra; M = measurement.

Table 5. Group Assignment Based on Mixture Model for Item Presentation Accommodation

	Accommodation status		
Latent class	Nonaccommodated	Accommodated	Total
Class 1	7,520	3,638	11,158
Class 2	2,504	6,386	8,890
Total	10,024	10,024	20,048

Discussion

The manifest group and mixture DIF analyses provided a unique understanding of the interplay between students who received accommodations and those who did not on the mathematics assessment used in the study. Modern interpretations of DIF provide a framework to consider how differences in test structure, (i.e., characteristics associated with items or testing situation) not relevant to ability should be interpreted. The intent of accommodations is to provide opportunities for SWD to demonstrate what they know and what they can do without being penalized by the structure of standard test protocols without invalidating inferences drawn from the test scores. As we discussed earlier, the equivocal results of test-level comparisons between student test performance under conditions of accommodations versus nonaccommodations relies on condition status to detect whether SWD improve over nonaccommodated students. Condition status is also used to inform decisions about the validity of score inferences raising questions about invalidly inflated scores for SWD. It is apparent that test-level comparisons are limited and to expect more than condition-related descriptive information, even when students with and without disabilities both take a test under standard and accommodated conditions, is unwarranted. The advantage of the design used in this study was the ability to examine the effects of calculator use and item presentation accommodations (grouped) on item-level performance without being constrained by accommodation status to make performance comparisons.

The interaction between item characteristics and when SWD used a calculator became clear when the item types and measurement constructs shifted and calculator use differentially influenced performance. First, it was significant that 14 of the 34 total items exhibited a nonnegligible magnitude DIF (>0.20 difference) and that students using calculators fared better on easier items whereas the nonaccommodated group fared better on more difficult items. Yet proportionally, accommodation status did not predict class membership equally as nearly 20% of the students using calculators scored in ways consistent with students not using calculators. This is a significant finding and draws attention to item difficulty and math skills and knowledge that play along when calculators are offered as an accommodation. Higher ability parameter estimates were clearly on the side of students using calculators who performed consistently like those not using calculators and unveils much about skill level, disability, and the degree to which the calculator accommodation removes the barrier posed by the disability. In this case, test performance may have been less influenced by any construct irrelevance caused by the item features or format and more of a factor related to math ability. Interestingly, similar findings for a subset of the DIF items found for calculator use were also revealed when item presentation was used as an accommodation. Ability level again appeared to more likely separate students receiving the accommodation, as the DIF associated with accommodation use was based on students with moderate to low ability. Higher ability scores for students receiving the item presentation accommodations were similar to the nonaccommodated group. The change across test item DIF for accommodated and nonaccommodated test takers was not sufficiently explained by accommodation status, and accommodation use can be considered as necessary but unlikely to explain these differences. Math ability, and ability that is not consistent across items that change features and constructs, is again vital to scoring well on the large-scale test used in this investigation.

Our findings are consistent with other item-level accommodation studies that have investigated single accommodations and their effect on specific types of items (e.g., Fuchs et al., 2000; Helwig et al., 2002). When ability levels vary, the influence of the accommodation varies as well and apparently is more likely to improve the performance of lower skilled students on items with fewer features (less complex) and that are less difficult. For example, as noted earlier in our review, less skilled readers with learning disabilities benefited when language-complex math items were read to them. In this study we categorized students into Class 1 and Class 2 based on total score DIF and uncovered that calculator use was more effective for items that were easier (from a criterion-referenced perspective) and featured basic arithmetic skills, number sense, and simple application requirements. Calculators did not benefit students on more difficult items and items that featured abstract thinking, symbolic manipulation, and the application of mathematical concepts unless the students using the calculators had higher math ability. Our analyses also pointed to the reality that accommodation group membership offered little to our understanding of performance differences and that when accommodation status was treated as a secondary nuisance dimension (Ackerman, 1992), ability and item type were more clearly linked to performance and not to whether students used calculators or had the items presented to them in an alternative fashion. Other mixture DIF studies that treated gender and ethnicity group membership as nuisance dimensions also found minimal association between group membership and DIF (e.g., Cohen &Bolt, 2005). From our results, with support from similar analyses, accommodations are access points that mediate the test performance of SWD, and the score differences are more associated with mathematical skills.

