

Effects of a Read-Aloud Modification on a Standardized Reading Test

Lindy Crawford

*Department of Special Education
College of Education*

University of Colorado at Colorado Springs

Gerald Tindal

*Department of Educational Leadership
College of Education*

University of Oregon

We investigated the effects of a read-aloud modification on students' performance on a reading comprehension test. A total of 338 students in Grades 4 and 5 participated; 76 of these students (22%) received special education services, the majority of whom were labeled learning disabled. Students completed a standardized reading comprehension test under two conditions: (a) standard administration, and (b) video administration. A repeated measures analysis of variance revealed a significant interaction between educational classification and administration format. We compared teacher judgments related to the importance of the read-aloud modification for individual students to students' actual test performance. Teacher judgment was the most accurate for students in special education. Findings were interpreted within three different contexts: (a) previous research on teacher decision making, (b) theoretical explanations of reading versus listening comprehension, and (c) the importance of research methodologies in answering test modification questions.

Including students with disabilities in statewide testing often involves the use of test accommodations that, in theory, maintain the attributes of the construct being tested and do not alter score meaning (Tindal & Fuchs, 1999). A growing body of research related to the effects of test accommodations is beginning to inform practice, partially evidenced by statewide test policies that include a comprehensive list of allowable accommodations (National Center on Educational Outcomes, 2002). On the other hand, far less research has explored the use of test modifications on large-scale assessments. Test modifications represent changes to test construction, administration, response, or scoring that alter the

nature of the assessed construct in a significant manner (Hollenbeck, 2002). The meaning of a test score under a modified condition is qualitatively different from the meaning of a test score under a standard condition. Because the assessed construct differs so dramatically from the original construct, the data derived from modified tests are often disaggregated from the majority group and analyzed separately.

Statewide assessment policies may actually encourage student participation in the standard assessment with accommodations or in the alternate statewide test but discourage student participation in statewide tests that are modified. Policies in North Carolina and Iowa, for example, aggregate the scores of students who complete standard statewide tests with accommodations and disaggregate (but still count) the scores of students who complete alternate assessments. Students completing the statewide test with modifications, however, are not even counted as participants, and in turn, no scores are reported for their performance. A similar scenario is true in Wyoming except that students are counted as participants, but their test scores are automatically converted to zero (Thurlow, Lazarus, Thompson, & Robey, 2002).

It may be that the differences between an alternate and standard assessment are more clearly demarcated than those differences between a modified and standard test, resulting in clearer score meaning and possibly greater acceptance. An informal review of the research reveals that more research has been conducted on the design and analysis of test accommodations and alternate assessments than has been conducted on tests that have been modified. It is unclear whether the low use of test modifications in the field has driven the lack of research or if the lack of empirically based information has driven the low use of test modifications in the field.

THEORETICAL BASIS

In this study, we examined the effects of a test modification—an oral presentation of a reading comprehension test. Reading a reading test aloud to a student fundamentally changes the skill being assessed (listening comprehension vs. reading comprehension) yet still provides information related to the underlying construct (text comprehension). Researchers have examined the relation between reading and listening comprehension with students in elementary school (Joshi, Williams, & Wood, 1998), middle school (Royer, Kulhavy, Lee, & Peterson, 1986), and college (Sinatra, 1990). They have consistently found that reading and listening comprehension are highly related and share many of the same cognitive processes. This relation between reading and listening comprehension was coined the “unitary view of comprehension” by Sticht (1979). According to the unitary view of comprehension, the only difference between reading and listening is the modality through which information is received (Sticht & James, 1984).

Correlation studies provide much of the research support for the unitary view of comprehension (Aaron, 1991; Joshi et al., 1998; Palmer, McCleod, Hunt, & Davidson, 1985; Rispen, 1990; Townsend, Carrithers, & Bever, 1987). In an attempt to improve on and expand the understanding of the relation between reading and listening comprehension, Joshi et al. used highly standardized and validated tests of reading and listening comprehension to assess the comprehension of 273 students in Grades 3 through 6. Students

completed two subtests (cloze tests) in the Woodcock Reading Mastery Test–Revised (WRMT–R: Woodcock, 1987). Joshi et al. reported strong correlations between reading and listening ranging from .61 at Grade 4 to .75 at Grade 6, with a mean correlation of .67 across all four grades.

In another study conducted by Joshi et al. (1998), 60 students in the third grade and 60 students in the fifth grade completed the reading and listening comprehension subtests of the Wechsler Individual Achievement Test (Wechsler, 1991) and the WRMT–R. Students also completed the Peabody Individual Achievement Test–Revised (Markwardt, 1989). The range of correlations for the measures spanned from .40 to .90 across both grades and all three tests, with the strongest correlations reported at Grade 5. Across all of the tests, listening comprehension accounted for greater variance in reading comprehension at Grade 5 than it did at Grade 3.

One conclusion made by Joshi et al. (1998) is that “listening comprehension is a better predictor of reading comprehension among older (fifth grade) than younger (third grade) children. This suggests that decoding skills may be relatively more important in accounting for variation in reading comprehension in younger children” (p. 325). Many other researchers also have reported that as students age, listening comprehension accounts for more variance in reading comprehension than does decoding (see meta-analysis by Gough, Hoover, & Peterson, 1996).

