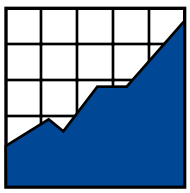
Synthesis Report 65

Item-level Effects of the Read-aloud Accommodation for Students with Reading Disabilities



NATIONAL CENTER ON EDUCATIONAL OUTCOMES

In collaboration with:

Council of Chief State School Officers (CCSSO)

National Association of State Directors of Special Education (NASDSE)

Synthesis Report 65

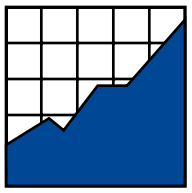
Item-level Effects of the Read-aloud Accommodation for Students with Reading Disabilities

Sara E. Bolt
Michigan State University

Martha L. Thurlow
National Center on Educational Outcomes

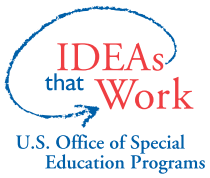
September 2006

Bolt, S. E., & Thurlow, M. L. (2006). *Item-level effects of the read-aloud accommodation for students with reading disabilities* (Synthesis Report 65). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.



NATIONAL CENTER ON EDUCATIONAL OUTCOMES

The Center is supported through a Cooperative Agreement (#H326G050007) with the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. This report was completed under the Center's previous Cooperative Agreement (#H326G000001) with the Office of Special Education Programs The Center is affiliated with the Institute on Community Integration at the College of Education and Human Development, University of Minnesota. Opinions expressed herein do not necessarily reflect those of the U.S. Department of Education or Offices within it.



NCEO Core Staff

Deb A. Albus
Manuel T. Barrera
Christopher J. Johnstone
Jane L. Krentz
Kristi K. Liu
Ross E. Moen
Michael L. Moore
Rachel F. Quenemoen
Dorene L. Scott
Karen Evans Stout
Martha L. Thurlow, Director

National Center on Educational Outcomes
University of Minnesota • 350 Elliott Hall
75 East River Road • Minneapolis, MN 55455
Phone 612/626-1530 • Fax 612/624-0879
http://www.nceo.info

Executive Summary

Research support for providing a read-aloud accommodation (i.e., having an individual read test items and directions aloud) to students with disabilities has been somewhat limited, particularly when merely examining effects of the accommodation on overall test scores for general groups of students with disabilities. We examined data on accommodated and non-accommodated performances of students with specific reading disabilities on various math test items anticipated to be highly sensitive to accommodation effects. Analyses were conducted across three consecutive years of data from an elementary and middle school statewide assessment program. Within the fourth grade dataset, items classified as reading-hard (RH) and those classified as mathematically easy but difficult to read (ME/RH) were positively affected by the accommodation. Marginally significant findings were obtained for the ME/RH item set at the eighth grade level. Limitations of the study, as well as implications, are discussed.

Table of Contents

Overview

A significant challenge associated with the development and implementation of K-12 statewide assessment and accountability systems has been determining how to most effectively include students with disabilities. Students with disabilities have been excluded from testing (Erickson, Thurlow, & Ysseldyke, 1996), from documents that report student performance (Klein, Wiley, & Thurlow, 2006), and from accountability formulas used to determine consequences for schools (Bolt, Krentz, & Thurlow, 2002). When students are excluded, it follows that these students may not benefit from the intended positive consequences of standards-based educational reform. School administrative decisions about the delegation of resources are not likely to target services to students with disabilities if their progress does not count.

One practice that has been associated with higher participation rates of students with disabilities in statewide assessment programs is the provision of testing accommodations (Olson & Goldstein, 1996). Many students with disabilities may struggle to access test content unless accommodations are provided. In other words, extraneous features of test presentation (e.g., excessive, complex text displayed in print format alone) may hinder them from demonstrating what they know and can do with respect to what the test is intended to measure (e.g., math skills, social studies, science knowledge, etc.). Accommodations are changes in standard test administration procedures that are intended to remove the associated construct-irrelevant variance (i.e., variance associated with extraneous features of test administration) for students with disabilities (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000). For example, students with visual impairments may not be able to demonstrate their achievement in a particular content area if they are tested using standard print. Provision of a large print or braille edition of the test may enable students with visual impairments to more accurately show what they know and can do. Although provision of accommodations to students with such visual or hearing impairments is often considered reasonable, accommodations for students with specific learning disabilities tend to be considered more controversial (Bolt, 2004). Federal legislation requires that "appropriate" accommodations be provided to students with disabilities to participate in state- and district-wide assessment; however, limited guidance is provided within this legislation about how to determine whether an accommodation is appropriate.

In determining whether accommodations are appropriate, Tindal (1998) suggests that it can be helpful to understand test alterations in terms of their placement on a continuum of validity. On this continuum, test changes are considered "accommodations" if they allow for measurement of the same construct as that measured among other students. Test changes are considered "modifications" if they more substantially alter the construct being measured. According to Hollenbeck (2002), who summarized the perspectives of various researchers on test alteration classification, accommodations (i.e., appropriate test alterations) are those that (a) involve a change to standard procedures that promote access for students to demonstrate knowledge,

(b) do not change the assessment construct, and (c) provide differentiated access. Although research investigations are accumulating to determine whether various test changes meet these criteria, there remains a lack of consistent findings across studies (Sireci, Scarpati, & Li, 2005; Thompson, Blount, & Thurlow, 2002).

Despite concerns about whether certain test alterations meet the criteria for being valid accommodations, they tend to be provided to a majority of students with disabilities. In a recent examination of public accountability reports across states, a research team found that across the states that reported the number of accommodated students with disabilities (which included between 29 and 35 states, depending on the grade level), an average of 61 to 66 percent of students with disabilities received accommodations (Thurlow, Moen, Altman, 2006). Unfortunately, research investigations of the extent to which accommodations are valid have failed to offer a clear answer as to whether this widespread provision of accommodations is justified.

Without conclusive empirical support for accommodations, policymakers and practitioners are forced to consult general testing guidelines for best practice in making testing accommodation decisions. In 1999, the most recent version of the testing standards were developed and published by a joint committee selected to represent the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). The published document (AERA, APA, & NCME, 1999) included the following statement about testing students with disabilities:

> In testing individuals with disabilities, test developers, test administrators, and test users should take steps to ensure that the test score inferences accurately reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement. (p. 118)

It is further indicated in these standards that those who make decisions about testing accommodations should be knowledgeable about the effects of disability status on test performance, and that modified tests should be pilot tested to students with the disability under consideration to investigate the validity of measurement with the suggested changes. Given the numerous possible test changes and disability types, meeting this standard will require considerable investigation. At this point, it seems most efficient to focus on those accommodations that are provided frequently to students (Bolt & Thurlow, 2004) and on those students with disabilities that are most common and closely tied to the accommodation under consideration.