Making valid inferences from large-scale test results when nonstandard procedures are allowed for some test takers has extraordinary importance to educational policy and classroom practice. Removing construct-irrelevant variance through accommodations is appropriate, reasonable, and equitable for giving SWD access to the test items but can distort how the accommodations affect performance. Ability (i.e., skills, knowledge, abstract reasoning) can be widely distributed across item types and difficulty level that makes it unclear when an accommodation distorts the measurement construct or when it simply removes the access barrier presented by a disability. Knowing when accommodations interact with test item features requires a level of precision that research can provide if test performance can be analyzed with item features used as part of the analysis. Questions aligned with the nature of the instruction and curriculum experiences of SWD can be answered by methods such as the mixture DIF approach used in this study. Modern interpretations of DIF provide a framework to consider how differences in test structure not relevant to ability may be interpreted. A more practical way of asking the same question is: What situational characteristics play a role when DIF is revealed? The multidimensional nature of this question argues for DIF analyses that go well beyond the manifest group level. This question may also bring into the discussion dimensions associated with test accommodations not typically included in a test-level DIF that may influence student performance. Most notable is the relationship between the amount and type of classroom instruction relevant to the construct of interest and the types of accommodations allowable on the large-scale test that are routine during classroom instruction and are vital to understanding performance differences. We might also be interested in the curriculum available to students, the research evidence that supports using the curriculum, and the level of teacher experience in implementing the curriculum. Validity then becomes an interpretation of student performance that is more informed at the test item level and less likely to be based on test-level inferences according to

accommodation status. Latent or mixture DIF models are not restricted by identifying groups before determining performance differences not associated with ability and are more effective in considering additional contextual variables (Zumbo, 2007; Zumbo & Gelin, 2005). Classroom instruction and other context variables must be taken into account when SWD are allowed to use various accommodations that link what takes place in the daily school routine with testing. Ketterin-Geller et al. (2007) posit that disability classifications per se should play a less prominent role when accommodations are assigned to students and that functional skills are a better way to link classroom-based accommodations with test-based accommodations.

The fundamental precept that accommodations level the playing field when SWD take on-demand tests should be revisited. Indeed, issues of equity through reasonable adjustments to standard test procedures are ensured and provide comparable access opportunities to a fair and just evaluation for all students. But the notion of a level playing field most likely ends after access to a test (and the items) is acquired via accommodations, and validity research should frame and test access questions unique to test-level comparability studies. We should not expect these studies to yield data that go beyond addressing access, access-related issues, and fundamental test structure and other factor-analytic concerns. Ultimately, the playing field is the classroom but high-quality instruction is the game. Validity and accommodation research will advance when classroom instructional routines, student ability, and test item types are considered equal players in the evaluation model.

Declaration of Conflicting Interests

The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The authors received no financial support for the research and/or authorship of this article.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*, 67–91.

Bolt, S. E. (2004, April). *Using DIF analyses to examine several commonly-held beliefs about testing accommodations for students with disabilities*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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.

Chui, C. W. T., & Pearson, P. D. (1999, June). *Synthesizing the effects of test accommodations for special education and limited English proficient students*. Paper presented at the National Conference on Large Scale Assessment, Snowbird, UT.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of DIF. *Journal of Educational Measurement*, *42*, 133–148.

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.

Cohen, A. S., & Kim, H. (1992). Detecting calculator effects on item performance. *Applied Measurement in Education*, *5*, 303–320.

Engelhard, G., Fincher, M., & Domaleski, C. S. (2006). *Examining the reading and mathematics performance of students with disabilities under modified conditions: The Georgia Department of Education Modification Research Study: Differential item functioning based on Rasch analyses*. Atlanta: Georgia Department of Education.

Fletcher, J. M., Francis, D. J., Boudousquie, A., Copeland, K., Young, V., & Kalinowski, S., et al (2006). Effects of accommodations on high-stakes testing for students with reading disabilities. *Exceptional Children*, *72*, 136–150.

