Does Use of Text-to-Speech and Related Read-Aloud Tools Improve Reading Comprehension for Students With Reading Disabilities? A Meta-Analysis

Sarah G. Wood, MS¹, Jerad H. Moxley, PhD¹, Elizabeth L. Tighe, PhD¹, and Richard K. Wagner, PhD¹

Abstract

Text-to-speech and related read-aloud tools are being widely implemented in an attempt to assist students’ reading comprehension skills. Read-aloud software, including text-to-speech, is used to translate written text into spoken text, enabling one to listen to written text while reading along. It is not clear how effective text-to-speech is at improving reading comprehension. This study addresses this gap in the research by conducting a meta-analysis on the effects of text-to-speech technology and related read-aloud tools on reading comprehension for students with reading difficulties. Random effects models yielded an average weighted effect size of \( d = .35 \), with a 95% confidence interval of .14 to .56, \( p < .01 \). Moderator effects of study design were found to explain some of the variance. Taken together, this suggests that text-to-speech technologies may assist students with reading comprehension. However, more studies are needed to further explore the moderating variables of text-to-speech and read-aloud tools’ effectiveness for improving reading comprehension. Implications and recommendations for future research are discussed.

Keywords

reading comprehension, text-to-speech, reading disabilities, technology, meta-analysis

Reading comprehension, which is defined as the ability to construct meaning from interacting with a text, is critical for students to succeed in today’s educational settings (Snow, 2002). For students with reading disabilities, reading comprehension is often difficult (Kim, Linan Thompson, & Misquitta, 2012). Most current theories argue that one of the primary causes of reading disabilities is a struggle to decode written text. This has a direct negative effect on reading comprehension by decreasing word reading accuracy and speed. Inefficient decoding may also tax cognitive resources, leaving fewer resources available for comprehension (Smythe, 2005).

Presenting reading material orally in addition to presenting it in a traditional paper format removes the need to decode reading material and therefore has the potential to help students with reading disabilities better comprehend written texts. There are several technologies for presenting oral materials (e.g., text-to-speech, reading pens, audiobooks). Previously text was available orally through books-on-tape and human readers. More recently, text-to-speech technology has been used widely in educational settings from elementary school through college. Unfortunately, its implementation has outpaced the lagging research base on the effects of using text-to-speech to support comprehension. The research literature is characterized by contradictory results, with some studies reporting improved reading and others not (Dalton & Strangman, 2006; Stetter & Hughes, 2010; Strangman & Hall, 2003). Due to these mixed findings, we wanted to conduct a current review of the literature. The goal of this meta-analysis is to synthesize the research literature on the effects of text-to-speech and related tools for oral presentation of material on reading comprehension for students with reading disabilities.

Individuals with reading disabilities have unexpected significant deficits in reading and its component skills (e.g., decoding, fluency) despite potential educational opportunities. This reading deficit is believed to have a neurobiological basis and is also characterized by a failure to respond to appropriate instruction and intervention (Fletcher, 2009; ²Florida State University, Tallahassee, USA

Corresponding Author:
Sarah G. Wood, MS, Department of Psychology, Florida State University, 1107 W. Call Street, Tallahassee, FL 32306-4301, USA.
Email: sgw12@my.fsu.edu
practitioners and education researchers have begun to use text-to-speech technology development began in the 1980s and has been rapidly increasing. With technological advances, text-to-speech and related tools are used in both interventions and compensation settings. We included both intervention and compensation studies in this review but also added a moderator variable to represent this distinction to determine whether comparable effects emerged from both kinds of studies. Additionally, some students with reading disabilities are relatively accurate at decoding but are slow readers and may be unable to keep up with reading assignments. This becomes increasingly true as students move into more advanced course work (i.e., college reading) (Jones, Schwilk, & Bateman, 2012). Text-to-speech may enable students to read material that is beyond their basic word-reading level while matching their interests and listening comprehension abilities.

**Intervention Versus Compensation**

Intervention studies seek to improve students’ reading skills independent of the technology. In contrast, compensation studies seek to provide students with an assistive tool to help them with their reading. Intervention-oriented studies use text-to-speech tools to improve unassisted reading skills, whereas compensation-oriented studies address the use of text-to-speech tools to compensate for word-level skill deficits and gain access to written material. It is often unclear how to tell when intervention has failed and when students should begin to use compensation (Edyburn, 2007). Although educators would like to improve students’ basics skills as much as possible, not all students are able to reach proficient reading skills. The goal of compensation is designed to help students access texts. If comprehension problems of students with reading disabilities stem at least in part from decoding deficits, then reducing or eliminating the decoding requirement should improve reading comprehension. Oral presentation of material including text-to-speech helps eliminate the decoding requirement by reading the words aloud to the student, thus enabling comprehension (Olson, 2000). Text-to-speech and read-aloud text presentation tools are used in both interventions and compensation settings. We included both intervention and compensation studies in this review but also added a moderator variable to represent this distinction to determine whether comparable effects emerged from both kinds of studies.