The read-aloud accommodation is one of the most commonly allowed accommodations on statewide math assessments (Clapper, Morse, Lazarus, Thompson, & Thurlow, 2005). At last count, there were 47 states that allowed statewide math tests to be read aloud to students (Clapper et al., 2005). Many students with disabilities are currently being provided this accommodation on statewide tests. Tests designed to measure math achievement typically include

word problems that may be difficult for students with specific learning disabilities in reading to understand, particularly if the items are presented in print alone. It has been suggested that reading requirements on tests that are not intended to measure reading skills may increase the amount of construct-irrelevant difficulty for students with reading difficulties (Messick, 1995). By allowing a teacher to read the test out loud, providing an audio-cassette or video version of the test, or offering a computerized screen-reader, it is anticipated that students with reading difficulties can more effectively demonstrate their math knowledge.

Because reading aloud test items often represents a change in standardized test administration procedures, it is important to verify that this does result in more valid measurement of the intended test construct. The test standards clearly indicate the need to document the effects of any changes to standardized test administration procedures (AERA, APA, & NCME, 1999). One team of researchers has indicated that accommodations are provided to many students who derive no substantial benefit from them (Fuchs et al., 2000). Given that additional resources are often needed to provide this accommodation, it is important to investigate whether the accommodation truly facilitates better access to the test.

Several different methods have been used to examine the validity of the read-aloud accommodation on non-reading tests. Three of these methods (differential boost, measurement comparability, and item-level effects) as well as results of studies that have used these methods are described in the following sections.

Differential Boost

Accommodations are typically provided to offset construct-irrelevant difficulty that is assumed to be present solely for students with disabilities. The source of this difficulty is believed to be an interaction of the students' disability and certain aspects of standard test administration procedures. If accommodations are to be offered only to students with disabilities, which is the case in many states (Clapper et al., 2005), it is important to demonstrate that they differentially affect the performance of students with disabilities when compared to students without disabilities. If both students with and without disabilities benefit, it might be argued that it is unfair to provide the accommodation only to students with disabilities. Phillips (1994) suggested that a valid accommodation should improve scores for students with disabilities, and have no effect on the scores of students without disabilities. Others have argued along similar lines, using the term "differential boost" to describe how a valid accommodation should have a more positive impact on scores for students with disabilities than students without disabilities (Fuchs et al., 2000).

Several investigations of the read-aloud accommodation have examined whether an interaction effect exists such that students with disabilities derive substantially greater benefit from

the accommodation than students without disabilities (McKevitt & Elliott, 2003; Fuchs et al., 2000; Kosciolek & Ysseldyke, 2000). Many research teams have specifically investigated the effects of the read-aloud accommodation on math tests. Some of these teams have shown that students with disabilities benefitted more from the read-aloud accommodation on math tests than students without disabilities (Burch, 2002; Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998; Weston, 1999). However, other studies failed to show such an effect (Meloy, Deville, & Frisbie, 2002; Schulte, Elliott, & Kratochwill, 2001).

It is important to note several potential limitations of using the differential boost approach to answer questions about the validity of testing accommodations. First of all, the differential boost approach assumes that increased access to test content provided by the accommodation will be evident through a significant improvement in overall test score. It can sometimes be the case that the accommodation facilitates test access, but that this access does not result in an improvement in overall test score. It could be the case that a read-aloud accommodation truly allows a student to more effectively engage in the intended math-related activities during testing, but that the student has not acquired the math skills that are being tested, and therefore the positive effect of the accommodation is not evident from an analysis of improvement in overall test score alone. Furthermore, it can be particularly difficult to detect positive effects of a read-aloud accommodation on math tests when a student has failed to receive the accommodation during instruction. In this case, lack of familiarity may hinder adequate use of the accommodation in the testing context. It may also limit the student's ability to access instruction, such that they fail to learn what is tested and therefore can not show that the accommodation is helpful based on the items presented on the test. If researchers fail to consider these possibilities when making judgments about the validity of accommodations, accommodations may be unfairly denied to students who would actually benefit from receiving them during both instruction and testing.

Additionally, differential boost approaches that use group designs may fail to recognize the positive effects that an accommodation has for select subsets of students with disabilities. Elbaum, Arquelles, Campbell, and Saleh (2004) found that there was wide variation in the benefits derived by students with disabilities from a student-read-aloud accommodation. To gain a better understanding of when accommodation provision may be needed and allow for testing that is valid, it seems necessary to examine more closely the effects for particular groups of students with disabilities that would likely benefit, and find ways to more specifically detect whether the accommodation improves test access apart from a mere examination of changes in overall test score.

Measurement Comparability

A second method that has been used to study the validity of the read-aloud accommodation involves an investigation of measurement comparability across accommodated and non-accom-

modated test administrations. If an accommodation successfully removes construct-irrelevant variance associated with a disability, it follows that certain test measurement characteristics should be similar for test administrations among accommodated students with disabilities and non-accommodated students without disabilities. Comparing factor structures across accommodated and non-accommodated tests can help in determining whether the corresponding test conditions help to measure the same general constructs, and is a method that has been recently used to study the validity of testing accommodations (Huynh & Barton, 2006; Pomplun & Omar, 2000). Another method for examining measurement comparability across testing conditions is analysis of differential item functioning (DIF) (Bolt & Ysseldyke, in press). When DIF is identified for a target group of students, there is concern that the item may be biased for that group. Although some researchers who have used these methods have indicated a high degree of measurement comparability across math tests that were read aloud to students with disabilities and math tests that were not read aloud to students without disabilities (Lewis, Green, & Miller, 1999; Pomplun & Omar, 2000), other similar investigators have suggested that the accommodation does not substantially improve measurement comparability for students with reading disabilities (Bielinski, Thurlow, Ysseldyke, Friedebach, & Friedebach, 2001; Bolt & Bielinski, 2002).

Item-level Effects of the Read-aloud Accommodation

In an effort to more carefully investigate how the read-aloud accommodation impacts measurement of tested skills, researchers have examined item-level effects of the read-aloud accommodation (Fuchs et al., 2000; Helwig, Rozek-Tedesco, Tindal, Heath, & Almond, 1999). If the read-aloud accommodation is serving to remove extraneous sources of reading difficulty for students such that they can demonstrate true math competence, one would anticipate that a read-aloud accommodation would have a greater impact on items with significant reading requirements. In line with this hypothesis, Fuchs et al. (2000) found that students with disabilities benefited significantly more than students without disabilities from the read-aloud accommodation on math problem solving items, but not on other problems.