Fuchs, L. S., & Fuchs, D. (2002). Mathematics problem-solving profiles of students with mathematics disabilities with and without comorbid reading disabilities. *Journal of Learning Disabilities*, *36*, 563–573.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C., Brinkley, E., & Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children*, *67*, 67–81.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C. L., & Karns, K. M. (2000). Supplementing teacher judgments of mathematics test accommodations with objective data, *School Psychology Review*,*29*, 65–85.

Helwig, R., Rozek-Tedesco, M. A., & Tindal, G. (2002). An oral versus standard administration of a large-scale mathematics test. *Journal of Special Education*, *36*, 39–47.

Helwig, R., Rozek-Tedesco, M. A., Tindal, G., Heath, B., & Almond, P. J. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth grade students. *Journal of Educational Research*, *93*, 113–125.

Huynh, H., & Barton, K. E. (2006). Performance of students with disabilities under regular and oral administrations of a high-stakes reading examination. *Applied Measurement in Education*, *19*, 21–39.

Johnstone, C. J., Altman, J., Thurlow, M. L., & Thompson, S. J. (2006). *A summary of research on the effects of test accommodations: 2002 through 2004* (Tech. Rep. No. 45). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Ketterlin-Geller, L. R., Yovanoff, P., & Tindal, G. (2007). Developing a new paradigm for conducting research on accommodations in mathematics testing. *Exceptional Children*, *73*, 331–347.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems.. *Journal of Educational Measurement*, *14*, 139–160.

Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education*, *7*, 93–120.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.

Shepard, L., Taylor, G., & Betebenner, D. (1998). *Inclusion of limited-English-proficient students in Rhode Island's Grade 4 mathematics performance assessment*. Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Sireci, S. G. (2005). Unlabeling the disabled: A psychometric perspective on flagging scores from accommodated test administrations. *Educational Researcher*, *24*, 3–12.

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction effect. *Review of Education Research*, *75*, 457–490.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). WinBugs (Version 1.4) [Computer software]. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

Thissen, D. (2003). MULTILOG for Windows (Version 7) [Computer software]. Chicago: Scientific Software.

Thompson, S., Blount, A., & Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001* (Tech. Rep. No. 34). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved January 2003 from http://education.umn.edu/NCEO/OnlinePubs/Technical34.htm

Tindal, G., & Fuchs, L. (2000). *A summary of research on test changes: An empirical basis for defining accommodations*, Lexington: University of Kentucky Mid-south Regional Resource Center, Interdisciplinary Human Development Institute.

Wells, C. S., Cohen, A. S., & Patton, J. (2008, March). *A range-null hypothesis approach for testing DIF under the Rasch model*. Paper presented at the annual conference for the National Council on Measurement in Education, New York.

Zenisky, A. L., & Sireci, S. G. (2007). *A summary of the research on the effects of test accommodations: 2005-2006* (Tech. Rep. No. 47). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG 3.0 [Computer software]. Lincolnwood, IL: Scientific Software International.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*, 223–233.

Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated (mediated) test item bias. *Journal of Educational Research and Policy Studies*, *5*, 1–23.

Zuriff, G. E. (2000). Extra examination time for students with learning disabilities: An examination of the maximum potential thesis. *Applied Measurement in Education*, *13*, 99–117.

About the Authors

Stanley E. Scarpati, EdD, is an associate professor in the Special Education Program at the University of Massachusetts at Amherst. His current research interests are test accommodations for students with disabilities and social perception of students with learning disabilities.

Craig S. Wells, PhD, is an assistant professor in the Research and Evaluation Methods Program at the University of Massachusetts at Amherst. His current interests include educational measurement, item response theory, and differential item functioning.

Christine Lewis, MEd, is a graduate student in the Research and Evaluation Methods Program at the University of Massachusetts at Amherst. Her current interests include differential item functioning, policy implications of large-scale assessments, and equity issues in educational testing.

Stephen Jirka, MA, is an Associate Research Scientist at Pearson Publishing. His current interests include differential item functioning, external validation of test scores, and item response theory.