A related body of research has investigated the relation between reading and listening comprehension across readers with different proficiency levels. For example, Royer, Sinatra, and Shumer (1990) used results of the CTBS Reading subtest to group third- and fourth-grade students into three reading levels—high, medium, and low. Royer et al. administered tests of comprehension in both reading and listening to all groups; they found no main effect for modality of presentation but reported a within-grade interaction between reading level and modality, with good readers performing significantly better on tests of reading comprehension than on tests of listening comprehension and poor readers performing better on tests of listening comprehension. Royer et al. concluded:

Listening will be superior to reading when the child either has very poor skills or when the materials exceed the capabilities of the student. Alternatively, reading will be superior to listening when the students have well developed reading skills and when the test materials are at or below the skill level of the student. (p. 194)

Similar results also were reported by Royer et al. (1986) who found that the reading comprehension of fourth- and sixth-grade students exceeded their listening comprehension scores when passages were below grade level but that listening comprehension exceeded reading comprehension when passages were above grade level.

Royer et al. (1990) relied on a framework developed by Kleiman and Schalert (1978) to explain the findings reported previously: When developing readers read easier (decodable) text, they are able to focus their attention on metacognitive skills such as self-monitoring, rereading, and adjusting their pace when necessary. They can focus on grasping the meaning of the text. However, when developing readers face textual material that is too hard for them in vocabulary and other syntactical features, they become fully absorbed in decoding the words and are not able to concentrate on the meaning of the text even when they are

given time to reread it. By reading this harder text aloud to the developing reader, he or she is relieved of the need to decode and can focus on comprehending the message.

In summary, students with strong decoding skills or students reading passages below their instructional level are likely to perform better on reading comprehension tasks than on listening comprehension tasks, but students without fully developed word recognition skills or who face difficult passages may perform better when information is read to them. It is important to note, however, that individual characteristics also affect an individual's strengths at listening versus reading, and blanket assumptions about a student's reading versus listening comprehension cannot be made. For example, if a student has poor decoding skills but even poorer auditory skills, he or she may comprehend information better when it is read rather than heard.

RESEARCH ON READ-ALoud MODIFICATIONS

Our primary goal in this study was to explore the effect of a read-aloud modification on students' comprehension of textual information. Our first research question asks: Does a read-aloud modification result in significantly better test scores than reading passages silently? An intuitive answer to this question is that reading passages aloud to students will improve their comprehension of the information. Whereas some researchers have found this to be true (Meloy, Deville, & Frisbie, 2002), others have not (Kosciolek & Ysseldyke, 2000), noting that the Kosciolek and Ysseldyke study had very few participants. Our second and related question asks: Does a read-aloud modification result in a "differential boost" (Fuchs & Fuchs, 2001) for individual students with disabilities? One assumption is that students with disabilities will benefit more from a read-aloud modification than students without disabilities (Kosciolek & Ysseldyke, 2000; Meloy et al., 2002), whereas others have reported that "reading the reading test" does not result in higher scores for all students with reading disabilities (Bielinski, Thurlow, Ysseldyke, Freidebach, & Freidebach, 2001). In fact, Bielinski et al. reported that in some cases, "reading the reading test made a bad situation worse" (p. 13). Results of initial research studies on test modifications as well as theory surrounding text comprehension do not lead us to believe that a read-aloud modification automatically increases the reading test performance of all students or even certain subgroups of students.

An oral presentation of information in the context of a reading test is considered a modification, whereas oral presentation of information in the context of a math test is viewed as an accommodation; in the former scenario, the construct being assessed is confounded with the presentation format. An increasing number of research studies have explored the effects of a read-aloud accommodation on students' math performance (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000; Helwig, Rozek-Tedesco, Tindal, Heath, & Almond, 1999; Hollenbeck, Rozek-Tedesco, Tindal, & Glasgow, 2000); however, researchers have not reported consistent results. Lack of consistency may be due to confounds associated with analyzing test performance in specific content areas while simultaneously investigating the effects of a read-aloud accommodation. An obvious need exists for more empirical studies into the effect of a read-aloud modification on statewide reading tests. In this study, the construct validity of our findings will be enhanced through our analysis of the effects of a read-aloud modification on a reading test.

TEACHER DECISION MAKING

One possible explanation for the lack of a “blanket effect” of read-aloud accommodations is provided by Fuchs and Fuchs (2001) in a summary of their research on the accuracy of teacher decision making related to the assignment of test accommodations. Similar to many test accommodations (e.g., extended time and use of a scribe), read-aloud accommodations are primarily used for students with learning disabilities. Yet, group designs are unable to represent the heterogeneity apparent in the population of students labeled as learning disabled. Results of group design studies, therefore, may mask the effectiveness (or noneffectiveness) of test accommodations on individual performance.