Such item-level studies of testing accommodation effects can help to improve sensitivity. A high-level of sensitivity is considered necessary because of the possibility that students have not acquired the math skills necessary to demonstrate that the accommodation improves their ability to demonstrate math knowledge on a test. Failure to acquire the tested math skills may have a variety of potential causes, one of which may include failure to access effective math instruction due to a disability. Although data are available that indicate accommodations are commonly provided to students for statewide tests (Thurlow et al., 2006), it remains unclear whether students receive the same accommodations during instruction. If the test accommodation being investigated was not adequately provided during instruction, a lack of positive accommodation

effects on a test may be related to lack of instructional access rather than due to failure of the accommodation to improve access to the content being tested. For example, if students are not receiving the necessary accommodations during instruction, it may be difficult for them to obtain the skills and knowledge that are tested. In this case, they may not be able to demonstrate on a test that the accommodation truly helps them access the test content because they do not have the tested math skills to perform better on the test under the accommodated condition.

Analysis of item level effects can help identify whether the read-aloud accommodation is helpful on items measuring skills to which students have most likely had prior instructional access (i.e., items that measure skills covered early in typical math curricula, and skills to which the students are more likely to have had repeated prior exposure). If item-level studies are not conducted, students may in the future be denied accommodations in both testing and instructional settings that in fact *do* help them better access instruction and better demonstrate knowledge on tests. A summary of findings from prior item-level analysis studies is provided in the following section.

Helwig et al. (1999) classified math items according to reading complexity and examined the performance of sixth-grade students across accommodated and non-accommodated conditions to determine whether effects of a video read-aloud accommodation varied according to the reading demands of test items. In their study, a literature review was first conducted to identify variables associated with reading complexity. The variables that were then addressed by the authors included number of words, number of syllables, number of difficult words in the item, number of multi-syllable words, number of verbs, number of mathematics vocabulary words, and passage complexity as measured in T-units, which are described as "complete expressions of thought that can stand alone in grammatically correct fashion" (Helwig et al., 1999, p. 118).

Helwig et al. (1999) also created student groups based on separate measures of oral reading fluency (ORF) and math achievement. Subgroups included: (1) high math, (2) low math, (3) high ORF, (4) low ORF, (5) high ORF-high math, (6) high ORF-low math, (7) medium ORF-high math, (8) low ORF-high math, (9) low ORF-low math, and (10) all students. Student disability status was not indicated in this study. Correlations were calculated for the relationship between each of the reading complexity variables and performance differences between standard and accommodated test administrations within each student group. Only two of these correlations were significant: verb count and performance differences between standard and accommodated conditions for students in the low ORF group ($r = .25$, $p < .05$), and verb count and performance differences between standard and accommodated conditions for students in the low ORF-high math group ($r = .29$, $p < .05$). Overall, there appeared to be only very weak relationships among various measures of reading complexity and the effect of the accommodation.

Helwig et al. (1999) also examined differences in the percentage of students scoring correctly across standard and video versions within the various student groups on six math items that were

classified as complex in reading. Items classified as complex reading items were those with 40 or more words, three or more difficult words, and five or more verbs. For two of the items, a significantly greater proportion of students scored correctly on the accommodated version than on the non-accommodated version across several student groups. These student groups included the high math group, low ORF-high math group, and all students. Within the low ORF-high math group, a greater proportion of students scored correctly on one additional complex reading item under the accommodated condition. The authors concluded that students may need to have a certain level of mathematical competence in order to benefit from the accommodation.

Another recent study similarly indicated that level of competence in the tested skill area played a role in the extent to which accommodations improved test score (Elliott & Marquart, 2004). In this study, extended time was found to influence scores particularly for students at-risk for math problems. The authors indicate ceiling effects may have contributed to their findings. This seems to suggest that additional attention is needed to examine effects in the context of item difficulty level in order to know whether accommodations are effective for groups of students, rather than relying on overall test score to provide an indication of effects.

Helwig, Rozek-Tedesco, and Tindal (2002) conducted a study similar to that of Helwig et al. (1999), in which they investigated the effects of a video read-aloud accommodation on math items classified as having demanding reading requirements, this time examining differential effects for students without disabilities and a combined group of students with either an identified reading disability or those who were deemed eligible for receiving this accommodation based on teacher discretion. Both elementary and middle school students participated in the study by taking the test with and without the accommodation. Results of the Analysis of Variance (ANOVA) did not support the hypothesis that the video read-aloud accommodation had a significantly greater impact on difficult reading items for students with reading difficulties than students without such difficulties. The authors noted that the reading complexity of the items present in the 2002 study was somewhat lower than that of the items present in the Helwig et al. (1999) study, which may have accounted for the different results.

Based on this review of current accommodation research methods and findings, it remains questionable whether accommodations are effective in removing construct-irrelevant variance associated with reading difficulties on math tests for students with disabilities. Various methods have been used to study accommodation effects; research findings have varied, even when the same research approach has been used. It has been argued that an examination of test score change across accommodated and non-accommodated conditions is not sufficient evidence to determine whether an accommodation is effective, given that many students who receive this accommodation on a math test may not have the math skills necessary to demonstrate the true effects of the accommodation. Helwig et al.'s (1999) finding that students with higher math skills tended to benefit more from the accommodation on complex reading items appears to

support this notion. In the current study, we did a similar analysis of the reading complexity of math items, in order to identify whether the read-aloud accommodation particularly improved performance for students with reading-related disabilities on items classified as difficult to read. In addition, we classified items according to the difficulty of the math involved. This classification scheme allowed us to examine the effects of the read-aloud accommodation for students according to both the reading and math complexity of the items. Through an investigation of accommodation effects on both easy and difficult math items with complex reading, we sought answers to the following research questions:

1) Do students with reading disabilities who receive a read-aloud accommodation perform better than those not receiving a read-aloud accommodation on items classified as difficult to read, after controlling for overall test performance?

2) Do students with reading disabilities who receive a read-aloud accommodation perform better than those not receiving a read-aloud accommodation on items classified as both difficult to read and mathematically easy, after controlling for overall test performance?

This study represents an extension and improvement of past research in this area given that it (a) focuses on a select group of students with disabilities that are expected to benefit from the read-aloud accommodation (i.e., students with learning disabilities and an individualized education program (IEP) intended to address reading skill deficits), (b) examines effects for items that are expected to be particularly sensitive to improvements for students with disabilities rather than relying solely on effects on overall test score, and (c) involves analysis of true testing data in which consequences for test performance were present (i.e., school ratings were being publicly reported and were based on student performance on this test), which can be considered to potentially enhance the ecological validity of the study.

Method

Participants

Data from three consecutive annual administrations of the math section of a statewide assessment program for fourth and eighth grade students were made available to the researchers for analysis. Students participated in this assessment in one of the following three ways: by taking the regular assessment, the regular assessment with accommodations, or an alternate assessment. IEP teams determined how individual students with disabilities participated and which testing accommodations each student received. These teams typically consisted of teachers and school support personnel. Parents and students were also urged to participate in this decision-making process. Student performance on the statewide assessment was used to determine school accountability ratings across all three years; no student-level consequences were attached to test performance.

