Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Fall 2014, Vol. 33, No. 3, pp. 17–30

A Meta-Analysis of Research on the Read Aloud Accommodation

Heather Buzick and Elizabeth Stone, *Educational Testing Service*

Read aloud is a testing accommodation that has been studied by many researchers, and its use on K-12 assessments continues to be debated because of its potential to change the measured construct or unfairly increase test scores. This study is a summary of quantitative research on the read aloud accommodation. Previous studies contributed information to compute average effect sizes for students with disabilities, students without disabilities, and the difference between groups for reading and mathematics using a random effects meta-analytic approach. Results suggest that (1) effect sizes are larger for reading than for math for both student groups, (2) the read aloud accommodation increases reading test scores for both groups, but more so for students with disabilities, and (3) mathematics scores gains due to the read aloud accommodation are small for both students with and without disabilities, on average. There was some evidence to suggest larger effects in elementary school relative to middle and high school and possible mode effects, but more studies are needed within levels of the moderator variables to conduct statistical tests.

Keywords: accommodations, disabilities, meta-analysis, read aloud

Testing accommodations for students with disabilities are intended to remove construct-irrelevant barriers to accessing or responding to test items while retaining the ability to measure the intended construct (e.g., Koretz & Hamilton, 2006, p. 562). The Individuals with Disabilities Education Act Amendments of 1997, 20 U.S.C (1997) introduced regulations for providing testing accommodations to students with disabilities on K-12 state assessments. The read aloud accommodation (herein referred to as *read aloud*), in which test content is read out loud to test takers, is one that is allowed by some states and prohibited by others on reading assessments (National Center on Educational Outcomes [NCEO], 2011). Read aloud can be delivered in a variety of modes including prerecorded audio, a human reader in an individual or group setting, or computerized text-to-speech. State policies dictate whether an entire test or parts of it (test directions, item stems, or word categories such as proper nouns) can be read aloud. Read aloud has the potential to be useful for students with print disabilities (e.g., blindness, low vision, and reading-based learning disabilities). While it is one of the most frequently allowed accommodations on statewide assessments (Christensen, Lazarus, Crone, & Thurlow, 2008), it is also one of the most controversial because it can alter the construct being measured (e.g., items that define decoding, reading fluency, or spelling as part of the measured construct on reading assessments; items with formulas, fractions, or graphs on a mathematics assessment).

Research on changes in test scores associated with accommodation use is part of a broader collection of approaches to evaluate score comparability and obtain validity evidence for making inferences about the proficiencies of test takers with disabilities. Phillips (1994) argued that in the context of high-stakes testing (such as employment and licensure testing), an accommodation given to only some students is not appropriate if nondisabled students would benefit from it. The notion that an appropriate and effective accommodation increases scores on average for students who need the accommodation while the scores for students who do not need the accommodation remain unchanged has been termed the maximum potential thesis (see Zuriff, 2000) or the interaction hypothesis (see Sireci, Scarpati, & Li, 2005). In the context of K-12 instruction and assessment, some have posited that an accommodation may be appropriate even if students without disabilities would benefit from it, as long as students with disabilities who need the accommodation would receive a significantly greater score increase, on average (i.e., differential boost; Fuchs & Fuchs, 1999; see Cahalan-Laitusis, 2007).

Previous qualitative syntheses of research on differences in test scores associated with read aloud have offered no consensus on its effect (Kettler, 2012; Lai & Berkeley, 2012; Laitusis, Buzick, Stone, Hansen, & Hakkinen, 2012; Rogers, Christian, & Thurlow, 2012; Sireci et al., 2005). As part of a larger study, Elbaum (2007) computed the average effect size difference between students with and without disabilities. The results suggested that elementary-school students with learning disabilities experienced a greater score boost from the accommodation than did students without disabilities, but that at the secondary-school level, students without disabilities benefited more from the accommodation than students with disabilities. The mixed results underscore the need to conduct a meta-analysis to combine evidence across studies quantitatively (Wise, 2010).

The grade-level effect of read aloud identified by Elbaum (2007) is one example of a moderator variable that has been

Heather Buzick, Center for Validity Research, Educational Testing Service, MS 10-R, 660 Rosedale Road, Princeton, NJ 08541; hbuzick@ets.org. Elizabeth Stone, Center for Validity Research, Educational Testing Service, MS 09-R, 660 Rosedale Road, Princeton, NJ 08541.

hypothesized to have an impact on the strength or direction of effect sizes (e.g., Crawford & Tindal, 2004; Helwig & Tindal, 2003; Helwig, Rozek-Tedesco, & Tindal, 2002; Laitusis, 2010). Read aloud delivery mode is another potential moderator that has been studied (e.g., Flowers, Kim, Lewis, & Davis, 2011; Harris, 2008; Miranda, Russell, & Hoffman, 2004). Students' disabilities may also have an impact on effect sizes—some researchers have used all students receiving special education services (e.g., Crawford & Tindal, 2004; Helwig & Tindal, 2003) while others have chosen students with specific disabilities related to reading (e.g., Fletcher et al., 2006; Laitusis, 2010).

While results of individual research articles may or may not contribute support for the use of accommodations, meta-analysis can introduce greater statistical power to detect effects and provide context and a measure of consistency across studies. Our goal in conducting a meta-analysis of research on read aloud was to see if we could find conclusive evidence of the effect of read aloud by quantitatively summarizing effect sizes for both students with and without disabilities from existing research, testing the consistency of effects, and describing the contributions of moderators. We focused on four research questions: (1) does read aloud increase test scores for students with disabilities?; (2) does read aloud increase test scores for students without disabilities?; (3) does read aloud increase test scores more for students with disabilities than for students without disabilities?; and (4) what factors contribute to effect size differences? We answered these questions separately for reading and mathematics.

Method

Search and Inclusion Criteria

Our initial search was focused on finding research reports between the years 1998 and 2013 that compared scores on reading or mathematics assessments taken with and without a read aloud accommodation by K-12 students in the United States. We chose 1998 as the lower bound because IDEA included regulations on the use of testing accommodations on state assessments starting in 1997 (Individuals with Disabilities Education Act Amendments of 1997, 20 U.S.C, 1997). We began by searching the National Center on Educational Outcomes (NCEO) Accommodations Bibliography (NCEO, 2013) using the advanced search with keywords: "read aloud" and "read directions," and research purpose: "compare scores." We then reviewed all NCEO syntheses of research on accommodations, specifically focusing on quantitative score comparisons of read aloud (Cormier, Altman, Shyyan, & Thurlow, 2010; Johnstone, Altman, Thurlow, & Thompson, 2006; Rogers et al., 2012; Zenisky & Sireci, 2007). This did not add any articles to our original search but provided information to develop and test our selection criteria and helped us to validate the NCEO Accommodations Bibliography. We then searched for additional technical reports and other published and unpublished articles, including dissertations, through Google and Google Scholar as well as multidatabase online search engines (EBSCO and JSTOR). Keywords included variations of "oral accommodation," "audio presentation," "read aloud," "text-to-speech," "presentation accommodation," "test accommodation," "auditory accommodation," and "reading pen."