Additionally, some students with reading disabilities are relatively accurate at decoding but are slow readers and may be unable to keep up with reading assignments. This becomes increasingly true as students move into more advanced course work (i.e., college reading) (Jones, Schwilk, & Bateman, 2012). Text-to-speech may enable students to read material that is beyond their basic word-reading level while matching their interests and listening comprehension abilities.

**Use of Text-to-Speech and Other Read-Aloud Tools in Education**

Text-to-speech technology development began in the 1980s and has been rapidly increasing. With technological advances, practitioners and education researchers have begun to use text-to-speech and related tools to help students with reading disabilities. In text-to-speech, the text is paired either with a synthesized computer or human recording of that same text (Biancarosa & Griffiths, 2012). As computer technology has improved over the past 20 years, the prevalence and quality of electronic versions of books and computer software with text-to-speech capabilities have also increased. Text-to-speech software development is occurring worldwide in diverse locations, including the United States, Egypt, and Europe (Smythe, 2005). Examples of such programs are DecTalk, ClassMate Reader, Texthelp Read & Write Gold, and Kurzweil 3000 (Berkeley & Lindstrom, 2011). The number of free and easily accessible text-to-speech software programs is increasing (Berkeley & Lindstrom, 2011). There are many statewide implementations, including examples from Kentucky and Missouri (Goddard, Kaplan, Kuehnle, & Beglau, 2007; Hasselbring & Bausch, 2005). These programs commonly contain voice options, custom pronunciation, creation of synthetic audio files, and other assistive tools such as text highlighting (Peters & Bell, 2007). Text-to-speech technology features can greatly affect the users’ experience and thus may affect its effectiveness. Some of these features include reading rate, voice type, document tagging (which affects reading order), and dynamic highlighting. Reading rate can greatly affect the users’ experience (Lionetti & Cole, 2004). When users choose text-to-speech software, a consistent comment is the desire to have software that follows natural speech and prosody. These subtle differences of reading rate, voice, document formatting, and dynamic highlighting between text-to-speech systems may affect overall user experience and thus may contribute to the mixed findings in the literature.

In addition to different features of text-to-speech tools, individual differences in users may affect the users’ text-to-speech experience. Individuals with clinical diagnoses of reading disabilities are heterogeneous. Some may be more skilled at decoding than others and thus may benefit differently from text-to-speech/read-aloud tools. Also, many have comorbid diagnoses, such as attention deficit/hyperactivity disorder, which may affect their reading performance. In addition, personality and social factors interacting with their disabilities may either facilitate or decrease text-to-speech use (Parette & Scherer, 2004; Parr, 2012). Finally, having supportive and knowledgeable technical help with the software, particularly at school, increases students’ ability to use and integrate this technology effectively into their daily lives (King Sears, Swanson, & Mainzer, 2011; Newton & Dell, 2009). These differences may lead to different compensation patterns, which may affect interactions with read-aloud/text-to-speech accommodations.

As previously mentioned, the results of several reviews of the literature have not yielded consistent findings about whether text-to-speech improves comprehension (e.g., Hall, Hughes, & Filbert, 2000; MacArthur, Ferretti, Okolo, &
Additionally, we considered small-studies, sometimes comprehensions for individuals with reading disabilities. The effects of text-to-speech and related tools on reading comprehensions. The focus of the current meta-analysis was scale standardized tests that were not specific to reading disabilities for students without disabilities. Their focus was on large-differences between the effects of students with and without disabilities, and what factors contribute to test scores more for students with disabilities compared to students without disabilities.

Li (2014b) was focused on two general questions of the effects of read-aloud accommodations for students with and without disabilities on reading assessments and which factors would influence those effects of read-aloud accommodations. Those factors were explored through moderators (e.g., disability status and content area), and effect sizes were compared utilizing hierarchical linear modeling. Therefore, they focused on using these moderators to explain the variability between student groups and topics. Having broad inclusionary criteria such as including both quasi-experimental and experimental studies as well as published and unpublished studies created a large enough pool, which enabled moderator analysis. The results revealed smaller effect sizes for read-aloud accommodations for math assessments compared to reading assessments regardless of student disability status. Additionally, students with disabilities did not benefit differentially on math assessments compared to students without disabilities.

Buzick and Stone (2014b) carried out a meta-analysis that addressed the effects of read-aloud accommodations for students with and without disabilities on standardized assessments, whether read-aloud accommodations improve test scores more for students with disabilities compared to students without disabilities, and what factors contribute to differences between the effects of students with and without disabilities for read-aloud accommodations. Their inclusionary criteria were stricter than those used by Li (2014a), resulting in a smaller number of studies included in the meta-analysis. The results supported those of Li (2014a) in that smaller effects were found for math assessments compared to reading assessments regardless of student disability status. Additionally, students with disabilities did not benefit differentially on math assessments compared to students without disabilities.