For the analysis, we limited all of our analyses to examine the performance of students who were receiving special education services (i.e., had an IEP) who were coded as having "learning disability" as their primary disability, and "reading" as the primary area of their IEP. This represents a much more targeted group of students who are likely to benefit from the accommodation than has been the focus of past investigations of accommodations. It is possible that students had additional disabilities or IEP areas; however, they all were documented to have particular difficulties with reading, as well as what can be considered a "mild disability." In order to investigate the effects of the read-aloud accommodation, our primary analysis involved the creation of the following two groups.

Group A: Students who did not receive any accommodations, or received merely setting (small group or individual administration) and/or timing accommodations ($N_{fourth} = 431$, $N_{eighth} = 720$).

Group B: Students who received the read-aloud accommodation, and may have additionally received small group and extended time accommodations ($N_{fourth} = 1406$, $N_{eighth} = 1878$).

It was considered necessary to include students receiving setting and timing accommodations because of the way the read-aloud accommodation is typically provided in practice. Rarely does it make sense to provide the read-aloud accommodation without allowing additional time (given that it takes extra time to read material out loud and for students to request portions to be re-read as necessary), or without a setting accommodation (small group administration) given that it may be disruptive to a large-group of students who can read the test on their own. Although one could study the read-aloud accommodation in isolation, it would not represent how students typically are provided this accommodation in practice, and results would therefore not generalize to real-world situations. At the same time, it is important to understand and try to separate out any effect that might result simply from extended time. We therefore ran secondary analyses in which students who received extended time and setting were compared to those who received no accommodations, and an analysis in which students who received only extended time were compared to those who received no accommodations. As for the primary analyses, we targeted only students with learning disabilities who had reading listed as their primary IEP area.

Item Classification

We selected only multiple choice math items for this analysis; the same multiple-choice items were administered to students across the three consecutive years of data that were analyzed. Thirty-two multiple-choice items were included for the fourth grade analysis; thirty-one multiple items were included for the eighth grade analysis. In order to identify items anticipated to be particularly sensitive to the read-aloud accommodation, we first classified items according to reading complexity. Items containing only numbers and mathematical symbols and no reading content (i.e., computation-only items; four such items for fourth grade, three such items for

eighth grade) were eliminated prior to our classification, although performance on these items was included in determining the total score, which was used as the covariate in the analysis. Items that were identified as not requiring reading skills to correctly answer were also eliminated from the analysis, even though text was present in these items. These items, however, were also included in the total score calculation that was used as a covariate in the analysis.

Reading. For the remaining test items, the number of words, syllables, multi-syllable words, and verbs were tallied. These particular item characteristics were identified by Helwig et al. (1999) as corresponding to the reading difficulty of math items. Words identified within the response choices were included in the counts. Arabic numbers identified within the stem of the items were counted as one word and one syllable; numbers that were isolated in the responses without surrounding text were not included in these counts. Section directions were not included, given that they were deemed unnecessary for correctly solving any of the individual items. Similarly, words located on graphs and figures were minimal, and were typically found in the stem of the corresponding items, and so were not included in the counts.

Values for each of the four reading complexity variables were standardized across items; for each item, a sum of these standard scores was calculated. The five items with the greatest sum of standard scores were classified as reading-hard (RH). The five items having the lowest sum of standard scores were classified as reading-easy (RE). RH items contained more than 25 words, more than 30 syllables, more than eight multi-syllable words, and most contained four or more verbs in the fourth grade test. RH items for the eighth grade test contained more than 50 words, more than 80 syllables, more than eight multi-syllable words, and four or more verbs. RE items contained less than 15 words, less than 20 syllables, less than five multi-syllable words, and less than four verbs for the fourth grade test. RE items for eighth grade test contained less than 14 words, less than 16 syllables, less than four multi-syllable words, and all contained just one verb.

Math. Next, items were classified as either math-hard (MH) or math-easy (ME). The National Council of Teachers of Mathematics' (NCTM's) *Principles and Standards for School Mathematics* indicates that addition and subtraction are typically covered in the early elementary grades, whereas multiplication, division, fractions, and whole number computation estimation are covered in greater depth in the later elementary school years (NCTM, 2000). We used these standards as a framework for classifying items within the fourth grade dataset as MH or ME, although we did include relatively simple multiplication items as ME items. Consequently, items that required single addition, subtraction, and multiplication computations were classified ME; those involving basic math vocabulary (e.g., odd vs. even numbers), knowledge of number, and basic number sentences were also considered ME on the fourth-grade test. Those items that involved division, fractions, estimation, and multiple operations were considered MH for the fourth-grade test. Three items that did not fit either of these classifications and appeared to test spatial reasoning skills were not included in this classification within the fourth grade test.

Although the NCTM standards were consulted for categorizing the eighth grade items, the standards provided somewhat less guidance for determining which items to classify as easy and hard given that most skills that were tested are described within the NCTM standards for grades 6-8, with no further specification given as to when they are typically taught. However, it was assumed that items that involved use of more complicated forms of algebra, probability, and multiple operations are typically more difficult than items that involve single operations and simple math vocabulary (e.g., which numbers are prime, knowledge of place value to the millions place, converting inches to feet, etc.). This assumption guided our classification of eighth grade items according to difficulty level.

Fourth and eighth-grade items were then cross-classified as MH/RH, MH/RE, ME/RH, and ME/RE using the sum of standard scores for reading complexity described in the previous section. The MH/RE and ME/RE item sets contained items for which the sum of standard scores for reading complexity was at or below the average. The MH/RH and ME/RH groups contained items for which the sum of standard scores for reading complexity was above the average. Each of these item sets contained a total of five items. Several items were not included in any of these four sets, in order to allow for sets with an equal number of items (i.e., five items).

Student performance on items that were included in the two mathematically easy groups (ME/RE and ME/RH) and the two mathematically hard groups (MH/RE and MH/RH) were examined to verify that our categorization scheme for math difficulty aligned with actual student performance on these items, based on item difficulties presented in the technical manual for the test examined. Our item difficulty categorization aligned well with item difficulty levels provided within the technical manual for the test. For the fourth grade test, the average item-difficulties for the items in the ME/RE and ME/RH groups were .75 and .74 and the average item difficulties for the items in the MH/RE and MH/RH groups were .45 and .47 respectively. For the eighth grade test, the average item-difficulties for the items in the ME/RE and ME/RH groups were .64 and .71 and the average item difficulties for the items in the MH/RE and MH/RH groups were .35 and .41 respectively. This information, along with our data on how students within the dataset analyzed performed supported our classification of items according to math difficulty.