We included articles in the meta-analysis that met the following criteria: (1) means and standard deviations for test scores from the accommodated (read aloud) and from the standard administration were reported for one or more groups of students or could be derived from other study statistics or computed from raw data, (2) the effect of read aloud could be isolated from other accommodations with the exceptions of timing and setting, (3) the research design was a repeated-measures design (i.e., a single- or independent-groups pretest-posttest design) with counterbalanced administration conditions and parallel test forms or an independent-groups posttest only design with random assignment, and (4) the assessment was a large-scale mathematics or reading assessment (such as NAEP or the regular state assessment) or was created from items that were similar to large-scale assessments for the general population (i.e., not alternative assessments). We included the full range of disability types as well as research reports that only had students without disabilities. We included both types of study designs to increase the number of articles in our meta-analysis (see Morris & DeShon, 2002).

Our rationale for the second criterion includes two components. First, read aloud is often provided in an accommodation bundle along with other accommodations such as extended time or a small-group setting. This is because the use of read aloud may require a longer testing time (e.g., read aloud may be slower than silent reading, students may request repeated reading of some text) and because read aloud was administered by humans in many of the articles we considered, which requires a separate space from the main administration to avoid distraction and inappropriate administration to students who do not need the accommodation. The second component has to do with many timing and setting accommodations being viewed by states as not altering the construct being measured on state summative assessments. For example, K-12 assessments are practically untimed in most states so speededness is not part of the construct. In contrast, we did not include articles with a read aloud accommodation bundled with another potentially construct-altering accommodation such as calculator use for mathematics tasks or modified language for reading tasks.

Our initial search resulted in 85 articles. We conducted an initial round of filtering to eliminate reports in which it was straightforward to determine ineligibility based on information from the NCEO Accommodations Bibliography and study abstracts. This included a number of reports on mode effects and alternative assessments as well as qualitative research and single studies that appeared multiple times. After filtering these, we were left with 45 articles. Applying our selection criteria to the 45 quantitative research reports resulted in 19 that we determined to be appropriate to include in our meta-analysis: 11 on mathematics, six on reading, and two (Meloy, Deville, & Frisbie, 2002; Olson & Dirir, 2010) reporting on both subjects. Two articles (Calhoon, Fuchs, & Hamlett, 2000; E. L. Higgins & Raskind, 2005) included only students with disabilities, and two articles (Geraghty & Vanderwood, 2007; Helwig, Rozek-Tedesco, Tindal, Heath, & Almond, 1999) included only students without disabilities.[1] We excluded many of the 45 initially identified quantitative research articles on read aloud based on our criteria but found them to include useful validity evidence for read aloud, so we also created a table of characteristics of interest (see Appendix A).

Individual Effect Sizes

We defined comparisons as groups of students with summary information available for the accommodated and standard administrations. From the articles that included studies on mathematics, there were 19 independent comparisons for students with disabilities and 20 for students without disabilities.[2] From the articles that reported on reading, there were nine independent comparisons for students with disabilities and eight for students without disabilities. Six of the selected articles included multiple grade levels; we chose to treat each grade level independently to provide more information.

For comparisons obtained with an independent-groups posttest-only design, Hedges (1981) proposed the following effect size: $d = (\bar{X}_A - \bar{X}_S)/SD_{pooled}$, where $SD_{pooled} = \sqrt{\frac{(n_A-1)SD_A^2+(n_S-1)SD_S^2}{(n_A-1)+(n_S-1)}}$. The index A denotes the accommodated administration (read aloud) group and the index S denotes the standard administration group. $\bar{X}$ (with subscript A or S) is the sample average within each group, SD_S denotes the standard deviation of scores in the standard administration group, and SD_A denotes the standard deviation of scores in the accommodated administration group. For repeated-measures designs, we used the effect size formula proposed by Becker (1988), $d = (\bar{X}_A - \bar{X}_S)/SDS$, because the standard deviation in the standard condition is unaffected by the treatment (i.e., accommodation use) and is, therefore, more consistent across research reports (Becker, 1988). Further, the resulting effect size is on the same metric as the effect size for independent groups, allowing us to combine results from articles with different designs with minimal bias (Morris & DeShon, 2002). We adjusted each individual effect size using Hedges's (1981) correction factor (see Appendix B). This adjustment corrects for bias due to small sample sizes in the control group.

There are two options for comparing effect sizes across students with and without disabilities. Seven out of the eight articles with reading comparisons and 10 out of 13 articles with mathematics comparisons included both students with and without disabilities. Consequently, this permitted us to use the more favorable approach, which is to subtract the effect size for students without disabilities from the effect size for students with disabilities within individual comparisons ($\Delta = d_{SWD} - d_{SWOD}$). In total, we computed eight effect size differences in reading and 18 effect size differences in mathematics. This approach minimizes bias by using each comparison as its own control (Borenstein, Hedges, Higgins, & Rothstein, 2009). The limitation of this approach is that we could not include the few reports that only included one student group in that particular set of analyses.

Random-Effects Meta-Analysis Approach

For math and reading separately, we estimated three overall effects of the read aloud accommodation—for students with disabilities, students without disabilities, and the difference between the two groups (i.e., six models) using a random-effects approach to meta-analysis. The models are the same for individual effect sizes and effect size differences (d is replaced by Δ in the following equations). For a given model, for the ith independent comparison, the observed effect, d_i, is equal to the true effect, δ_i, plus sampling error e_i:

$$d_i = \delta_i + e_i, \tag{1}$$

where $\delta_i \sim N(\mu, \tau)$, with μ equal to the overall effect and τ equal to the variability in the true effects for individual comparisons. The key difference between a fixed effects and a random effects approach to meta-analysis is the assumption that the true effects may vary due to factors other than random error ($\tau \geq 0$), whereas in an fixed effects approach, no additional variability is assumed ($\tau = 0$). Consequently, in a random effects approach, there are two sources of error assumed: within and between comparisons (observed effect = overall effect + between comparison error + within comparison error).We estimated the overall effect in each model, μ, by computing a weighted average $\bar{d}$ of the observed effects defined as

$$\bar{d} = \frac{\sum w_i d_i}{\sum w_i}, \tag{2}$$

where the weight, w_i, is the inverse of the sum of the within-comparison variance, σ_{e_i}, and the between-comparison variance, τ. See Appendix B for the variance formulas.

In our meta-analysis, the sources of the within-comparison variability include sampling error (small samples of students, convenience sampling, and any clustering of sample students in classrooms, schools, or regions that is not accounted for in the study designs), the use of test forms that are not parallel across standard and accommodated administrations, and measurement error. For effect sizes from an independent-groups posttest-only design, an additional source of within-comparison variance is the random assignment of test takers to accommodation conditions—that is, the groups may differ on their average scores under the standard administration by chance. Studies with a repeated-measures design and counterbalanced administration conditions do not have this additional source of within-comparison error. However, we had to estimate the correlation between standard and accommodated administrations when it was not reported, which is an additional source of error from repeated measures comparisons (see Appendix B).