The two previous meta-analyses included students regardless of disability type, and their focus was on large-scale standardized tests that were not specific to reading comprehension. The focus of the current meta-analysis was the effects of text-to-speech and related tools on reading comprehension for individuals with reading disabilities. Additionally, we considered small-n studies, sometimes known as single-case studies, which are commonly used in the field of special education. Recently new methods have become available to calculate a standardized effect for single-case studies, which can be combined with between-subjects studies (Shadish & Lecy, 2015). These methods are now recommended for use as described in a recent paper commissioned by the National Center for Education Research and the National Center for Special Education Research (Shadish, Hedges, Horner, & Odom, 2015).

We hope to extend the findings from the previous meta-analyses by (a) calculating effect sizes for just students with reading disabilities using read-aloud tools on reading comprehension assessments and (b) including additional moderator variables (e.g., delivery method, intervention vs. compensation, amount of reading material presented using the read aloud). Three questions were addressed in this meta-analysis:

1. What is the average weighted effect size of the use of text-to-speech and related tools on reading comprehension for students with reading disabilities?
2. Are there identifiable moderators of the effects of text-to-speech and related tools on reading comprehension?
3. What is the overall quantity and quality of the current research base on students with reading disabilities using text-to-speech software to assist with reading comprehension, and are there important gaps in the literature?

Method

Inclusionary Criteria

Inclusionary criteria were developed to reflect the diversity of technology and settings in which participants used oral presentations of text. The below were used in this meta-analysis:

1. Reading comprehension must be measured at the sentence, paragraph, or passage level. Both researcher-designed and norm-referenced assessments were included.
2. Only studies in which the effect sizes could be calculated for students with dyslexia, reading disabilities, or learning disabilities (subtype reading) were included. Due to the diversity and lack of consensus in the definition of reading disabilities (Quinn & Wagner, 2013), an inclusive definition was used. Students could have more than one disability but had to have a reading disability as well.
3. All studies must include a condition with oral presentation of the reading material. Presentation type may be any of the following: human recorded audio,
human readers, or a variety of technology, including synthesized text-to-speech, reading pens, or other eBooks with text-to-speech or oral capabilities.

4. Studies must be reported in English.

Additionally, in order to gain a complete quantitative synthesis with minimal publication bias, this meta-analysis included all types of literature: peer-reviewed studies, unpublished reports, theses, dissertations, conference proceedings, and poster presentations. Because of variability in the grade at which students with reading disabilities are identified, we restricted our meta-analysis to studies of third grade and above, the grade by which most students with reading disabilities have been identified. The search was not limited by quantitative study design (experimental, quasi-experimental, or observational), utilization of preexisting data, or year; however, qualitative studies were excluded.

Search Procedures

Studies were gathered through a comprehensive and systematic search that utilized multiple methods: searching databases, checking previously published studies’ bibliographies, and contacting experts in the text-to-speech research field. The following databases were used: Education Resources Information Center, Dissertation and Theses, PsycINFO, Linguistic and Language Behavior Abstracts, and ComDisDom provided by Proquest; Education Full Text and Academic Search Complete provided by EBSCO; Web of Science provided by Thomson Reuters; Expanded Academic ASAP provided by GALE; and Google Scholar provided by Google.

Three groups of search terms describing the outcome variable (reading comprehension), the population of interest (students with reading disabilities/dyslexia), and the oral modality of reading material (read aloud/text-to-speech) were used in various combinations, depending on each database’s thesaurus. The topic of “reading comprehension” was searched with the following terms: “reading comprehension,” or “comprehension tests,” or “reading comprehension tests.” Next, various terms referring to dyslexia were connected to the reading comprehension topic using the Boolean AND operator including “response to intervention,” or “learning dis*,” “reading disabilities,” “dyslex*,” or “special education students,” or “disabilities.” These search statements were then combined using the Boolean AND operator with various terms referring to the pairing of audio and visual representations of the text simultaneously. These terms reflect the diversity of text delivery including text-to-speech technology, human readers, and other read-aloud accommodations: “text-to-speech synthesizer,” or “text to speech,” or “text-to-speech,” or “speech synthes*,” or “synthes* speech” or “speech synthesizers,” or “oral reading,” or “read aloud,” or “book on tape,” or “audio books,” or “assistive technology,” or “examinations,” or “academic accommodations,” or “testing accommodations.” Collectively all of the various search combinations yielded 2,933 records, which were reviewed, and finally duplicates were deleted. Initial screening of the search results involved reading titles and abstracts to assess their fit with this review of text-to-speech technology for individuals with reading disabilities. “Reading disabilities” was broadly defined so as not to exclude potentially useful studies. This review of titles and abstracts excluded 2,874 records, leaving 59 records remaining.