“Consequently, [Fuchs & Fuchs] have been developing and studying an assessment tool that teachers can use to supplement their judgments about whether an individual student with LD [learning disabilities] should be awarded a specific accommodation” (Fuchs & Fuchs, 2001, p. 178). The tool that Fuchs and Fuchs designed acts as a “pre-test” investigating the effects of various accommodations on students’ test performance before actually assigning accommodations during high-stakes testing. The criterion Fuchs and Fuchs used to calculate whether or not an assigned accommodation results in a differential boost for students with learning disabilities is calculated by computing the mean difference between the standard and accommodated test scores attained by students without disabilities and then adding one standard deviation to that number. Students with learning disabilities who demonstrate such a large difference between test scores on the standard and accommodated versions are described as “profit[ing] substantially” (Fuchs & Fuchs, 2001, p. 177) from the pretest accommodation and thus receive the accommodation on high-stakes tests. Across both math and reading accommodations, Fuchs and Fuchs found that their data-based tool resulted in more accurate assignment of test accommodations than did teacher-based decisions made in the absence of data.

Research has demonstrated a need to develop this type of data-based tool, as teachers do not consistently make accurate decisions when assigning test accommodations (Fuchs & Fuchs, 2001; Helwig & Tindal, 2003). Teacher decision making related to the assignment of a read-aloud modification for a statewide reading test is the focus of our third research question: Is there a difference between teachers’ rating of students’ reading abilities and students’ performance on the standard administration of the reading test? Our fourth and final research question continues this line of inquiry by asking: Is there a difference between teacher judgments of whether or not students will benefit from a read-aloud modification and students’ actual benefits?

METHOD

Participants

Teachers and students from two states (Oregon and North Carolina) participated in this study. Participating teachers were not randomly selected. They had agreed to participate

in a larger research study investigating the effect of an oral presentation of a math state-wide test and self-selected into this smaller, related study.

Each participating building was represented by two general education teachers with classrooms of approximately 25 to 30 students and one special education teacher with all students in Grades 4 and 5 whose individualized education programs (IEPs) contained academic goals and objectives. A total of 357 students participated; analyses were conducted on 338 students (those students who completed both test formats). Twenty-six percent of the students received Title One support ($n = 89$) and, due to our purposeful oversampling, 22% of the students received special education services ($n = 76$). The majority of students in special education were labeled as learning disabled, and all students on IEPs were partially or fully included in the general education classroom (see Table 1 for demographic information about participants).

Conditions

To compare students' performance on both reading and listening comprehension measures, we used a within-factor, or crossed, design. In this design, every participant engages in every condition, thus reducing threats associated with the lack of a random sample.

Students completed the reading test under two conditions in random order. The first condition was form. To avoid practice effects, we created two alternate forms of the test that contained similar but not identical passages and questions (Form A and Form B). Students were preassigned forms. The second condition was administration format (standard or video). Students were preassigned into either the standard or the video administration so that one half of the students taking the video version were assigned Form A, and the other one half of the students were assigned Form B. Conditions were precoded on the individual test packets teachers received. To maintain treatment integrity, we explained to teachers that they must give each reading test to each child in the manner that we assigned.

Measures

Test passages and questions were drawn from a larger sample of items previously developed by one of the participating states. Four test forms were created across both grades. A test booklet was printed in both a standard version (dimensions = 8.5 × 11 in.) and a video version (dimensions = 4.25 × 5.5 in.). The standard version had text displayed across opposing pages, whereas the video version had text on the right side with the opposing (left) page left blank. This system resulted in four different test booklets (Form A, standard and video, and Form B, standard and video). We printed all copies using a color system so teachers could easily distinguish matching forms (e.g., a green booklet for Form A–video and blue booklet for Form B–standard). All test materials (booklets and directions) were distributed at a training workshop in which teachers received standardized administration directions to read to students.

TABLE 1
Student Demographics

<i>Demographic</i>	<i>Fourth Grade</i>	<i>Fifth Grade</i>
Gender		
Male	36	131
Female	38	133
Grade total	74	264
Educational classification		
General education	25	148
Title One	39	50
Special education	10	66
Grade total	74	264
Ethnicity		
Caucasian	65	162
African American	8	82
Hispanic	0	8
Asian/Pacific Islander	0	3
American Indian/Native Alaskan	1	0
Multiracial	0	8
Missing data	0	1
Grade total	74	264
Disability		
Mental retardation	1	2
Speech or language	1	12
Orthopedic impairments	0	0
Traumatic brain injury	0	0
Specific learning disability	6	35
Serious emotional disturbance	0	1
Hearing impairment	0	1
Visual impairment	0	0
Autism	0	0
Other health impairment	2	4
Learning disability and speech	0	6
Other health impairment and speech	0	2
Missing data	0	3
Grade total	10	66

Administration Procedures

Standard administration. The standard test consisted of five passages, each followed by five to eight questions. Students were allowed 45 min to complete the test. For the standard test administration, teachers read the following script:

You are about to take a 30-question reading test similar to other reading tests you have taken. You will read each passage and then each comprehension question. For each question, choose the *one* best answer from the four choices given. Your answers should be marked on the separate answer sheet, which is provided. The test will last for 45 minutes. I will announce when you have 20, 10, and 5 minutes left

and 1 minute left. You are not expected to know the answer to every question. Some of the questions may involve ideas that you have not talked about yet in your reading classes. If you come to a question that you do not know how to answer, think carefully and then choose the answer that seems correct. You may go back at any time and change an answer or complete a question that you have skipped.