The between-comparison variability in our meta-analysis comes from a number of factors directly controlled by the researchers—grade level, types of assessment items, read aloud mode, study design, types of disabilities, and content read aloud, as well as other student factors such as reading and math ability, and other unobservable factors such as students' motivation. For students with disabilities, students without disabilities, and the difference between each group within each content area, we estimated the degree of heterogeneity among the effect sizes. We computed I^2 (J. Higgins, Thompson, Deeks, & Altman, 2003) to estimate the proportion of the observed variation in effect sizes that represents true differences. We also tested the hypothesis that all comparisons within groups and content areas share a common effect size with the Q statistic (Borenstein et al., 2009). We tested the difference in average effect sizes between mathematics and reading for both students with and without disabilities using z-tests (Borenstein et al., 2009). In computing the z-test statistics, we treated the comparisons in the report by Olson and Dirir (2010), which included both math and reading, as independent because they were conducted in separate states under different conditions, but removed the comparisons from Meloy et al. (2002) because the samples were the same across content areas.

Moderators

We hypothesized a priori that grade level (operationalized as elementary school [3–5], middle school [6–8], and high school [9–12]), read aloud delivery mode, disability type, extra time provided only on the accommodated administration, and content read aloud (e.g., entire test vs. directions) would be important moderator variables. We also expected that these factors might interact among each other. There were not enough comparisons within the levels of each of the moderators to test any statistical hypotheses; consequently, we used forest plots to visually display moderators and to inspect trends in the effect sizes. Given the possibility of interactions among the moderators, we decided that graphical display was a better approach than statistical tests which are likely to capitalize on chance when the number of research reports is small (e.g., Hunter & Schmidt, 2004).

As an exploratory analysis, we summarized effect sizes within grade bands. The grade-level variable, among all of our hypothesized variables, is the one most likely to be of interest in policy decisions because decoding is often part of the measured construct in the early grades. Due to the smaller number of comparisons within some of the grade bands, we used a fixed-effects approach to that analysis (e.g., Borenstein et al., 2009, p. 84). We obtained a description of the average effect within grade bands for the comparisons in our analysis but did not estimate true grade band effects and cannot generalize to conditions other than those represented in the articles included in our analysis. Differences in average effects across grade bands would suggest the need for more research at particular grade levels.

Results

Figure 1 is a forest plot showing effect sizes for the 56 comparisons across reading and math for both students with and without disabilities. Figure 2 shows the difference in effect sizes between groups as well as the average boost in reading and math. In each of the forest plots, point estimates of effect sizes and average effect sizes are denoted by squares. The error bars for represent the 95% confidence intervals, which indicate the range of sampling error for each estimated effect size (Figure 1) and effect size difference (Figure 2). The comparisons in the forest plots are listed by read aloud mode within grade band (i.e., elementary school, middle school, and high school). The values of the other moderators as well as the study design are also listed for all comparisons in Figure 1.

Read aloud increased reading assessment scores for both student groups in all comparisons: bias-corrected effect sizes ranged from .29 to 1.10 for students with disabilities and .03 to .70 for students without disabilities. The difference between groups, shown in Figure 2, ranged from –.12 to .75—in one study (Fletcher et al., 2009), students without disabilities benefitted more than students with disabilities. Across mathematics comparisons, bias-corrected effect sizes ranged from –.44 to .80 for students with disabilities, from –.19 to .46 for students without disabilities, and from –.64 to 1.01 for the difference between groups. The read aloud accommodation decreased average mathematics assessment scores for students with disabilities in two comparisons and for students without disabilities in four comparisons. Students without disabilities benefitted more than students with disabilities on the mathematics assessment in two comparisons.

The top of Table 1 shows the average effect sizes for mathematics and reading. The average effect sizes were significantly higher for reading than for math for both students with disabilities ($z = 4.94, p < .01$) and students without disabilities ($z = 2.31, p = .022$). For reading, average effects were positive for both students with and without disabilities ($\bar{d} = .56$ and .21) and for the difference in score gains between groups ($\overline{\Delta} = .34$). None of the 95% confidence intervals for the average effect contained zero; these effect sizes and confidence intervals for reading are consistent with the differential boost hypothesis, but do not satisfy the interaction hypothesis. For mathematics, we found significant but small, positive effects for students with and without disabilities ($\bar{d} = .13$ and .08) but no significant difference in score gains between groups ($\overline{\Delta} = .03$). This suggests that read aloud for mathematics assessments meets neither the interaction hypothesis nor the differential boost hypothesis.

The heterogeneity test results are reported at the bottom of Table 1. For both reading and math, the value of the Q statistic is statistically significant for both groups and also for the difference between groups, indicating heterogeneity among effect sizes. For reading, the computed I^2 indicates that 71% and 72% of the variation in effect sizes for students with and without disabilities, respectively, and 57% of the variation in the difference in effect sizes between the two groups is due to study characteristics (rather than random error). For math, we estimate that 73% and 79% of the variation in effect sizes for students with and without disabilities and 70% of the variation in effect sizes between the two groups is due to study characteristics (rather than random error). These percentages are considered high according to guidelines described in J. Higgins et al. (2003).

The video mode of delivery for mathematics had a small and sometimes negative effect relative to the other two modes; consequently, we also computed the overall effect sizes for math excluding comparisons with video (using random-effects). The results are reported in Table 1. Removing the comparisons with video as the read aloud mode (nine for students with disabilities, 10 for students without disabilities) slightly increased the average effect sizes for both student groups, but the difference between groups was still small ($\overline{\Delta} = .08$). Excluding comparisons that used video as the delivery mode decreased the percentage of variation explained due to factors other than random error only for students with disabilities.

Among the comparisons included in this meta-analysis, the effect sizes (computed via a fixed effects model) for both students with and without disabilities taking reading assessments were slightly larger for elementary school ($\bar{d} = .59$ and .17) than for middle school ($\bar{d} = .40$ and .09). The fixed-effects average effect size difference between students with and without disabilities was also slightly larger for elementary school ($\overline{\Delta} = .43$) than for middle school ($\overline{\Delta} = .30$). For mathematics, there was little difference between elementary ($\bar{d} = .15$ and 0) and middle school ($\bar{d} = .05$ and .05) for students with and without disabilities, respectively, but the fixed-effects average effect size difference between groups was larger for elementary school ($\overline{\Delta} = .17$) than for middle school ($\overline{\Delta} = -.05$).