Additionally, “backwards mapping” was utilized by identifying 14 other relevant studies from the references list. Of these studies 3 did not have the appropriate population or disability categories (Crawford & Tindal, 2004; Koretz & Hamilton, 2000; Randall & Engelhard, 2010). The 11 remaining studies were retained for further review. Also, experts in the field were contacted for additional studies. This resulted in finding 6 additional studies, of which 3 were already included (Boyle et al., 2003; Hodapp, Judas, Rachow, Munn, & Dimmitt, 2007; Roberts, Takahashi, Park, & Stodden, 2013). The last 2 studies of the 6 identified were not included because 1 was not about reading comprehension (Orr & Parks, 2007) and we were not able to extract data on the treatment groups for the other study (Swan, Kuhn, Groff, & Roca, 2005).

Collectively, all of the search methods yielded 81 articles, which were reviewed in full text. Studies were excluded from the meta-analysis based on the following six exclusionary criteria: qualitative studies, synthesis studies (literature reviews and meta-analyses), multiple interventions without reading comprehension reported separately from other interventions, missing a silent reading condition, not conducted in the students’ native language, and inaccessible full-text versions of studies. More details are available upon request from the first author about methods, including which specific studies were excluded for which reasons. The remaining 43 records were then reviewed for possible inclusion within the meta-analysis.

Coding Procedure

Before coding began, studies for which the effect sizes could not be calculated were removed (see Note 1). In an attempt to get more information about the studies, we contacted the studies’ authors. Next, the following characteristics of the studies were coded: What was read aloud (divided between partial versus entire passages and whether comprehension questions were read aloud), type of oral modality (categorized into one of the three terms: synthesized text-to-speech, human readers, recorded human voice), reading material difficulty (categorized into at the participants’ grade levels or reading levels), intervention (defined as using oral
presentation to improve students’ basic deficits so they can become better readers independent of technology), compensation (defined as utilizing technology as a long-term reading solution, like glasses for fixing vision problems), ability to repeat oral reading material, decoding (defined as students’ decoding skills being needed for any part of the passage or comprehension assessment), participant grade (defined as a student’s grade level represented in the sample [see Note 2]), and type of publication (categorized into either published or unpublished). For between-subjects studies, the following three additional characteristics were coded: randomization (defined as an equal chance of participants’ receiving random treatment or control condition), differential sample attrition, and group equivalence (defined as intervention and control groups’ performing similarly on measured pretest variables). To ensure coding accuracy a second rater coded 20% of the studies. Overall reliabilities were 0.95 for total interrater agreement and 0.94 for overall Kappa. The range of the interrater agreement across the codes was 0.90 to 1.00 (except for decoding at 0.60). Kappa was 1.00 for all codes except decoding (0.28), intervention (0.65), and what was read aloud (0.73), respectively. Coding discrepancies were resolved through discussion except for the decoding moderator, which was coded by a third rater. The third rater was in agreement with the first rater with Kappa = 1.00 and not in agreement with the second rater with Kappa = 0.40. We went with the codes provided by the first and third raters as the second rater appeared to have trouble rating this variable.

Results

Effect Size Calculations

Effect sizes were calculated for the 22 remaining studies (see Note 3). We calculated effect sizes for all studies, except single-case studies, using the equations in Cooper, Hedges, and Valentine (2009). We used the standardized mean difference (Hedges $g$, which is Cohen’s $d$ with a correction, noted in this paper as $\tilde{d}$) as our effect size measure as it corrects for small sample size bias. For the within-subject studies, the paired samples correlations were estimated using other information provided in the publication. Effect sizes that are comparable to those of group studies are available for some case study designs, such as multiple baseline and ABAB designs (Shadish, 2014a, 2014b), and these were included where possible.

Data Analyses

The analyses were conducted using the metafor package (Viechtbauer, 2010) as implemented in R. The package reports restricted maximum likelihood (Harville, 1977) estimates of the pooled effect size and also reports several measures of heterogeneity including $Q$ ($Q$ represents the amount of total heterogeneity in the study and follows a $\chi^2$ distribution with $k−1$ degrees of freedom) (J. P. T. Higgins & Thompson, 2002) and remaining unexplained heterogeneity (for the moderator analysis) (see Note 4), which we reported. We also estimated the effects of potential publication bias using the trim and fill method (Duval & Tweedie, 2004) as well as funnel plots (Cooper et al., 2009).

The results show that the use of text-to-speech tools has a significant impact on reading comprehension scores with $\bar{d} = .35$, 95% confidence interval (CI) (.14, .56), $p < .01$ (see Table 1 and Figure 1, respectively). The sampling variance was significant $Q (22) = 2051.19, p < .001$. Visual confirmation of this heterogeneity using a funnel plot is presented in Figure 2. A trim and fill analysis suggested that the effect size would be $\bar{d} = .24$, 95% CI (.02, .45), $p = .03$. Additionally, a series of sensitivity analyses were conducted.