Video administration. Students watched a television monitor where each passage in the test was read aloud by a research assistant; students could choose to follow along in their test booklets. After the textual information was read, students listened as each question and its answer choices were presented. As each answer choice was read aloud, it changed color on the television monitor. After the last answer choice was read, the screen went blank and teachers told students, "Choose the correct answer now."

We required that teachers pace the administration. Findings of Curtis and Kropp (1961) helped justify this decision; they found significantly higher performance when items were projected in a paced manner on a large screen (either one or three at a time) relative to taking the test with a traditional booklet and answer sheet. To avoid making the presentation too fast or too slow, we directed teachers to pause the videotape and allow approximately 30 sec for students to respond. Teachers had the discretion to lengthen the times for each question (up to twice the length of the suggested times), but they were told never to shorten the suggested time. When the allotted time was up, students were told to turn the page and the teacher would again begin the videotape. For the videotaped administration, teachers read the following script:

You are about to take a 30-question reading test that may be different from other reading tests you have taken. Each passage will be read aloud on a videotape by an actor. Each passage also is printed in your test booklet exactly as it is being read. You may watch the video monitor as the passage is read, or read the passage silently to yourself while the actor reads it. After the passage is read, a question and four answer choices will be shown on the screen. These choices also are printed in your test booklet.

For each question, choose the *one* best answer from the four choices given. Your answers should be marked on the separate answer sheet that is provided. Each question is printed on a separate page. *Do not* turn the page to the next question until you are told to do so. You *may not* go back and change answers, so think carefully before you make your choice. You are not expected to know the answer to every question. Some of the questions may involve ideas that you have not talked about yet in your reading class. If you come to a question that you do not know how to answer, think carefully and then choose the answer that seems correct.

DATA ANALYSIS OF RESEARCH QUESTIONS

Our research questions are grouped into two categories, the first having to do with the effects of a read-aloud modification and the second set pertaining to teacher decision making related to assigning test modifications.

Effects of Modification

- Research question 1: Does a read-aloud modification result in significantly better test scores than reading passages silently?
- Research question 2: Does a read-aloud modification result in a differential boost for individual students with disabilities?

We completed five different analyses when ascertaining the effects of the modification. First, we investigated mean group differences across Grades 4 and 5 using students' raw scores on both the standard test and the video administration. Second, we investigated the effects of order on performance. Third, we tested for a form effect (A or B). Fourth, we conducted a 2×2 analysis of variance (ANOVA) with one between factor (educational classification of student) and one within factor (administration format). Fifth, we calculated individual gain scores to investigate the possibility of a differential boost for students with disabilities.

Teacher Decision Making

- Research question 3: Is there a difference between teachers' rating of students' reading abilities and students' performance on the standard administration of the reading test?
- Research question 4: Is there a difference between teacher judgment on whether or not students will benefit from a read-aloud modification and students' actual benefits?

To answer questions related to teacher decision making, we first conducted a chi-square analysis to investigate possible differences between teachers' rating of students' reading abilities and student performances on the reading assessment. Second, we used the Fuchs and Fuchs (2001) differential boost criterion to analyze whether or not teachers were accurate in their judgments concerning who would benefit from the test modification.

RESULTS

Grade Effect

No grade effect was found when analyzing differences between standard administration raw scores obtained by students in Grades 4 and 5, $t(336) = .354, p = .724$. Similarly, there were no significant differences on the video administration across both grades, $t(336) = -.885, p = .377$ (see Table 2 for mean scores across both conditions). Due to unequal sample sizes, post hoc Sheffé F tests were calculated but also revealed nonsignificant p values. Because grade did not reveal itself as a confound, we combined scores at Grades 4 and 5 for the remainder of the analyses.

TABLE 2
Mean Scores for Students in Grades 4 and 5

	<i>n</i>	<i>M</i>	<i>SD</i>
Standard administration			
Grade 4	74	19.73	5.17
Grade 5	264	19.47	5.79
Video administration			
Grade 4	74	20.66	4.29
Grade 5	264	21.20	4.71

TABLE 3
Mean Scores for Forms A and B

	<i>n</i>	<i>M</i>	<i>SD</i>	<i>SE</i>
Standard administration				
Form A	159	19.70	5.18	0.41
Form B	179	19.37	6.06	0.45
Video administration				
Form A	179	20.35	4.72	0.35
Form B	159	21.91	4.38	0.35

Order Effect

In all of the comparisons, the superscripted number refers to the order of administration (1 = first, 2 = second). For the standard presentation, an ANOVA revealed no significant difference for the combined order and form combinations, $F(3, 334) = 1.03, p = .379$. A follow-up pairwise comparison (Sheffé F) holding form constant revealed no significant difference for XXDEFINITION OF SA (SA)¹ versus SA² ($p = .692$), nor was a significant difference found for XXDEFINITION OF SB (SB)¹ versus SB² ($p = .718$). For the video presentation, an ANOVA revealed a significant difference for the combined order and form combinations, $F(3, 334) = 4.44, p = .005$. Pairwise comparisons holding form constant, however, revealed no order effect for XXDEFINITION OF VA (VA)¹ versus VA² ($p = .342$) and XXDEFINITION OF VB (VB)¹ versus VB² ($p > .999$).