Among the reading comparisons, there are no clear trends with respect to content read aloud, read aloud mode, or types of disabilities, but there are few comparisons within each moderator. Figure 2 illustrates that students with disabilities received a greater score increase than students without disabilities in most comparisons but that none of the differences in effect sizes were large (the point estimates were all less than .1). Although there are differences between elementary

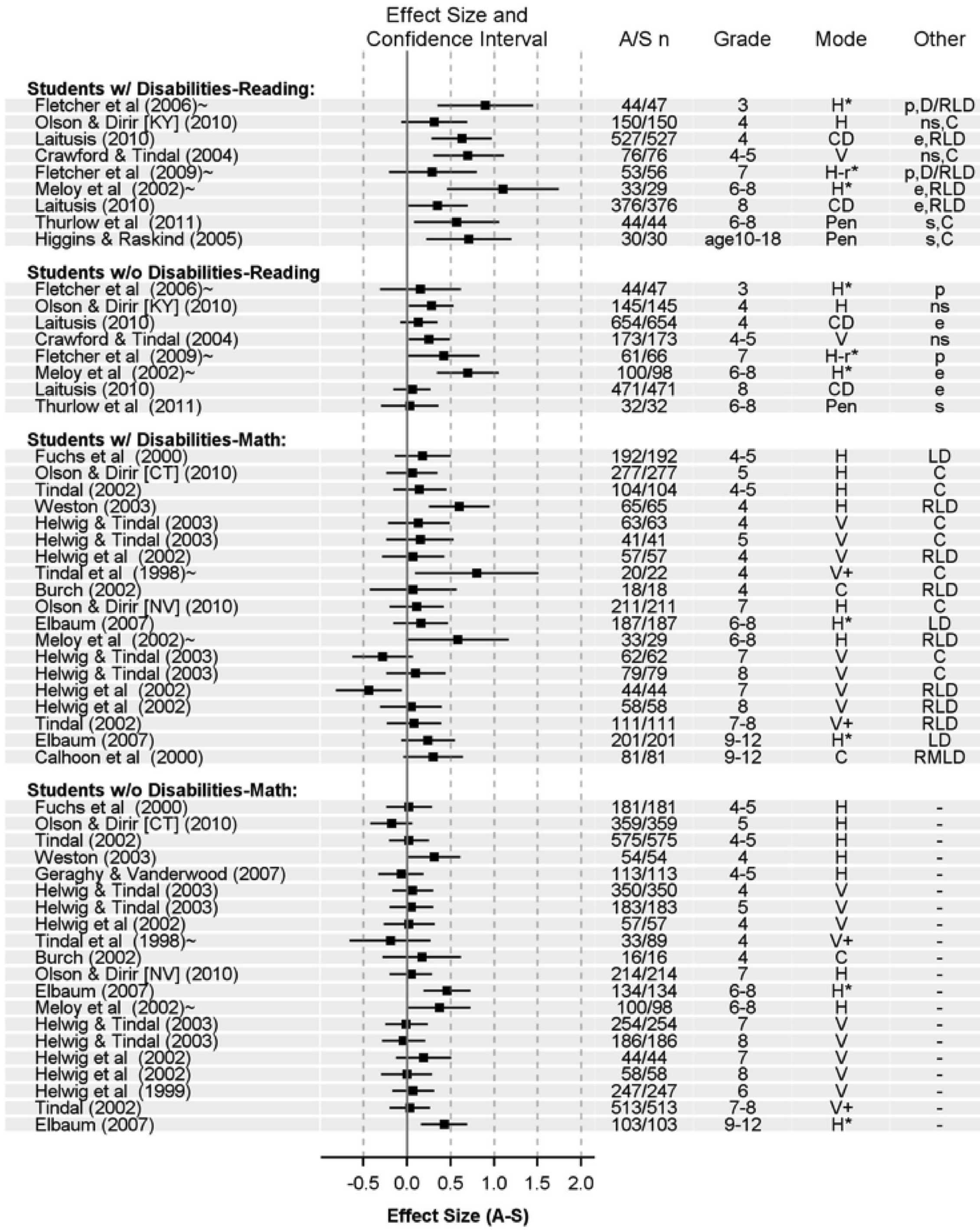

FIGURE 1. Effect sizes for individual comparisons.
Note: A = accommodated; S = standard; n = sample size. Mode: C = computer; CD = audio CD; H = human reader; H-r = human reader with restricted content; Pen = reading pen; V = video; V+ = video plus highlighting. *extra time on accommodated administration only. Other: content read aloud—ns = not specified; p = proper nouns and comprehension stems; e = entire test; s = student choice; disabilities—C = combined group with various types, IEP in reading and possibly math; D = dyslexia; LD = learning disabilities, type not specified; RLD = learning disabilities with IEP in reading or teacher-recommended reading accommodation; RMLD = learning disabilities with IEP in reading and math ~Independent groups posttest only design (blank is repeated measures).

school and middle school on average, this trend is not obvious across the individual effect size differences. For the mathematics comparisons, most of the confidence intervals for effect size differences between students with and without disabilities cover zero. Human read aloud improved scores for students with disabilities relative to students without disabilities more in elementary school than middle school. There were no other obvious trends. Based on the forest plots, we conclude that the moderator variables are interacting with each other or there are other unmeasured factors causing variation in effect size differences between students with and without disabilities.