Data Analysis

Analysis of Covariance (ANCOVA) was used to analyze group differences in performance on the item sets for students with reading disabilities receiving (Group B) and not receiving (Group A) the read-aloud accommodation, as well as for the secondary analyses of those receiving extended time without the read-aloud accommodation. In order to control for existing group differences in math performance, we used overall performance on the test as a covariate in the analysis. Levene's test was used to examine the homogeneity of variance prior to conducting each of the ANCOVA analyses. Because we anticipated specific effects on just two of the six

item groups (i.e., RH and ME/RH), we considered it appropriate to maintain a Type I error rate of $\alpha = .05$ for each test, despite the fact that multiple statistical tests were conducted. Sato (1996) describes the various merits and drawbacks of controlling for Type I error when multiple tests are conducted, and we concluded that failure to detect important effects of an accommodation could lead to reduced test access for students with disabilities in the future, which would be a potentially worse error to make than what would result if we found an accommodation effect by mistake.

Results

Demographic information for both student groups is provided in Table 1. Groups A and B were similar with respect to gender and race, although for the eighth grade comparison, Group B (read-aloud group) was comprised of slightly fewer non-white students than the corresponding Group A (non-read-aloud group). Information on math performance on the entire test (not just on the subset of multiple-choice items included in the analyses) for Groups A and B is also provided in Table 1. In both the fourth and eighth grade datasets, Group A (non-read-aloud group) tended to perform better on the math test than Group B (read-aloud group).

Table 1. Demographic Characteristics for Groups

	4th Grade		8th Grade	
	No Read-aloud (A)	Read-aloud (B)	No Read-aloud (A)	Read-aloud (B)
N	431	1406	720	1878
Percent female	34.3	33.1	37.2	33.8
Percent non-white	19.8	22.9	26.1	20.8
Percent received timing accommodation	9.5	64.5	11.3	53.7
Percent received setting accommodation	22.7	94.7	31.0	94.3
Statewide assessment mean percentile (standard deviation)	43.9 (26.4)	40.2 (24.5)	34.4 (23.5)	28.9 (24.2)

Fourth grade group mean scores for each of the item sets are presented in Table 2, in addition to mean scores that were adjusted for overall student performance on the multiple-choice section of the test. The corresponding values for the eighth grade groups are presented in Table 3. The ANCOVA results for each item set are presented in Table 4.

Table 2. Fourth Grade Mean and Adjusted Mean Group Performance on Item Sets

	Mean (sd)		Adjusted Mean	
	No Read-aloud (A)	Read-aloud (B)	No Read-aloud (A)	Read-aloud (B)
Total Score	21.7 (6.2)	21.0 (5.9)	-	-
Math Hard/Reading Hard	2.6 (1.5)	2.4 (1.4)	2.5	2.5
Math Hard/Reading Easy	2.5 (1.4)	2.3 (1.4)	2.4	2.4
Math Easy/Reading Hard**	3.9 (1.2)	3.9 (1.1)	3.8	4.0
Math Easy/Reading Easy	4.0 (1.2)	4.0 (1.2)	3.9	4.0
Reading Easy**	2.9 (1.4)	2.7(1.4)	2.9	2.7
Reading Hard*	3.6 (1.3)	3.6 (1.2)	3.5	3.6

*Adjusted mean difference between groups is significant at $p < .05$, two-tailed.
**Mean difference between groups is significant at $p < .01$, two-tailed.

Table 3. Eighth Grade Mean and Adjusted Mean Group Performance on Item Sets

	Mean (sd)		Adjusted Mean	
	No Read-aloud (A)	Read-aloud (B)	No Read-aloud (A)	Read-aloud (B)
Total Score	15.5 (5.3)	14.3 (5.4)		
Math Hard/Reading Hard	1.7 (1.2)	1.6 (1.1)	1.6	1.6
Math Hard/Reading Easy	1.6 (1.2)	1.4 (1.1)	1.4	1.4
Math Easy/Reading Hard	2.8 (1.4)	2.6 (1.4)	2.6	2.7
Math Easy/Reading Easy	3.2 (1.4)	2.9 (1.4)	3.0	2.9
Reading Easy*	2.7 (1.3)	2.3 (1.4)	2.5	2.4
Reading Hard	2.7 (1.3)	2.5 (1.3)	2.6	2.6

*Mean difference between groups is significant at $p < .05$, two-tailed.

Significant differences between groups after controlling for overall performance were identified for four of the twelve item sets across the two grade levels that were analyzed. In two of these cases, Group B (read-aloud group) performed significantly better than Group A. This occurred in the fourth grade comparison for the reading hard (RH) item set (F_{fourth} (1, 1834) = 5.2, $p < .05$) and for the math easy/reading hard (ME/RH) item set (F_{fourth} (1, 1834) = 11.0, $p < .01$). These results align with our expectation that positive effects of the read-aloud accommodation for students with reading disabilities would be most clearly evident on items that are difficult to read and those that are difficult to read and mathematically easy. Although the same significant differences were not identified within the eighth grade comparisons, the results for the ME/RH item set within this grade-level approached significance (F_{eighth} (1, 2595) = 3.3, $p < .07$), and represented the only item set in which the Group B adjusted mean was higher than that of Group

Table 4. ANCOVA Results for Read-aloud/No Read-aloud Groups

	4th Grade	8th Grade
Math Hard/Reading Hard	F(1, 1834) = 0.9	F (1, 2595) = 1.6
Math Hard/Reading Easy	F(1, 1834) = 2.4*	F (1, 2595) = 0.1
Math Easy/Reading Hard	F(1, 1834) = 11.0, p < .01, partial eta-squared = .006	F (1, 2595) = 3.3, p < .07
Math Easy/Reading Easy	F(1, 1834) = 2.8	F (1, 2595) = 1.8
Reading Easy	F(1, 1834) = 7.5, p < .01, partial eta-squared = .004	F (1, 2595) = 6.6, p < .05 partial eta-squared = .003
Reading Hard	F(1, 1834) = 5.2, p < .05, partial eta-squared = .003	F (1, 2595) = 1.1

*Levene's test for homogeneity of error variance resulted in p < .05, which suggests that there were unequal error variances among groups. However, the ANCOVA procedure is considered relatively robust to this violation when the larger variance is associated with the larger group sample size (Johnson & Rakow, 1994), and therefore results are reported.

A among the eighth grade comparisons. The remaining two significant differences in Group A and Group B performances occurred on the reading easy (RE) item sets (F_{fourth} (1, 1834) = 7.5, p < .01; F_{eighth} (1, 2595) = 6.6, p < .05), with Group A (non read-aloud group) having higher adjusted means than Group B (read-aloud) across both grade level comparisons.

Demographics of the student groups for the secondary analyses, as well as results for the secondary analyses intended to examine any effects that might be due to extended time are provided in Tables 5 through 8. For only one of the target item groups were significant effects identified in these analyses. This was found in the eighth grade analysis for RH items, in which the estimated mean for this item group was actually higher for the non-accommodated group of students. The results of these secondary analyses seem to suggest that effects of the read-aloud accommodation could not be considered due merely to any extended time that was provided.