Form Effect

Holding administration format constant, we investigated possible score differences on Form A versus Form B (see Table 3 for mean scores). We found no significant difference between Form A and Form B during the standard presentation, $F(1, 336) = 0.29, p = .594$; however, a significant difference was found between Form A and Form B during the video presentation, $F(1, 336) = 9.91, p = .002$.

Administration Format (Standard or Video)

Our primary analysis investigated the variance across three groups of students (general education, special education, and Title One) and was based on a repeated measures

TABLE 4
Mean Scores Across All Educational Classifications

	<i>n</i>	<i>Standard Administration</i>		<i>Video Administration</i>		<i>Difference</i>
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
General education	173	21.71	4.39	22.83	3.59	1.12
Title One	89	20.35	4.83	21.08	4.00	0.73
Special education	76	13.58	4.96	17.11	4.96	3.53

ANOVA. A significant interaction was found between student classification and administration format, $F(2, 335) = 13.47, p < .0001$. As is apparent in Table 4, all students performed better under the video presentation, with students in special education benefiting the most. Cohen's d effect sizes (Rosenthal, 1994), using the standard administration as the control group, revealed a large effect size for students with disabilities (.71), small effect size for students in general education (.28), and relatively no effect size for students in Title One (.17).

Differential Boost

Using the Fuchs and Fuchs (2001) formula for analyzing the effects of an accommodation, we computed the mean difference between the standard administration and the video administration attained by the group of students in general education (difference = 1.12) and then added one standard deviation (video presentation, $SD = 3.59$) to that number, resulting in a substantial difference between the two test scores of 4.71, which we rounded to 5 points.

We then calculated gain (or loss) scores for each student across both administration formats. Thirty-three percent of students with disabilities demonstrated an increase of 5 points or greater on the video presentation versus the standard presentation ($n = 25$); 12% of Title One students also demonstrated a substantial increase ($n = 11$) as well as 13% of students in general education ($n = 23$). Scores of 14 students (4%) demonstrated the reverse effect (those favoring the standard administration by at least 5 points); 4 of these students received special education, 9 students were in general education, and 1 student was in Title One.

Teacher Ratings

Teachers were asked to rate students' reading proficiency along a scale ranging from 1 to 5, with a score of 1 being the least proficient and 5 being the most. Students' test performance was grouped into three categories: (a) weak (0–10 correct), (b) average (11–20 correct), and (c) strong (21–30 correct).

A chi-square analysis revealed a significant difference between teachers' rating of students' reading proficiency and students' actual performance on the standard administration of the reading test, $\chi^2(8, N = 337) = 177, p < .0001$. Discrepancies are apparent in the

TABLE 5
Teacher Rating of Students' Reading Proficiency and Student Performance
on Standard Administration

<i>Proficiency</i>	<i>Low Performance</i>	<i>Average Performance</i>	<i>High Performance</i>	<i>Totals</i>
Very low	18	25	5	48
Low	8	52	9	69
Fair	3	51	55	109
High	0	10	53	63
Very high	0	2	46	48
Totals	29	140	168	337 ^a

^aOne missing score for teacher rating.

observed frequencies (Table 5) in that of the 48 students that teachers ranked “very low” in reading proficiency, only 18 (38%) scored in the lowest third of the test scores. A similar but not so pronounced pattern continues through the next two levels of reading proficiency, but teachers were very accurate in correctly rating students at the final two levels (96% of those students rated as “highly proficient” in reading scored in the top one third of the test, and 100% of those students rated as “proficient” scored in the top two thirds on the standard administration).

Teacher Judgment

Teacher judgments related to the importance of the read-aloud modification on the performance of individual students were compared to students' actual boost. Teachers identified a total of 135 students who they believed would substantially benefit from the modification (the modification was identified as having “high” or “very high” importance), but only 59 students met the +5 criterion established for the video administration.

Teacher judgments related to the boost displayed by students on IEPs were very accurate, with 100% (25 of 25) of students who met the criterion being judged by their teachers as needing the modification (high or very high importance).

Two hundred seventy-eight students did not demonstrate a substantial boost from the video administration; 99 of these students (35%) were judged by their teachers as needing the modification (high or very high importance), whereas 57 students (21%) were judged by their teachers as possibly needing the modification (fair importance). (One of the 278 students was not rated by his teacher.) One hundred twenty-one students (44%) were accurately judged by their teachers as not needing the modification (low or very low importance).