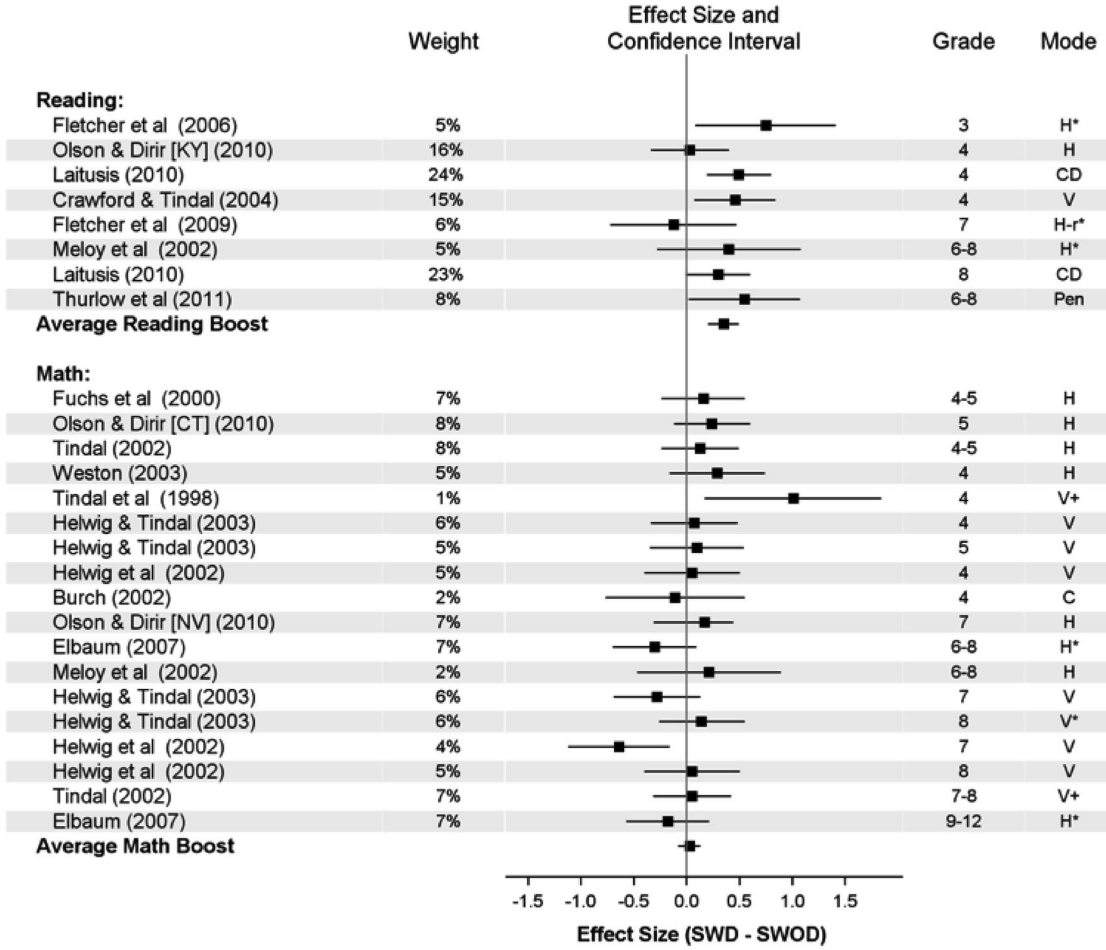

FIGURE 2. Forest plot for the difference in effects between students with and without disabilities.
Note: Effect sizes are equal to the effect size for students with disabilities (SWD) minus the effect size for students without disabilities (SWOD). The average effects for reading and math were computed using a random-effects approach. Error bands reflect the confidence intervals. Weights are the inverse of the sum of the within- and between-study variance. Mode: C = computer; CD = audio CD; H = human reader; H-r = human reader with restricted content; Pen = reading pen; V = video; V+ = video plus highlighting. *extra time on accommodated administration only.

Discussion

In this study, we quantitatively summarized previous research on read aloud used by students with and without disabilities on mathematics and reading assessments. We found that for reading assessments, although students with disabilities received a differential score boost of .34 standard deviation units, on average, scores increased by .21 standard deviation units for students without disabilities when using read aloud. These results call into question the fairness of read aloud for making high stakes decisions about individual students taking K-12 reading assessments, if only some students are permitted to use the accommodation. Policymakers ultimately decide the appropriate threshold for determining that an accommodation is acceptable for a given test purpose; however, our summary of research does not unequivocally support the use of read aloud on reading assessments by only some students if the test scores are being interpreted and used in the same way for all students, without additional validity evidence.

Read aloud on the reading assessment does appear to be effective at raising test scores for students with disabilities (by an average of .56 standard deviation units). This suggests that more evidence is needed on the validity of inferences based on the accommodated administration if test scores are used for purposes other than describing aggregate achievement levels. For example, if the state assessment is being used as a graduation requirement, then a predictive validity study would be needed to provide evidence that scores are comparable and have the same meaning for students using the read aloud accommodation on the reading assessment relative to those who do not use the read aloud accommodation.

For mathematics assessments, read aloud also increased scores for both student groups, but the average score gains were small for both groups (.13 and .08 standard deviation units, respectively). We found no evidence of differential boost from read aloud on mathematics assessments. The results from the 19 comparisons with video as the delivery mode suggest that, on average, video read aloud does not improve test scores on mathematics assessments for either student group. The small average effects of read aloud for mathematics assessments suggests that future work should focus on ways to improve the efficacy of read aloud for mathematics assessments, if it is to be administered at all (cf. Kettler, 2012).

Table 1. Effect Size Summary and Heterogeneity Test Results

	Students With Disabilities	Students Without Disabilities	Difference Between Students With and Without Disabilities
Average effect size:			
Reading	.56	.21	.35
	[.42, .70]	[.11, .31]	[.20, .49]
Math	.13	.08	.03
	[.05, .21]	[.01, .14]	[–.08, .13]
Math (Excluding comparisons with video)	.23	.14	.08
	[.13, .34]	[–.02, .30]	[–.09, .25]
Heterogeneity of effects:			
Reading			
Q	34.4	31.9	20.9
	($p < .01$)	($p < .01$)	($p < .01$)
I^2	71%	72%	57%
Math			
Q	73.1	102.1	63.1
	($p < .01$)	($p < .01$)	($p < .01$)
I^2	73%	79%	70%
Math (Excluding comparisons with video)			
Q	3432.5.7	93.9	35.6
	($p < .01$)	($p < .01$)	($p < .01$)
I^2	66%	88%	72%

Note: 95% confidence interval in brackets.

There was a significant amount of unexplained variation in effect sizes for both reading and mathematics, indicating that average effects should be interpreted with caution. Our study points to the need for more creative ways to both study and administer read aloud. In terms of research, looking at item-level effects, for example, rather than total score, may provide information about particular item types that are more or less influenced by read aloud. This will become possible as test delivery increasingly moves to computer- or tablet-based platforms that can record accommodation use at the item level. More studies in each grade band would make it possible to estimate the true effect size within each grade band, which would be useful in deciding on the appropriateness and efficacy of read aloud in certain grades. Research on computer-delivered read aloud (i.e., text-to-speech) would also be useful since there are currently not enough eligible studies to provide a meaningful summary and this mode is being increasingly considered for use operationally.

There are several important limitations to this meta-analysis. Our decision to treat grade level independently and include multiple comparisons from the same study in each of our six models may have contributed to a downward bias on the standard errors. As with any meta-analysis, the influence of past studies on subsequent studies also may have caused a dependency among comparisons and biased standard errors. While we minimized other dependencies across comparisons by estimating random effects models separately for students with and without disabilities by content area, the cost of this decision was that the number of studies was too small to perform statistical tests of the influence of moderator variables on the overall effect sizes from the six models. Finally, unlike randomized trials or studies with national samples, most of the research on accommodations comes from convenience samples (e.g., students from several schools); consequently, the estimated average effects of read aloud may not generalize to all students in the United States.

Given the increased use of technology to deliver assessments, read aloud delivered via digital or human-based text-to-speech will likely be the most common delivery mode on future assessments—because of the limited number of studies on this mode, our results may not generalize to future uses of read aloud delivered via text-to-speech. However, as studies increasingly employ this delivery method, they should be incorporated into future meta-analyses. Finally, because of the limited number of comparisons within each level of the theoretically important moderator variables, we were not able to statistically test the impact of disability status, delivery mode, or grade level on the effect of read aloud, nor were we able to model their interaction.

Describing changes in test scores is just one aspect of the validity argument in support of administering read aloud. Appropriate use of read aloud requires multiple types of validity evidence to support each specific test purpose. To further provide policymakers with information about the validity of inferences drawn from accommodated assessments, we encourage researchers to consult Appendix A when designing future studies on read aloud. Because of the large percentage of variation in effect sizes, our understanding of the effect of read aloud would benefit from additional studies (e.g., an experimental study on text-to-speech and one that compares grade-level effects on the reading assessment) that replicate some of the factors in previous research. Further, it is important that researchers are as thorough as possible when describing their research designs. As is clear from Figure 1, some studies did not specify features such as disability subtype or read aloud mode, and this reduced our ability to identify possible trends. In collecting validity evidence, it is also critical to know what parts of the test were read aloud in order to determine whether the accommodation changes the construct being measured.This information will be useful in collecting further evidence for evaluating the appropriateness of testing accommodations.