Table 5. Demographic Information for Extended Time and Setting/No Accommodation Group

	4th Grade		8th Grade	
	Ext. Time/Sett.	No Acc.	Ext. Time/Sett.	No Acc.
N	108	323	247	473
Percent female	26.9	36.8	35.2	37.8
Percent non-white	24.8	18.7	25.8	26.3
Percent received timing accommodation	38.0	0	33.2	0
Percent received setting accommodation	90.7	0	91.5	0

Table 6. ANCOVA Results for Extended Time and Setting/No Accommodation Group

	4th Grade	8th Grade
Math Hard/Reading Hard	F(1, 428) = 0.0	F (1, 717) = 0.1
Math Hard/Reading Easy	F(1, 428) = 0.2	F (1, 717) = 5.6, p < .05 Partial eta-squared = .008 (acc. perf. better)
Math Easy/Reading Hard	F(1, 428) = 0.1	F (1, 717) = 1.7
Math Easy/Reading Easy	F(1, 428) = 1.1	F (1, 717) = 3.1
Reading Easy	F(1, 428) = 2.6	F (1, 717) = 0.4
Reading Hard	F(1, 428) = 0.1	F (1, 717) = 5.9, p < .05 Partial eta-squared = .008 (non-acc. performed better)

Table 7. Demographic Information for Extended Time /No Accommodation Group

	4th Grade		8th Grade	
	Ext. Time	No Acc.	Ext. Time	No Acc.
N	41	323	82	473
Percent female	29.3	36.8	36.7	37.8
Percent non-white	21.1	18.7	28.0	26.3
Percent received setting accommodation	75.6	0	70.7	0

Table 8. ANCOVA Results for Extended Time/No Accommodation Group

	4th Grade	8th Grade
Math Hard/Reading Hard	F(1, 361) = 1.0	F (1, 552) = 0.0
Math Hard/Reading Easy	F(1, 361) = 0.0	F (1, 552) = 1.1
Math Easy/Reading Hard	F(1, 361) = 0.3	F (1, 552) = 1.1
Math Easy/Reading Easy	F(1, 361) = 7.5, p < .01 (non-acc. performed better) Partial eta-squared = .643	F (1, 552) = 0.6
Reading Easy	F(1, 361) = 0.1	F (1, 552) = 0.1
Reading Hard	F(1, 361) = 0.1	F (1, 552) = 2.1

For one of the item set comparisons, Levene's test for homogeneity of variance was significant, suggesting that this ANCOVA assumption was not met. However, additional analysis indicated that in this case, the largest variance was associated with the larger group. Johnson and Rakow (1994) suggest that the ANCOVA is relatively robust under this circumstance.

Discussion

The intent of this study was to identify whether the read-aloud accommodation had the intended effect of eliminating extraneous reading difficulty for students with disabilities specific to reading on a math test. Previous research has not sufficiently addressed the possibility that students may not have the math skills and knowledge necessary to demonstrate whether this accommodation assists them in accessing test content. Research has also tended to focus on students with disabilities in general, and not on those students that may have a clear need for the accommodation based on their specific academic needs. The effects of the read-aloud accommodation were examined by first classifying items according to both reading and math difficulty, and then measuring differences in item set performance between students with reading disabilities receiving and those not receiving the read-aloud accommodation on elementary and middle school statewide math tests. It was anticipated that group performance differences would be greatest on those items identified as difficult to read, and that those students with reading disabilities receiving the accommodation would score considerably higher on these items. It was also anticipated that positive effects of the accommodation would be most clearly evident on items that were considered difficult to read, but mathematically easy. In order to control for differences in math skills among those receiving and not receiving the read-aloud accommodation, overall score on the multiple-choice math items was used as a covariate in the analysis.

On two item sets, both of which were included within the fourth grade dataset, the group receiving the read-aloud accommodation performed significantly better than the group not receiving the read-aloud accommodation after controlling for overall performance. This occurred on the item set classified as reading hard (RH), and on the item set that was classified as math easy/reading hard (ME/RH). On the item sets classified as RE, students who did not receive the read-aloud accommodation performed better than those who did receive this accommodation, after controlling for overall performance.

Overall, we found support for our first hypothesis within the fourth grade dataset. We anticipated that students with reading disabilities receiving the read-aloud accommodation would perform better than those not receiving the accommodation on items with significant reading content. The significant ANCOVA results for the comparison between Group A and Group B performances on reading hard (RH) item sets indicated that Group B (read-aloud group) tended to perform better on these items than those who had not received the accommodation. However, the same result was not obtained within the eighth-grade comparison.

We also found support for our second hypothesis within the fourth-grade dataset, and came very close to finding support for this hypothesis within the eighth-grade dataset. After controlling for overall score, fourth-grade students with reading disabilities who received the read-aloud accommodation performed better than those not receiving the accommodation on items that

were difficult to read, but mathematically easy. In the eighth grade dataset, this item set was the only one for which the adjusted mean was higher for those receiving the accommodation than for those who did not receive the accommodation, with the ANCOVA results approaching significance ($p < .07$).

Altogether, these findings support the idea that students with reading disabilities who receive the read-aloud accommodation may not be able to demonstrate that the accommodation is truly helpful to them in accessing test content unless the math items are particularly easy. It is likely the case that students who have poor reading skills also have poor math skills, and therefore are only able to show that the accommodation has an effect on those items that contain math content which they have already mastered. Failure to identify accommodation effects on difficult math items would align well with this idea. More research will need to be conducted to verify this, and also to determine why students with reading difficulties also have math difficulties. It may be the case that learning disabilities are pervasive across content areas for many students. It could also be the case that limited reading skills are limiting student access to math instruction. One would need to investigate the extent to which the read-aloud accommodation is provided during instruction to know whether the latter explanation holds.

However, it is important to consider more carefully the findings that students with reading disabilities who did not receive the accommodation tended to perform better on items that were categorized as easy to read. Although this finding aligns with our general expectation that items with more difficult text would be particularly positively affected by the read-aloud accommodation, it raises the question of whether the read-aloud accommodation may add some sort of barrier for students in answering items that are otherwise very easy to read. It may be the case that students only need certain items read to them, in which case, it might be important to allow students to choose when they need test items read aloud to them and when they do not need them read aloud. This appears to align well with previous research on pacing during test administration that included a video read-aloud accommodation (Hollenbeck, Rozek-Tedesco, Tindal, & Glasgow, 2000). This research team found that students with disabilities performed better under a student-paced accommodation condition in which they could select when to use an accommodation than under a teacher-paced accommodation condition in which they received the accommodation for all test items. In our study, it was not possible to control how the accommodation was administered, and we do not have information on how it was actually provided. It may be the case that it was administered in different ways for different students. For example, some readers may have read faster or slower, some may have been more or less apt to re-read questions at student request, and some may have read only some of the test items to students.