DISCUSSION

We have chosen to interpret our findings within three different contexts: (a) previous research on teacher decision making, (b) theoretical explanations of reading versus listen-

ing comprehension, and (c) importance of research methodologies in answering test modification questions.

Previous research has found teacher decision making related to the assignment of test accommodations inaccurate (Fuchs et al., 2000; Helwig & Tindal, 2003; Weston, 1999). Specifically, Helwig and Tindal reported that teachers inaccurately recommended a read-aloud accommodation (in math) 45% of the time. Similarly, we found that in 35% of the cases in which teachers judged a read-aloud modification as being “important” or “highly important” to students’ test performance, students did not benefit from the modification, and in 21% of the cases, teachers inaccurately judged the modification to be “fairly important.”

One explanation for the preceding findings lies in the fact that teachers overidentify students as needing test accommodations or modifications (Fuchs & Fuchs, 2001). In our study, teachers judged that 40% of all students would greatly benefit from the modification (high or very high importance), but only 17% of the students benefited (according to the Fuchs & Fuchs differential boost criteria). These findings align with previous research highlighting teachers’ tendencies to make Type I errors.

We found that teachers were even less accurate when rating students’ reading proficiency as measured by student performance on a standard reading comprehension test. The combination of these findings highlight teachers’ tendency to underestimate students’ skills and overestimate their need for help. In the context of large-scale testing, it may be better that teachers underestimate the skills of low-performing students to ensure that all students who need test accommodations or modifications are provided with them. On a more practical level, however, it is disconcerting that teachers may make instructional decisions based on inaccurate perceptions of students’ skills.

Our findings also support theoretical assumptions about the construct of text comprehension (people understand information better when it is read aloud if they have weak decoding skills or if the passages are above their reading level) in that a greater percentage of students with disabilities profited from the test being read aloud than did students without disabilities. Results of our ANOVA also align with the theoretical premise that as people become older, they perform better on tests of reading comprehension than on tests of listening comprehension (Gough et al., 1996). Students in this study were not old (Grades 4 and 5), and on average, all three groups of students performed better on the read-aloud (video) administration. Similarly, when data were analyzed on an individual level, only 4% of the students displayed a differential boost on the standard (read silently) administration.

The results of our group analyses indicate that reading a reading test aloud improves the scores of students in Grades 4 and 5 regardless of their educational classification. One interpretation of our findings is that we changed the construct from reading comprehension to listening comprehension because we found significant score differences across both administration formats and across students with and without disabilities. Because the read-aloud improved the scores for students with and without disabilities, it would be judged by many as an inappropriate accommodation for reading tests (Phillips, 1994).

Yet, unlike Meloy et al. (2002) who reported only main effects for a read-aloud modification, we also found an interaction in which students with disabilities improved signif-

icantly more than did students in either general education or Title One. In a sense, the modification “leveled the playing field” but only after it improved the average scores of each group of students. Interpretation of this finding is complicated and may lie in the context of education policy as opposed to educational research. Critical to this discussion is the definition of comprehension.

Individual states define reading comprehension within their content and performance standards to which tests of reading comprehension are anchored. As an example, the state of Oregon’s reading benchmarks include those attributes inherent in fluent decoding (e.g., accuracy, natural phrasing, smooth flow). However, in Grade 3, students must only “determine meanings of words using contextual clues and illustrations” (Oregon Department of Education, 2002, p. 1). At Grades 5, 8, and 10, students must “determine meanings of words using contextual and structural clues and other reading strategies” (Oregon Department of Education, 2002, p. 1). After Grade 3, no reference is ever made to actually reading (decoding), and in the more specific language of the performance standards, the active verb is silent about the behavior of physically reading. For example, students are required to locate information, retell and summarize events and main ideas, identify cause and effect relations, and analyze and evaluate information. Again, the physical act of decoding is not specified. In Oregon, then, in the intermediate grades and higher, it would be possible to argue for reading a reading test and consider it an accommodation rather than a modification.

In North Carolina, the act of decoding is explicitly included in the language arts curriculum at most grade levels. Interestingly, however, is North Carolina’s emphasis on oral communication as illustrated by the following goal at fourth grade: “The learner will apply strategies and skills to comprehend text that is read, heard, and viewed” (North Carolina Public Schools, 1999a, ¶3). There is a different goal at second grade: “The learner will apply strategies and skills to create oral, written, and visual texts” (North Carolina Public Schools, 1999b, ¶5). These examples and the examples from Oregon’s standards extend the construct of reading comprehension far beyond a simple “understand what you read” definition.

Our third and final context for interpretation lies in the complexities associated with research design in trying to answer questions about the effects of test modifications; “to provide the most convincing empirical support for an accommodation, students with a specific need have to be compared to others without such a need who are otherwise comparable in achievement” (Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998, p. 442). However, what if mean scores improve significantly across various groups and an interaction exists? Do we choose to interpret these findings in a different light, such as the large differences in effect sizes (almost three fourths of a standard deviation improvement for students in special education as opposed to one fourth of a standard deviation difference for students in general education) or perhaps in the gain scores of individual students?