Notes

[1]Helwig et al. (1999) combined students without disabilities with students receiving special education services in mathematics. Students with disabilities represented 12% of the sample—we chose to categorize this study as students without disabilities only, because students were not receiving special education services for reading.

[2]Helwig et al. (2002) and Helwig and Tindal (2003) report means and standard deviations separately by form and order of administration (standard and accommodated). Groups were randomly assigned; therefore, we combined the means and pooled the standard deviations to compute single comparisons within grade levels. Calhoon et al. (2000) administered three different types of read aloud to all students: human reader, text-to-speech delivered via computer, and text-to-speech delivered via computer in addition to short video clips with extra information. Because the comparisons were not independent, we chose to use the text-to-speech delivered via computer comparison in the meta-analysis because this will likely be the most common delivery mode on future assessments.

Appendix A

Quantitative Studies on Read Aloud Not Included in Meta-Analysis

Reading						
Report Authors	**Design**	**Assessment**	**Grade(s)**	**Disabilities**	**Type of Read Aloud**	**What was Read Aloud?**
Bielinkski et al. (2001)[a]	OD	State Test Items: Reading Comprehension	3	Reading-Based Learning Disabilities	Human Reader in Small Group or Individual Setting, Possibly with Extended Time	Test Items and Response Options
Bolt & Ysseldyke (2006)[a]	OD	State Test Items: Reading/Language Arts	3, 7, 11	Not Specified	Not Specified	Not Specified
Cook et al. (2010)[a]	OD	State Test Items: ELA	4	Learning Disabilities	Not Specified	Not Specified
Cook et al. (2009)[a]	OD	Large-Scale Reading Comprehension	4, 8	Reading-Based Learning Disabilities	Prerecorded Audio CD	Entire Test
Elliott et al. (2010)[b]	RM	Modified Assessment: Reading Comprehension and Vocabulary	8	Not Specified	Recorded Voice	Directions and Items, Excluding Key Vocabulary Words
Flowers et al. (2011)[a]	OD	State Assessment: Reading	3–11	Any	Human Reader/Digital Text Readable with Text or Screen Reader	Entire Test
Harker & Feldt (1993)[c]	RM	Interpretation of Literary Materials; Vocabulary	High School	Students without Disabilities	Prerecorded Audio Cassette	Entire Test
Harris (2008)[a]	OD	State Assessment: ELA	6–8	Any	Human Reader or Audio CD	Entire Test
Huynh & Barton (2006)[a]	OD	High School Exit Exam	10	Any	Human Reader or Audio CD	Not Specified
Kettler et al (2011)[b]	RM	Modified Assessment: Reading Comprehension and Vocabulary	8	Not Specified	Recorded Voice on a Computer-Based Test	Directions and Items, Excluding Key Vocabulary Words

(Continued)

Reading						
Report Authors	**Design**	**Assessment**	**Grade(s)**	**Disabilities**	**Type of Read Aloud**	**What was Read Aloud?**
Kosciolek & Ysseldyke (2000)[d]	RM	Large-Scale Reading Comprehension	3–5	Any	Prerecorded Audio Cassette	Entire Test
McKevitt & Elliott (2003)[e]	RM	Large-Scale Reading Assessment	8	Reading-Based Learning Disabilities	Prerecorded Audio Cassette	Entire Test
Middleton (2007)[e]	IG	Large-Scale Reading Comprehension	4, 8	Reading-Based Learning Disabilities	Prerecorded Audio CD	Entire Test
Olson & Dirir (2010)[c]	RM	Items Similar to State ELA Assessment	7	Students Eligible for a Text Reader Accommodation	Text-to-speech (Computerized)	Not Specified
Randall & Engelhard (2010)[c]	IG	State Assessment: Reading	4–7	Any Disabilities, Mostly Learning Disabilities and a Large Percentage of Speech-Language Impairments in Grade 4	Human Reader	Entire Test

Note: OD = observational data, IG = independent groups, RM = repeated measures. Operational data indicates study evaluated observational test data, with test takers assigned accommodations based on need as required by the state. ELA is English Language Arts. Reasons for exclusion in the meta-analysis:

[a]students not randomly assigned to groups.
[b]effect of read aloud was conflated with characteristics of the modified assessment.
[c]standard and accommodated conditions not counterbalanced.
[d]not enough information to compute effect sizes.
[e]same data as Laitusis (2010).
[f]read aloud bundled with other accommodations.

Mathematics						
Report	**Design**	**Assessment**	**Grade(s)**	**Disabilities**	**Type of Read Aloud**	**What Was Read Aloud?**
Anjorin (2009)[a]	OD	High School Exit Exam: Mathematics	10	Learning Disabilities	Human Reader	Directions
Bielinski et al. (2001)[a]	OD	State Test Items: Mathematics	4	Students whose Primary Disability is in Reading	Human Reader in Small Group or Individual Setting, Possibly with Extended Time	Entire Test
Bolt & Ysseldyke (2006)[a]	OD	State Test Items: Mathematics	4, 8, 10	Subtypes Not Specified	Not Specified	Not Specified
Elliott et al. (2010)[b]	IG	Modified Assessment: Mathematics	8	Not Specified	Recorded Voice on a Computer-Based Test	Directions and Items, Excluding Key Vocabulary Words
Elliott et al. (2009)[c]	RM	Four Math Performance Tasks	4	Not Specified	Human Reader	Entire Test
Huynh et al. (2004)[a]	OD	High School Exit Exam: Mathematics	10	Multiple Types, and Significant Reading Disabilities	Multiple Types	Not Specified
Johnson (2000)[d]	RM	State Test Items: Mathematics	4	Reading-Based Learning Disabilities	Human Reader	Entire Test
Kettler et al. (2011)[b]	IGs	Modified Assessment: Mathematics	8	Types Not Specified	Recorded Voice on a Computer-Based Test	Directions and Items, Excluding Key Vocabulary Words

(Continued)

Mathematics						
Report	**Design**	**Assessment**	**Grade(s)**	**Disabilities**	**Type of Read Aloud**	**What Was Read Aloud?**
Ketterlin-Geller et al. (2007)[d]	RM	Items Aligned with State Mathematics Standards	3	Various Disability Types	Audio-Recorded Human Reader Available via Speaker Button in Computer-Based Platform	Text-Based Item Stems and Response Options
Lee & Tindal (2000)[e]	IG	State Test Items: Mathematics	4	Students in Special Education Classrooms with Various IEPs	Human Reader	Entire Test
Pomplum & Omar (2000)[a]	OD	State Test Items: Mathematics	4	Learning Disabilities	Human Reader	Not Specified
Schnirman (2005)[d]	RM	State Test Items: Math Concepts, Problem Solving	6, 7	Learning Disabilities	Human-Recorded Audiocassette	Entire Test
Schulte et al. (2001)[c]	RM	Items Aligned with National Council of Teachers of Mathematics Standards	4	Various Disability Subtypes, Including Learning Disabilities	Human Reader	Entire Test

Note: OD = observational data, IG = independent groups, RM = repeated measures. Operational data indicates study evaluated observational test data, with test takers assigned accommodations based on need as required by the state. ELA is English Language Arts. Reasons for exclusion in the meta-analysis:
[a]students not randomly assigned to groups.
[b]effect of read aloud was conflated with characteristics of the modified assessment.
[c]read aloud bundled with other accommodations.
[d]standard and accommodated conditions not counterbalanced.
[e]groups include both students with and without disabilities.