We also did not have data on how much extra time students who received the read-aloud accommodation tended to receive; it may be the case that only a certain amount of time should be allocated for students receiving the read-aloud accommodation unless the student has additional

reasons for needing extended time. More research should be conducted to determine how much extra time is necessitated through use of the read-aloud accommodation. Based on our secondary analyses, extended time did not appear to affect our target item groups, but it may have an impact on overall scores.

It is important to note a variety of other factors that may have confounded our results, given its non-experimental nature. Most notably, students were not randomly assigned to accommodated or non-accommodated conditions. The groups may therefore differ in important ways. For instance, students receiving the accommodation did not perform as well as students not receiving accommodation on overall score (particularly in the eighth grade dataset), suggesting that the accommodated students were most likely less proficient in math. Although we attempted to control for differences in math skills using the overall score as a covariate, analysis of covariance will not control for all variables on which two groups differ (Howell, 1997). Particularly in the eighth grade data set, it seems as if the total score covariate was not sufficient to control for differences between groups; students not receiving the accommodation scored equal to or higher on all item sets, with the exception of the ME/RH item set.

Another potential limitation to our study given the lack of control over how accommodations decisions were made is that the two groups (i.e., accommodated and non-accommodated students) may have differed in their need for the accommodation. We did not have access to an independent measure of students' reading skills; however, given that all students had a primary IEP area in reading, we assumed that they all would have benefited from some sort of accommodation to address their reading skill deficits. However, it may be the case that those who actually received the accommodation had the greatest need for the accommodation. Research on accommodation decision-making seems to suggest that educators are not very accurate in making decisions about who does and does not benefit from accommodations (Helwig & Tindal, 2003), and that often student need for an accommodation is not necessarily the deciding factor as to whether they actually receive an accommodation (Shriner & DeStefano, 2003). These research findings suggest that decisions about who did and did not receive accommodations on the test analyzed were not perfectly aligned with student need, which suggests that there was most likely considerable variation in student reading skills within each group. The fact that we obtained significant effects in the direction that we anticipated on our target item groups seems to indicate that this potential limitation did not substantially reduce the power of our study.

Despite our lack of experimental control, it is important to note some important benefits of our approach that could not have been attained using experimental methods. In contrast to experimental designs that may only examine data from a handful of districts with unique characteristics, extant data were available for our analysis from the entire statewide student population. Results from our study therefore generalize to all students with the given characteristics

(i.e., learning disability and primary IEP area in reading) in the state. As schools are required to collect more and more data on student outcomes over time to address legal requirements, it seems only reasonable to analyze these data to help in addressing important research questions wherever possible. Furthermore, with advances in measurement and statistical techniques, it is becoming more and more feasible to examine causal factors using extant data, making it less necessary to pull students out of classroom instruction to participate in experimental research studies. Our study therefore represents an example of using existing data in an efficient manner to address important research questions.

As higher stakes are being attached to student scores on statewide assessments, it is likely that more and more students will be offered testing accommodations in an attempt to increase student performance. In order to ensure that resources are used wisely, it is important that accommodations are only provided to those who are determined to need them based on documented benefits. At the same time, it is important to ensure that students have access to necessary accommodations in both instruction and testing, and that they are not denied accommodations simply because there is no identified improvement in overall test score. In some cases, high-stakes decisions are based on proficiency as measured on tests, and performance on a very small number of items may distinguish students who are considered "proficient" and those considered "non-proficient." Item-level analyses such as the one represented in this study appear to be extremely important to know whether accommodations truly help students with specific disabilities access test content.

Although we found support for the hypothesis that the read-aloud accommodation is particularly beneficial to students on items with complex reading among the fourth grade dataset, we did not find this to be the case for the eighth grade dataset. This finding seems to align with that of Helwig et al. (1999), who suggested that students may need a certain level of mathematical competence to demonstrate accommodation effects on a test. In our study, students appeared to demonstrate accommodation effects particularly on those items for which only a lower level of competence was necessary. It is important to recognize that failure to identify positive effects within the eighth grade dataset could merely be due to lack of math knowledge among these students, which could be due to any number of reasons, including failure to access effective instruction over several years.

It is clearly important for students with reading disabilities to have access to quality math content and instruction. Special educators can work to ensure such access by helping students become familiar with using accommodations, particularly before testing situations, and promoting advocacy skills among students with disabilities to ensure that they have access to necessary supports in the future. Even if students with reading disabilities do not presently demonstrate

a substantial increase in test scores under an accommodated condition, it may be the case that with greater access to quality instruction in math, combined with continued accommodation on math tests, they will demonstrate an overall test score improvement with the read-aloud accommodation on math tests in the future.

In most states, accommodations are made available only to students with identified disabilities (Clapper et al., 2005); it is therefore considered necessary to document that accommodations have a positive impact on the test scores of students with disabilities alone using some sort of differential boost approach. This method of investigation unfortunately neglects the possibility that certain accommodations may actually help all students better show what they know and can do with respect to what the test is intended to measure. A read-aloud accommodation may allow both students who are low-achieving and students with identified disabilities better access to test content. As research on the effects of testing accommodations for students with disabilities has accumulated, some researchers have begun to question whether it might simply be more efficient for test developers to consider the needs of students with disabilities during the initial stages of test development rather than trying to accommodate their needs after standardized test administration procedures have been developed and applied. This area of work has been termed "universal design for assessment" (Johnstone, 2003; Thompson, Johnstone, & Thurlow, 2002). Using principles of universal design, accommodations can be incorporated into tests from the beginning stages of test development and standardization, and made available to all students. This approach can eliminate the need to show that accommodations are differentially effective for students with disabilities. For example, if tests are computerized, and access to screen-reading tools is made available, this might allow all students, even those without identified disabilities, to access math test content that they find difficult to read on an individual basis. In this way, consideration of the unique needs of students with reading disabilities may actually improve the development of tests such that they better measure the intended construct for all students, even those who have not been identified as having a specific reading disability.

Based on findings from past studies, as well as the results of the current investigation, it appears that additional research on the effects of testing accommodations is warranted to identify optimal testing conditions for students with various educational difficulties. Both experimental approaches and carefully-thought out analyses of the growing number of extant large-scale assessment databases can clearly aid in identifying the effects of testing accommodations for both students with and without disabilities.

References

AERA, APA, & NCME (1999). *Standards for educational and psychological testing:* Washington, DC: American Educational Research Association.

Bielinski, J., Thurlow, M., Ysseldyke, J. E., Freidebach, J., & Freidebach, M. (2001). *Read-aloud accommodation: Effects on multiple-choice reading and math items* (Technical Report No. 31). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Available at http://education.umn.edu/nceo/OnlinePubs/Technical31.htm

Bolt, S. E. (2004). Examining empirical evidence for several commonly-held beliefs about testing accommodations for students with disabilities (Doctoral dissertation, University of Minnesota, 2004). *Dissertation Abstracts International, 65*, 1659.