As we found in this study, group results mask the effect of a read-aloud modification on the performance of individuals. Our findings (33% of students with disabilities demonstrated a differential boost) echo those of Fuchs and colleagues (as discussed in Fuchs & Fuchs, 2001) who found that, as a group, both students in general and special education benefited from a read-aloud modification on math tests, but 12% of students with

learning disabilities demonstrated a differential boost over their general education peers. Partial answers to a valid label for a read-aloud administration (modification or accommodation) may lie in single subject rather than group design research.

Strengths and Weaknesses of Study

Results of this study need to be interpreted with an understanding of important design features. First, teachers implemented all experimental procedures following a 1-day training session. We received overwhelmingly positive feedback from the teachers after the 1-day training, and teachers unanimously reported that they felt capable of running the study in their classrooms, which included scheduling all testing in various classrooms, managing the group administration of the tests, and operating the videotape. Yet, despite the positive evaluations received after the training sessions, all of the teachers did not conduct the testing within the 1-week window we suggested, and some teachers did not test until after a significant delay (sometimes up to 2–3 weeks later). Further complicating these delays was the fact that we did not monitor teacher implementation and have to assume that they were proficient.

Our sampling plan was neither random nor stratified for teachers or students. Rather, we solicited teachers who had participated in a related, larger study on test accommodations in the area of mathematics. We cannot ascertain the degree to which the teachers in this study were representative of other teachers in their state, but we have reported demographic information related to the student sample.

Such limitations notwithstanding, our methodology controlled for many critical variables: (a) Teachers followed a uniform process for administering the tests; (b) the videotape controlled for reading rate, ethnicity, and gender; (c) television monitors were large with easily visible screens, and teachers used a remote control allowing them to roam among the students while proctoring the tests; and (d) the booklet was designed so that each question being read was presented individually to the student, ensuring the treatment was being implemented. Finally, by using a repeated measures design while controlling for order effects, we increased the internal validity of our findings.

Implications

Our findings highlight the importance of investigating the benefits of a read-aloud administration for individual students before adopting this procedure as a read-aloud accommodation on content area tests. Many teachers would profit from a standardized procedure for making these types of test administration decisions. One method for improving teacher decision making may be to train them on how to compute the score needed to justify a substantial difference when the read-aloud procedure is used with a reading comprehension test. Then, with this information, teachers could more knowingly assign a read-aloud accommodation for content area tests.

Carlisle's (1991) suggestion of classifying students into two groups according to their reading comprehension difficulties may be most appropriate. Group 1 consists of students with "specific reading disabilities" (Carlisle, 1991, p. 18), or students who have adequate linguistic skills but slow or inaccurate decoding skills. Group 2 is made up of stu-

dents with “general comprehension problems” (Carlisle, 1991, p. 19), or students who perform poorly in both reading and listening comprehension tasks. To measure the usefulness of a read-aloud accommodation on a content area test, it may be necessary to first decide which students have difficulties with comprehension due to their poor decoding and which students have difficulties with comprehension regardless of the modality in which information is presented. By asking students to first complete a reading comprehension test presented both orally and in writing, teachers will be able to make the distinction Carlisle proposed.

An investigation into the reasons why teachers’ overassign test modifications also may be warranted. Although it may be true that teachers’ lack knowledge (Helwig & Tindal, 2003), it is just as likely that teachers feel pressured to assign test modifications to students “on the bubble” to avoid aggregating the scores of low-performing students with the scores of students in general education. As we previously discussed, overassignment of test modifications can result in negative consequences to individual students, but with the passage of the No Child Left Behind Act of 2001, underassignment of modifications could result in severe negative consequences for entire schools.

As with teachers, researchers too must not lose sight of the importance of carefully analyzing the effects of test modifications and accommodations on individual students. For 51 of 76 students in special education, the read aloud did not substantially (+1 *SD*) boost their scores. These results highlight the importance of the need for more empirical research in this area and cautions us against embracing the assumption that reading a reading comprehension test aloud will automatically increase the performance of students in special education and have no effect on students in general education (13% of whom showed a substantial boost under this condition).

Finally, an interesting subfinding was the appearance of an effect for Form B under the video administration. Researchers have only begun to explore possible interactions between individual questions and administration formats (e.g., see the work of Bielinski et al., 2001, on differential item functioning). Our finding of a form effect for the video administration leads us to suggest further studies investigating how different questions interact with different media. Research in this area will be especially important for item bias teams as they attempt to create comparable questions that students complete via various electronic formats.

REFERENCES

- Aaron, P. G. (1991). Can reading disabilities be diagnosed without using intelligence tests? *Journal of Learning Disabilities, 24*, 178–186, 191.
- Bielinski, J., Thurlow, M., Ysseldyke, J., Freidebach, J., & Freidebach, M. (2001). *Read-aloud accommodations: Effects on multiple-choice reading and math items* (Tech. Rep. No. 31). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved November 3, 2002, from <http://education.umn.edu/NCEO/OnlinePubs/Technical31.htm>
- Carlisle, J. F. (1991). Planning an assessment of listening and reading comprehension. *Topics in Language Disorders, 12*(1), 17–31.
- Curtis, H. A., & Kropp, R. P. (1961). A comparison of scores obtained by administering a test normally and visually. *Journal of Experimental Education, 29*, 249–260.