Our effect size is consistent with the previous meta-analyses (Buzick & Stone, 2014b; Li, 2014b) although with a somewhat smaller average weighted effect size. A possible contributing factor to our smaller average effect size was that we had three within-subject studies with negative effect sizes in our pool that were not in either of the previous meta-analyses (E. L. Higgins & Raskind, 1997; Hodapp et al., 2007; Thurlow, Moen, Lekwa, & Scullin, 2010). Without these studies the results are more positive and remain significant, $\bar{d} = .49 (.34, .58), p < .001$.

In contrast to previous meta-analyses, our study contained postsecondary and adult subjects, but if we only analyzed studies for K–12 we still obtained similar results, $\bar{d} = .36 (.13, .58) p < .01$. Also, the previous meta-analyses included studies measuring reading for students with disabilities published in or after the 2000s. In this meta-analysis, when only the studies published after 2000 were analyzed the average weighted effect size became $\bar{d} = .40 (.17, .64), p < .001$. This pattern remained true for both intervention and compensation studies, $\bar{d} = .54 (.20, .89), p < .001, \bar{d} = .35 (.12, .59), p < .001$, respectively. Overall, the effect of read-aloud tools on reading comprehension found here was significant and positive. However, consistent heterogeneity was found and explored through moderator analyses.

For the moderator analyses, only a single, significant moderator emerged, namely, whether the study design was a between-subjects or within-subject study. The average weighted effect size for between-subjects studies alone was $\bar{d} = .61$, 95% CI (.39, .83), $p < .001$. For within-subject studies, the average weighted effect size dropped to $\bar{d} = .15$, 95% CI (.−.13, .43), $p > .05$. Previous meta-analyses asking what is the impact of presentation accommodations (segmented text and read aloud) on high-stakes tests for
### Table 1. Summary of Included Articles.