Bolt, S. E., & Bielinski, J. (2002, April). *The effects of the read-aloud accommodation on math test items.* Paper presented at the annual conference of the National Council on Measurement in Education, New Orleans, LA.

Bolt, S., Krentz, J., & Thurlow, M. (2002). *Are we there yet? Accountability for the results of students with disabilities* (Technical Report No. 33). Minneapolis, MN: University of Minnesota. Available at http://education.umn.edu/nceo/OnlinePubs/Technical33.ht

Bolt, S. E., & Thurlow, M. L. (2004). Five of the most commonly allowed accommodations in state policy: Synthesis of research. *Remedial and Special Education, 25,* 141–152.

Bolt, S. E., & Ysseldyke, J. E. (in press). Comparing DIF across math and reading/language arts tests for students receiving a read-aloud accommodation. *Applied Measurement in Education, 19,* xx–xx.

Burch, M. A. (2002). Effects of computer-based test accommodations on the math problem-solving performance of students with and without disabilities. *Dissertation Abstracts International, 63* (03), 902A. (UMI No. 3047429)

Clapper, A. T., Morse, A. B., Lazarus, S. S., Thompson, S. J., & Thurlow, M. L. (2005). *2003 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 56). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Available at http://education.umn.edu/NCEO/OnlinePubs/Synthesis56.html

Elbaum, B., Arquelles, M., Campbell, Y., & Saley, M. (2004). Effects of student reads-aloud accommodation on the performance of students with and without disabilities on a test of reading comprehension. *Exceptionality, 12, 2,* 71–87.

Elliott, S. N., & Marquart, A. M. (2004). Extended time as a testing accommodation: It's effects and perceived consequences. *Exceptional Children, 70,* 349-367.

Erickson, R. N., Thurlow, M. L., & Ysseldyke, J. E. (1996). *Neglected numerators, drifting denominators, and fractured fractions: Determining participation rates for students with disabilities in statewide assessment programs* (Synthesis Report No. 23). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Available at http://education .umn.edu/NCEO/OnlinePubs/Synthesis23.html

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 sources. *School Psychology Review, 29*(1), 65–85.

Helwig, R., Rozek-Tedesco, M. A., & Tindal, G. (2002). An oral versus a standard administration of a large-scale mathematics test. *The Journal of Special Education, 36*(1), 39–47.

Helwig, R., Rozek-Tedesco, M. A., Tindal, G., Heath, B., & Almond, P. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth-grade students. *The Journal of Educational Research, 93*(2), 113–125.

Helwig, R., & Tindal, G. (2003). An experimental analysis of accommodation decisions on large-scale mathematics tests. *Exceptional Children, 69*, 211–225.

Hollenbeck, K. (2002). Determining when test alterations are valid accommodations or modifications for large-scale assessment. In G. Tindal & T. Haladyna (Eds.), *Large Scale Assessment Programs for All Students.* (pp. 109–148). Mahwah, NJ: Lawrence Erlbaum Associates.

Hollenbeck, K, Rozek-Tedesco, M. A., Tindal, G., & Glasgow A. (2000). An exploratory study of student-paced versus teacher-paced accommodations for large-scale math tests. *Journal of Special Education Technology*, *15*(2), 27–36.

Howell, D. (1997). *Statistical Methods for Psychology (4th ed.).* Belmont, CA: Wadsworth Publishing Co.

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

Johnson, C., & Rakow, E. (1994, November). Effects of violations of data set assumptions when using the analysis of variance and covariance with unequal group sizes. *Paper presented at the annual meeting of the Mid-South Educational Research Association*, Nashville, TN.

Johnstone, C. J. (2003). *Improving validity of large-scale tests: Universal design and student performance* (Technical Report 37). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Available at http://education.umn.edu/NCEO/OnlinePubs/Technical37.htm

Klein, J. A., Wiley, H. I., & Thurlow, M. L. (2006). *Uneven transparency: NCLB tests take precedence in public assessment reporting for students with disabilities* (Technical Report 43). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Available at http://education.umn.edu/nceo/OnlinePubs/Technical43.html

Kosciolek, S., & Ysseldyke, J. E. (2000). *Effects of a reading accommodation on the validity of a reading test* (Technical Report 28). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Available at http://education.umn.edu/nceo/OnlinePubs/Technical28.htm

Lewis, D., Green, D. R., & Miller, L. (1999, June). Using differential item functioning analyses to assess the validity of testing accommodated students with disabilities. *Paper presented at the national conference on large-scale assessment,* Snowbird, UT.

McKevitt, B. C., & Elliott, S. N. (2003). Effects and perceived consequences of using read-aloud and teacher-recommended testing accommodations on a reading achievement test. *School Psychology Review, 32*(4), 583–600.

Meloy, L. L., Deville, C., & Frisbie, D. A. (2002). The effect of a read-aloud accommodation on test scores of students with and without a learning disability in reading. *Remedial and Special Education, 23* (4), 248–255.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9)*,* 741–749.

National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics*. Reston, VA: Author. Retrieved March 3, 2003 from http://standards.nctm.org/document/prepost/copyr.htm

Olson, J., & Goldstein, A. (1996). Increasing the inclusion of students with disabilities and limited English proficiency in NAEP. *Focus on NAEP, 2* (1), 1–5.

Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7*(2), 93–120.

Pomplun, M., & Omar, M. H. (2000). Score comparability of a state mathematics assessment across students with and without reading accommodations. *Journal of Applied Psychology, 85*(1), 21–29.

Sato, T. (1996). Type I and Type II errors in multiple comparisons. *Journal of Psychology, 130,* 293.

Schulte, A. G., Elliott, S. N., & Kratochwill, T. R. (2001). Experimental analysis of the effects of testing accommodations on students' standardized achievement test scores. *School Psychology Review, 30*(4), 527–547.

Shriner, J. G., & DeStefano, L. (2003). Participation and accommodation in state assessment: The role of individualized education programs. *Exceptional Children, 69,* 147–161.

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

Thompson, S., Blount, A., & Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001* (Technical Report 34). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Available at http://education.umn.edu/NCEO/OnlinePubs/Technical34.htm

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L., (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Available at http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Thurlow, M.L., Moen, R., & Altman, J. (2006). *Annual Performance Reports (APR): 2003-2004 state assessment data*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Available at http://education.umn.edu/nceo/OnlinePubs/APR2003-04.pdf

Tindal, G. (1998). Models for understanding task comparability in accommodated testing. A publication for the Council of Chief State School Officers, Washington, DC. Retrieved May 19, 2006, from http://education.umn.edu/nceo/OnlinePubs/Accomm/TaskComparability.htm

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An experimental study. *Exceptional Children, 64,* 439–450.

Weston, T. (1999, April). *The validity of oral presentation in testing.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.