- Fuchs, L. S., & Fuchs, D. (2001). Helping teachers formulate sound test accommodation decisions for students with learning disabilities. *Learning Disabilities Research and Practice, 16*, 174–181.
- Fuchs, L. S., Fuchs, D., Eaton, S., Hamlett, C. L., & Karns, K. (2000). Supplementing teachers' judgments of mathematics test accommodations with objective data sources. *School Psychology Review, 29*, 65–85.
- Gough, P. B., Hoover, W. A., & Peterson, C. (1996). Some observations on the simple view of reading. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties* (pp. XX–XX). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- 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.
- 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. M. Haladyna (Eds.), *Large-scale assessment programs for all students* (pp. 395–426). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- 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 [Electronic version]. *Journal of Special Education Technology, 15*(2), 1–78.
- Joshi, R. M., Williams, K. A., & Wood, J. R. (1998). Predicting reading comprehension from listening comprehension: Is this the answer to the IQ debate? In C. Hulme & R. M. Joshi (Eds.), *Reading and spelling: Development and disorders* (pp. 319–327). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Kosciulek, S., & Ysseldyke, J. E. (2000). *Effects of a reading accommodation on the validity of a reading test* (Tech. Rep. No. 28). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved December 2, 2002, from <http://education.umn.edu/NCEO/OnlinePubs/Technical28.htm>
- Markwardt, F. C. (1989). *Peabody Individual Achievement Test—Revised*. Circle Pines, MN: American Guidance Services.
- 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*, 248–255.
- National Center on Educational Outcomes. (2002). *Special topic area: Accommodations for students with disabilities*. Retrieved December 4, 2002, from <http://education.umn.edu/nceo/TopicAreas/Accommodations/StatesAccomm.htm>
- North Carolina Public Schools. (1999a). *English language arts curriculum: Fourth grade*. Retrieved March 20, 2003, from <http://www.ncpublicschools.org/curriculum/languagearts/gradefour.htm>
- North Carolina Public Schools. (1999b). *English language arts curriculum: Second grade*. Retrieved March 20, 2003, from <http://www.ncpublicschools.org/curriculum/languagearts/gradetwo.htm>
- Oregon Department of Education. (2002). *Oregon standards (English)—School year 2002–2003*. Retrieved January 21, 2002, from <http://www.ode.state.or.us/tls/english/standards/contentstandards.pdf>
- Palmer, J., McCleod, C., Hunt, E., & Davidson, J. (1985). Information processing correlates of reading. *Journal of Memory and Language, 24*, 59–88.
- Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7*, 93–120.
- Rispens, J. (1990). Comprehension problems in dyslexia. In D. A. Balota, G. B. Flores d'Arcais, & K. Rayner (Eds.), *Comprehension processes in reading* (pp. 603–620). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Sage.
- Royer, J. M., Kulhavy, R. W., Lee, J. B., & Peterson, S. E. (1986). The sentence verification technique as a measure of listening and reading comprehension. *Educational and Psychological Research, 6*, 299–314.
- Royer, J. M., Sinatra, G. M., & Shumer, H. (1990). Patterns of individual differences in the development of listening and reading comprehension. *Contemporary Educational Psychology, 15*, 183–196.
- Sinatra, G. M. (1990). Convergence of listening and reading processing. *Reading Research Quarterly, 15*, 115–130.
- Sticht, T. G. (1979). Applications of the Audread model to reading evaluation and instruction. In L. B. Resnik & P. A. Weaver (Eds.), *Theory and practice of early reading* (Vol. 1, pp. 209–226). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

- Sticht, T. G., & James, J. H. (1984). Listening and reading. In P. D. Pearson (Ed.), *Handbook of reading research quarterly* (pp. 293–317). New York: Longman.
- Thurlow, M. L., Lazarus, S., Thompson, S., & Robey, J. (2002). *2001 state policies on assessment participation and accommodations* (Synthesis Rep. No. 46). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved December 28, 2002, from <http://education.umn.edu/NCEO/OnlinePubs/Synthesis46.html>
- Tindal, G., & Fuchs, L. (1999). *A summary of research on test changes: An empirical basis for defining accommodations*. Lexington, KY: Mid-South Regional Resource Center.
- 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.
- Townsend, D. J., Carrithers, C., & Bever, T. G. (1987). Listening and reading processes in college- and middle-school age readers. In R. Horowitz & S. J. Samuels (Eds.), *Comprehending oral and written language* (pp. 217–242). New York: Academic.
- Wechsler, D. (1991). *Wechsler Individual Achievement Test*. San Antonio, TX: Psychological Corporation.
- Weston, T. J. (1999). Investigating the validity of the accommodation of oral presentation in testing. *Dissertation Abstracts International*, *60*(XX), 1083A.
- Woodcock, R. W. (1987). *Woodcock Reading Mastery Tests—Revised*. Circle Pines, MN: American Guidance Services.