<table>
<thead>
<tr>
<th>Citation</th>
<th>$\bar{d}$</th>
<th>$v$</th>
<th>$n$</th>
<th>Average grade</th>
<th>Delivery method</th>
<th>Goal of study</th>
<th>What was read aloud</th>
<th>Reading material difficulty</th>
<th>Decoding needed</th>
<th>Peer reviewed</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Between-subjects study design</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bielinski, Thurlow, Ysseldyke, Freidebach, and Freidebach (2001)</td>
<td>0.28</td>
<td>0.003</td>
<td>1,261</td>
<td>3.0</td>
<td>HR</td>
<td>C</td>
<td>CP and Q/T</td>
<td>GL</td>
<td>No</td>
<td>No</td>
<td>MAP</td>
</tr>
<tr>
<td>Boyle et al. (2003)</td>
<td>0.65</td>
<td>0.108</td>
<td>38</td>
<td>10.5</td>
<td>HR and RHV</td>
<td>C</td>
<td>CP and Q/T</td>
<td>GL</td>
<td>No</td>
<td>Yes</td>
<td>RD</td>
</tr>
<tr>
<td>Fasting and Lyster (2005)</td>
<td>0.92</td>
<td>0.083</td>
<td>52</td>
<td>6.0</td>
<td>TTS</td>
<td>I and C</td>
<td>CP</td>
<td>GL</td>
<td>Yes</td>
<td>Yes</td>
<td>LS60</td>
</tr>
<tr>
<td>Fletcher, Francis, Boudousquie, and Copeland (2006)</td>
<td>0.90</td>
<td>0.048</td>
<td>91</td>
<td>3.0</td>
<td>HR</td>
<td>C</td>
<td>PP and Q/T</td>
<td>GL</td>
<td>Yes</td>
<td>Yes</td>
<td>TAKS</td>
</tr>
<tr>
<td>Fletcher et al. (2009)</td>
<td>0.50</td>
<td>0.065</td>
<td>62</td>
<td>7.0</td>
<td>HR</td>
<td>C</td>
<td>PP and Q/T</td>
<td>GL</td>
<td>No</td>
<td>Yes</td>
<td>NARA$^a$</td>
</tr>
<tr>
<td>Lange, McPhills, Mulhern, and Wylie (2006)</td>
<td>0.77</td>
<td>0.083</td>
<td>54</td>
<td>9.5</td>
<td>TTS</td>
<td>C</td>
<td>CP and Q/T</td>
<td>GL</td>
<td>No</td>
<td>Yes</td>
<td>TKS</td>
</tr>
<tr>
<td>Lundberg and Olofsson (1993)</td>
<td>−0.32</td>
<td>0.241</td>
<td>15</td>
<td>5.6</td>
<td>TTS</td>
<td>I</td>
<td>PP</td>
<td>RL</td>
<td>Yes</td>
<td>Yes</td>
<td>RD</td>
</tr>
<tr>
<td>Meloy, Deville, and Frisbie (2002)</td>
<td>1.10</td>
<td>0.073</td>
<td>62</td>
<td>7.0</td>
<td>HR</td>
<td>C</td>
<td>CP and Q/T, RA</td>
<td>GL</td>
<td>No</td>
<td>Yes</td>
<td>ITBS</td>
</tr>
<tr>
<td>Fletcher et al. (2009)</td>
<td>0.45</td>
<td>0.039</td>
<td>164</td>
<td>9.0</td>
<td>TTS</td>
<td>C</td>
<td>CP</td>
<td>GL</td>
<td>Yes</td>
<td>No</td>
<td>GMRT</td>
</tr>
<tr>
<td>Shany and Biemiller (1995)</td>
<td>0.81</td>
<td>0.209</td>
<td>18</td>
<td>3.5</td>
<td>HR</td>
<td>I</td>
<td>CP</td>
<td>RL</td>
<td>Yes</td>
<td>Yes</td>
<td>SAT$^a$</td>
</tr>
<tr>
<td><strong>Within-subject study design</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dolan, Hall, Banerjee, Chun, and Strangman (2005)</td>
<td>1.02</td>
<td>0.575</td>
<td>9</td>
<td>11.5</td>
<td>TTS</td>
<td>C</td>
<td>CP and Q/T</td>
<td>GL</td>
<td>No</td>
<td>Yes</td>
<td>NAEP</td>
</tr>
<tr>
<td>Elkind, Black, and Murray (1996)</td>
<td>0.10</td>
<td>0.139</td>
<td>50</td>
<td>17.0</td>
<td>TTS</td>
<td>C</td>
<td>CP and Q/T</td>
<td>GL</td>
<td>No</td>
<td>Yes</td>
<td>NDRT</td>
</tr>
<tr>
<td>Floyd and Judge (2012)</td>
<td>1.09</td>
<td>0.102</td>
<td>6</td>
<td>12.0</td>
<td>TTS</td>
<td>C</td>
<td>CP and Q/T</td>
<td>GL</td>
<td>No</td>
<td>Yes</td>
<td>SAT$^a$</td>
</tr>
<tr>
<td>E. L. Higgins and Raskind (1997)</td>
<td>−0.12</td>
<td>0.019</td>
<td>37</td>
<td>13.0</td>
<td>TTS</td>
<td>C</td>
<td>CP and Q/T, RA</td>
<td>GL</td>
<td>No</td>
<td>Yes</td>
<td>FRI</td>
</tr>
<tr>
<td>Hodapp, Judas, Rachow, Munn, and Dimmitt (2007)</td>
<td>−0.74</td>
<td>0.034</td>
<td>27</td>
<td>7.5</td>
<td>TTS</td>
<td>C</td>
<td>CP and Q/T</td>
<td>GL</td>
<td>No</td>
<td>No</td>
<td>JRS</td>
</tr>
<tr>
<td>Kosiolek and Ysseldyke (2000)</td>
<td>0.53</td>
<td>0.073</td>
<td>14</td>
<td>4.0</td>
<td>RHV</td>
<td>C</td>
<td>CP and Q/T</td>
<td>GL</td>
<td>No</td>
<td>No</td>
<td>CAT/5</td>
</tr>
<tr>
<td>Laitusis (2010a), Sample 1</td>
<td>0.57</td>
<td>0.000</td>
<td>527</td>
<td>4.0</td>
<td>RHV</td>
<td>C</td>
<td>CP and Q/T, RA</td>
<td>GL</td>
<td>No</td>
<td>Yes</td>
<td>GMRT</td>
</tr>
<tr>
<td>Laitusis (2010b), Sample 2</td>
<td>0.32</td>
<td>0.000</td>
<td>376</td>
<td>8.0</td>
<td>RHV</td>
<td>C</td>
<td>CP and Q/T, RA</td>
<td>GL</td>
<td>No</td>
<td>Yes</td>
<td>GMRT</td>
</tr>
<tr>
<td>McCoy (2012)</td>
<td>0.53</td>
<td>0.337</td>
<td>3</td>
<td>10.0</td>
<td>TTS</td>
<td>I and C</td>
<td>CP</td>
<td>GL</td>
<td>Yes</td>
<td>No</td>
<td>STAR</td>
</tr>
<tr>
<td>Meyer and Bouck (2014)</td>
<td>−0.21</td>
<td>0.143</td>
<td>3</td>
<td>7.6</td>
<td>TTS</td>
<td>C</td>
<td>CP and Q/T</td>
<td>GL</td>
<td>No</td>
<td>Yes</td>
<td>SWP</td>
</tr>
<tr>
<td>Schmitt, Hale, McCallum, and Mauck (2011)</td>
<td>0.23</td>
<td>0.063</td>
<td>25</td>
<td>7.0</td>
<td>TTS</td>
<td>C</td>
<td>CP</td>
<td>GL</td>
<td>Yes</td>
<td>Yes</td>
<td>TRL</td>
</tr>
<tr>
<td>Sorrell, Bell, and McCallum (2007)</td>
<td>−0.16</td>
<td>0.078</td>
<td>4</td>
<td>4.0</td>
<td>TTS</td>
<td>C</td>
<td>CP</td>
<td>RL</td>
<td>Yes</td>
<td>Yes</td>
<td>AR</td>
</tr>
<tr>
<td>Thurlow, Moen, Lekak, and Scullin (2010)</td>
<td>−0.32</td>
<td>0.011</td>
<td>44</td>
<td>6.7</td>
<td>RP</td>
<td>C</td>
<td>PP and Q/T</td>
<td>GL</td>
<td>Yes</td>
<td>No</td>
<td>GSRT</td>
</tr>
</tbody>
</table>