Appendix B

We used Microsoft Excel for all computations in the meta-analysis.

Bias Correction

We multiplied each effect size by the approximation to the degrees of freedom (*df*)-dependent bias-correction factor proposed by Hedges (1981) to produce an unbiased estimator of the population effect size. The correction factor, $c(df)$, is calculated as $c\ (df) = 1 - \frac{3}{4df - 1}$ where $n =$ number of students and $df = n$-1 for repeated-measures designs, and $n = (n_A n_S) / (n_A + n_S)$ and $df = n_A + n_S - 2$ where n_A is the number of students taking the accommodated administration and n_S is the number of students taking the standard administration for independent-groups designs.

Within-Study Sampling Variance

We computed the within-study sampling variance for independent groups (Hedges, 1981) as

$$\hat{\sigma}^2_{e_i} = \left(\frac{1}{n}\left(\frac{N-2}{N-4}\right)\left(1 + n\mu^2\right)\right) - \frac{\mu^2}{\left[c\ (N-2)\right]^2}, \qquad (3)$$

where μ is the overall effect, N is the combined number of students in the groups and $c(df)$ and n as previously defined. We computed the within-study sampling variance for repeated measures as

$$\hat{\sigma}^2_{e_i} = \left[\frac{2\,(1-\rho)}{n}\right]\left(\frac{n-1}{n-3}\right)\left[1 + \frac{n}{2\,(1-\rho)}\mu^2\right] - \frac{\mu^2}{\left[c\ (n-1)\right]^2}, \qquad (4)$$

with μ and n as previously defined, and ρ the correlation between scores from the standard and accommodated administrations. Note that the formula is misprinted in Becker (1988). We used the correct formula from Morris and DeShon (2002).

In the variance formulas, we estimated the overall effect size, μ, with the unweighted average effect size from the individual comparisons (Morris & DeShon, 2002) and estimated the population correlation, ρ, by r, the Pearson correlation coefficient.

Between-Study Sampling Variance

We estimated the between-study sampling variance (Borenstein et al., 2009) as $= \frac{Q - df}{C}$, where $df =$ (number of comparisons-1), $C = \sum_i w_i - \frac{\sum_i w_i^2}{\sum_i w_i}$ (scaling factor), weight $w_i = 1/\hat{\sigma}^2_{e_i}$, $Q = \sum_i w_i d_i^2 - \frac{\left(\sum_i w_i d_i\right)^2}{\sum_i w_i}$ is the total observed variance, and d_i is the observed effect.

Of the six comparisons from repeated-measures studies for reading for students with disabilities, three reported r. Two studies reported information to compute r using the formula (Morris & DeShon, 2002),

$$r = \frac{SD_S^2 + SD_A^2 - SD_D^2}{(2)\,(SD_S)\,(SD_A)} \qquad (5)$$

where SD_A and SD_S are the standard deviation of accommodated and standard administration scores, respectively, and $SD_D =$ standard deviation of difference scores, computed as $SD_D^2 = \frac{n\left(\bar{X}_A - \bar{X}_S\right)^2}{t^2_{RM}}$ where $\bar{X}_S$ is the mean score under the standard administration, $\bar{X}_A$ is the mean score under the accommodated administration, and t_{RM} is the *t*-statistic for a repeated-measures *t*-test. One reading study (Crawford & Tindal, 2004) did not report a correlation nor did it report additional statistics to derive the correlation. We estimated the correlation following Hedges and Olkin (1985; see also Morris & DeShon, 2002), with details available from the first author of this study. For students without disabilities, correlations were available for three comparisons, and two had to be estimated.

For the math studies, we used r as reported in the studies for seven comparisons out of 18 for students with and without disabilities. For the remaining 11 within each group, the studies did not provide enough additional information to compute the correlation. We applied the same approach as for the reading studies to estimate the correlation. For both reading and math, the meta-analysis results were robust to a wide range of values for estimates of ρ.

References

*Denotes reports that are included in the meta-analysis.

Anjorin, I. (2009). *High-stakes tests for students with specific learning disabilities: Disability-based differential item functioning*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses databases. (UMI No. 3390841)

Becker, B. J. (1988). Synthesizing standardized mean change measures. *British Journal of Mathematical and Statistical Psychology*, *41*, 257–278.

Bielinski, J., Thurlow, M., Ysseldyke, J., Friedenbach, J., & Friedenbach, M. (2001). *Read-aloud accommodations: Effects on multiple-choice reading and math items*. Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Bolt, S. E., & Ysseldyke, J. E. (2006). Comparing DIF across math and reading/language arts tests for students receiving a read aloud accommodation. *Applied Measurement in Education*, *19*, 329–355.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

*Burch, M. A. (2002). *Effects of computer-based test accommodations on the math problem-solving performance of students with and without disabilities* (Doctoral dissertation). Retrieved from ProQuesr Dissertations and Theses database. (UMI No. 3047429)

Cahalan-Laitusis, C. (2007). Validity and accommodations: A variety of approaches to accessible assessments. In C. Cahalan-Laitusis & L. L. Cook (Eds.), *Large scale assessment and accommodations: What works* (pp. 71–83). Washington, DC: Council for Exceptional Children.

*Calhoon, M. B., Fuchs, L. S., & Hamlett, C. L. (2000). Effects of computer-based test accommodations on mathematics performance assessments for secondary students with learning disabilities. *Learning Disability Quarterly*, *23*, 271–282.

Christensen, L. L., Lazarus, S. S., Crone, M., & Thurlow, M. L. (2008). *2007 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 69). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Cook, L., Eignor, D., Sawaki, Y., Steinberg, J., & Cline, F. (2010). Using factor analysis to investigate accommodations used by students with disabilities on an English-language arts assessment. *Applied Measurement in Education*, *23*, 187–208.

Cook, L., Eignor, D., Steinberg, J., Sawaki, Y., & Cline, F. (2009). Using factor analysis to investigate the impact of accommodations on the scores of students with disabilities on a reading comprehension assessment. *Journal of Applied Testing Technology*, *10*(2). (Retrieved from http://www.jattjournal.com/index.php/atp/index

Cormier, D. C., Altman, J. R., Shyyan, V., & Thurlow, M. L. (2010). *A summary of the research on the effects of test accommodations: 2007–2008* (Technical Report 56). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Crawford, L., & Tindal, G. (2004). Effects of a read aloud modification on a standardized reading test. *Exceptionality*, *12*, 89–106.