Note. $\bar{d}$ = effect size; $v$ = variance; $n$ = number of subjects; HR = human reader; C = compensation; CP = complete passages; Q/T = questions/test(s); GL = grade level; MAP = Missouri Assessment Program; RHV = recorded human voice; RD = researcher developed; TTS = text-to-speech; I = intervention; LS60 = Test of Silent Sentence Reading; PP = partial passages; TKS = Texas Assessment of Knowledge and Skills; NARA = Modified Neale Analysis of Reading Ability II; RL = student’s reading level; ITBS = Iowa Test of Basic Skills the reading comprehension part; RA = explicitly stated that they could request oral reading material on questions, passages, or comprehension tests; GMRT = Gates-MacGinitie Reading Tests; SAT = SAT Critical Readings; NAEP = NCES U.S. History and Civics tests; NDRT = Nelson-Denny Reading Test; FRI = Formal Reading Inventory; JRS = Jamestown Reading Series; CAT/S = California Achievement Test 5th Edition comprehension part; STAR = STAR Reading Assessment Test; SWP = Six Way Paragraphs Middle Level; TRL = Timed Readings in Literature; AR = Accelerated Reader; RP = reading pen; GSRT = Gray Silent Reading Test.

$^a$Researcher modified.
students with disabilities also found a lower effect size for within-subject study effect sizes as compared to between-subjects studies. Li (2014b) also found this pattern, but it did not remain significant after all of the predictors were added into the model.

Discussion

The impact of text-to-speech and related read-aloud tools for text presentation on reading comprehension for individuals with reading disabilities was explored in this meta-analysis. Three questions were explored and will be discussed in turn.

Average Effect Size of the Use of Text-to-Speech on Measures of Reading Comprehension

Our meta-analysis provides a more focused approach than previously taken as we looked exclusively at students with reading disabilities using read-aloud tools on reading comprehension assessments. This is motivated by a previous meta-analysis finding that indicated that presentation accommodations (segmented reading and read aloud) had a significant positive effect for students with learning disabilities but did not have an effect for students requiring special education services (Vanchu-Orosco, 2012). Both meta-analyses published in 2014 included many types of disabilities without distinguishing among them, whereas in the current study we focus exclusively on students with reading disabilities. Another difference is that the previous meta-analysis (Buzick & Stone, 2014b) focused mostly on state assessments and other large standardized assessments. Many large-scale assessments test general English skills, not solely reading comprehension. Here, we specifically chose to examine reading comprehension tests. Through this narrow focus on students with reading disabilities and reading comprehension assessments, our model hopes to expand the previous studies to provide new information about moderators affecting the effectiveness of read-aloud tools on reading comprehension.

The effects of text-to-speech and related read-aloud tools indicate that oral presentation of text for students with disabilities helps their reading comprehension test scores. Our finding of an average weighted effect size of .35 is consistent with the recent meta-analysis results on read-aloud accommodations for students with disabilities (Buzick & Stone, 2014b; Li, 2014b). Taken together, this
meta-analysis is useful as a starting point to begin quantifying the wide range of published and unpublished read-aloud literature.

Amount of Variability Present and Identifiable Moderators in Effect Sizes

There was significant heterogeneity present. The moderator of between-subjects versus within-subject study design was found to be significant. Possible reasons for this may include regression to the mean, attrition, and order effects due to lack of counterbalancing treatments (Li, 2014b). Interestingly, none of the other moderators were significant in the random effects model. Potential carry-over or training effects of reading with assistive technology may contribute to these differences. Nonsignificant moderators included what was read aloud, type of oral modality, reading material difficulty, intervention, compensation, ability to repeat oral reading material, decoding needed, and grade level. We were surprised that grade level did not moderate the effect of the text-to-speech/read-aloud presentation. This is in contrast to the previous meta-analyses, which found grade level a significant moderator (Buzick & Stone, 2014b). However, depending on how the variation in effect sizes between studies is conceptualized, it affects the amount of variance accounted for by the grade-level moderator (see Table 2; see Note 5).