*Elbaum, B. (2007). Effects of an oral testing accommodation on the mathematics performance of secondary students with and without learning disabilities. *Journal of Special Education*, *40*, 218–229.

Elliott, S. N., Kettler, R. J., Beddow, P. A., Kurz, A., Compton, E., McGrath, D.,. . .Roach, A. T. (2010). Effects of using modified items to test students with persistent academic difficulties. *Exceptional Children*, *76*, 475–495.

Elliott, S. N., Kratochwill, T. R., McKevitt, B. C., & Malecki, C. K. (2009). The effects and perceived consequences of testing accommodations on math and science performance assessments. *School Psychology Quarterly*, *24*, 224–239.

*Fletcher, J. M., Francis, D. J., O'Malley, K., Copeland, K., Mehta, P., Caldwell, C.J., & Vaughn, S. (2009). Effects of a bundled accommodations package on high-stakes testing for middle school students with reading disabilities. *Exceptional Children*, *75*, 447–463.

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

Flowers, C., Kim, D.-H., Lewis, P., & Davis, V. C. (2011). A comparison of computer-based testing and pencil-and-paper testing for students with a read-aloud accommodation. *Journal of Special Education Technology*, *26*, 1–12.

Fuchs, L. S., & Fuchs, D. (1999). Fair and unfair testing accommodations. *School Administrator*, *56*(10), 24.

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

*Geraghty, C. A., & Vanderwood, M. (2007). *Effects of a mathematics read aloud accommodation for students with high and low reading skills*. University of California, Riverside. Retrieved July 29, 2014, from http://www.stanswartz.com/Presentations_Cathleen/JSEarticlesubmission.pdf

Harker, J. K., & Feldt, L. S. (1993). A comparison of achievement test performance of nondisabled students under silent reading and reading plus listening modes of administration. *Applied Measurement in Education*, *6*, 307–320.

Harris, L. W. (2008). Comparison of student performance between teacher read and CD-ROM delivered modes of test administration of English language arts tests (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3321402)

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

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

*Helwig, R., Rozek-Tedesco, M. A., Tindal, G., Heath, B., & Almond, P. (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.

*Higgins, E. L., & Raskind, M. H. (2005). The compensatory effectiveness of the Quicktionary Reading Pen II on the reading comprehension of students with learning disabilities. *Journal of Special Education Technology*, *20*, 31–40.

Higgins, J., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analysis. *British Medical Journal*, *327*, 557–560.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.

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

Huynh, H., Meyer, J. P., & Gallant, D. J. (2004). Comparability of student performance between regular and oral administrations for a high-stakes mathematics test. *Applied Measurement in Education*, *17*, 39–57.

Individuals with Disabilities Education Act Amendments of 1997, 20 U.S.C. § 1400 (1997).

Johnson, E. S. (2000). The effects of accommodations on performance assessments. *Remedial and Special Education*, *21*, 261–267.

Johnstone, C. J., Altman, J., Thurlow, M. L., & Thompson, S. J. (2006). *A summary of research on the effects of test accommodations: 2002 through 2004* (Technical Report 45). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

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

Kettler, R. J. (2012). Testing accommodations: Theory and research to inform practice. *International Journal of Disability, Development and Education*, *59*(1), 53–66.

Kettler, R. J., Rodriguez, M. C., Bolt, D. M., Elliott, S. N., Beddow, P. A., & Kurz, A. (2011). Modified multiple-choice items for alternate assessments: Reliability, difficulty, and differential boost. *Applied Measurement in Education*, *24*, 210–234.

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education and Praeger.

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.

Lai, S. A., & Berkeley, S. (2012). High-stakes test accommodations: Research and practice. *Learning Disability Quarterly*, *35*, 158–169.

*Laitusis, C. C. (2010). Examining the impact of audio presentation on tests of reading comprehension. *Applied Measurement in Education*, *23*, 153–167.

Laitusis, C., Buzick, H., Stone, E., Hansen, E., & Hakkinen, M. (2012). Literature review of testing accommodations and accessibility tools for students with disabilities. Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/08/Smarter-Balanced-Students-with-Disabilities-Literature-Review.pdf

Lee, D., & Tindal, G. (2000). *Differential item functioning (DIF) as a function of test accommodation*. Newark: University of Delaware Education Research and Development Center.

McKevitt, B., & 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*, 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*, 248–255.

Middleton, K. V. (2007). *The effect of a read-aloud accommodation on items on a reading comprehension test for students with reading-based learning disabilities* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3281388)

Miranda, H., Russell, M., & Hoffman, T. (2004). Examining the feasibility and effect of a computer-based read-aloud accommodation on mathematics test performance. Retrieved July 29, 2014, from http://www.bc.edu/research/intasc/researchprojects/enhanced_assessment/PDF/EAP_ReadAloud.pdf

Morris, S. B, & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*(1), 105–125.

National Center on Educational Outcomes (NCEO). (2011, April). *Developing common accommodations policies: Discussion points for consortia* (NCEO Brief #2). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

National Center on Educational Outcomes (NCEO). (2013). Accommodations bibliography. Retrieved July 29, 2014, from https://apps.cehd.umn.edu/nceo/accommodations/AdvancedSearch.aspx

*Olson, J. F., & Dirir, M. (2010). *Technical report for studies of the validity of test results for test accommodations—Establishing the validity of test accommodations and score interpretations for students with disabilities: A collaboration of state-based research*. Washington, DC: Council of Chief State School Officers.

Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education*, *7*, 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*, 21–29.

Randall, J., & Engelhard, G., Jr. (2010). Performance of students with and without disabilities under modified conditions. *Journal of Special Education*, *44*, 79–93.

Rogers, C. M., Christian, E. M., & Thurlow, M. L. (2012). *A summary of the research on the effects of test accommodations: 2009–2010* (Technical Report 65). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Schnirman, R. K. (2005). *The effect of audiocassette presentation on the performance of students with and without learning disabilities on a group standardized math test* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3180285)

Schulte, A. G., Elliott, S. N., & Kratochwill, T. R. (2001). Effects of testing accommodations on standardized mathematics test scores: An experimental analysis of the performances of students with and without disabilities. *School Psychology Review*, *30*, 527–547.

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

*Thurlow, M., Moen, R., Lekwa, A., & Scullin, S. (2011, April). *Exploring a partial auditory accommodation for reading assessment*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

*Tindal, G. (2002). *Accommodating mathematics testing using a videotaped, read-aloud administration*. Washington, DC: Council of Chief State School Officers.

*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. J. (2003). *The validity of oral accommodation in testing: NAEP Validity Studies*. Washington, DC: National Center for Education Statistics.

Wise, L. L. (2010). Accessible reading assessments for students with disabilities: Summary and conclusions. *Applied Measurement in Education*, *23*, 209–214.

Zenisky, A. L., & Sireci, S. G. (2007). *A summary of the research on the effects of test accommodations: 2005–2006* (Technical Report 47). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

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