In addition, we were surprised that whether the study used text-to-speech/read-aloud tools in an intervention setting did not emerge as a significant moderator. Using the text-to-speech/read-aloud tools as intervention is theoretically and practically different than using text-to-speech/read-aloud tools in a compensation setting. There were very few studies that were truly interventions despite many studies claiming to be interventions. More research comparing intervention versus compensation approaches needs to be conducted.

The Quantity and Quality of Research on Students Using Text-to-Speech and Related Read-Aloud Tools for Reading Comprehension

Initial search results suggest that this is an active and growing area of research. We found 81 articles for full-text review while applying exclusionary criteria. We found a small number of studies from the 1980s and 1990s. The literature is growing, and the rate of new studies being conducted is increasing with time. Interestingly, the amount of studies using the computerized text-to-speech read-aloud tool is increasing. For example, the studies from the 2010s all used text-to-speech technology. This may reflect the trend of wider access to improved text-to-speech technology.

Limitations and Future Directions

However, study quality remains a concern, with many studies missing key components such as proper controls and adequate statistical reporting. Additionally, many studies used convenience samples (e.g., Boyle et al., 2003). It is important to have randomized treatment-controlled studies to make causal inferences about the effect of text-to-speech/read-aloud tools in an educational context, thus increasing ecological validity by capturing the diversity of the current classroom. However, there are trade-offs between internal and external validity that are difficult to balance. It is ideal to have large sample sizes, but they should also be as homogeneous as possible, allowing the best possible assessment of the read-aloud/text-to-speech presentation. However, small-sample studies could be useful for identifying students’ disabilities and then allowing intensive study of these individual differences. The downside to them is external validity because you are not able to have a large enough sample size to extend to various types of learning disabilities. However, the studies’ high internal validity is beneficial to this field. Taken together, new studies with clear and complete reporting of both between-subjects and within-subject study designs will be beneficial.

Table 2. Moderator Analysis.

<table>
<thead>
<tr>
<th>Moderator</th>
<th>K</th>
<th>N</th>
<th>B</th>
<th>QM</th>
<th>QE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between-subjects</td>
<td>23</td>
<td>2,942</td>
<td>.48</td>
<td>6.17</td>
<td>2049.60**</td>
</tr>
<tr>
<td>Grade level</td>
<td>23</td>
<td>2,942</td>
<td>&lt;.01</td>
<td>0.01</td>
<td>168.13**</td>
</tr>
<tr>
<td>Intervention</td>
<td>23</td>
<td>2,942</td>
<td>.20</td>
<td>0.47</td>
<td>2049.02**</td>
</tr>
<tr>
<td>Difficulty</td>
<td>23</td>
<td>2,942</td>
<td>.30</td>
<td>0.67</td>
<td>2049.14</td>
</tr>
<tr>
<td>Delivery method (2 modes below)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Recorded voice</td>
<td></td>
<td></td>
<td></td>
<td>.29</td>
<td></td>
</tr>
<tr>
<td>b. Human reader</td>
<td></td>
<td></td>
<td></td>
<td>.44</td>
<td></td>
</tr>
<tr>
<td>Decoding needed</td>
<td>23</td>
<td>2,942</td>
<td>-.02</td>
<td>0.01</td>
<td>2041.46**</td>
</tr>
<tr>
<td>Partly oral passages</td>
<td>23</td>
<td>2,942</td>
<td>-.16</td>
<td>0.32</td>
<td>2032.51**</td>
</tr>
<tr>
<td>Ability to repeat</td>
<td>23</td>
<td>2,942</td>
<td>.11</td>
<td>0.16</td>
<td>2040.92**</td>
</tr>
<tr>
<td>oral reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peer reviewed</td>
<td>23</td>
<td>2,942</td>
<td>.41</td>
<td>3.49</td>
<td>2028.56</td>
</tr>
<tr>
<td>Randomized</td>
<td>23</td>
<td>2,942</td>
<td>.28</td>
<td>1.34</td>
<td>2046.64</td>
</tr>
<tr>
<td>Baseline</td>
<td>3</td>
<td>1,337</td>
<td>.59</td>
<td>1.86</td>
<td>1.46</td>
</tr>
</tbody>
</table>

Note. K = number of studies included; B = change in effect size of one-unit increase in moderator; QM = heterogeneity explained by the moderator; QE = remaining unexplained heterogeneity.

*p < .05. **p < .01.

There are a few limitations that are worth noting. First, our study included a relatively small sample size (n = 22 studies); however, this is still larger than previous findings. Second, reading comprehension measures and text-to-speech interventions were diverse and were applied at
different intensities. Analyzing the importance of intensity of interventions seems particularly interesting, but due to inconsistent reporting, we were unable to develop a meaningful measure of intensity. Third, placebo effects and motivation have yet to be explored or explicitly controlled for. Currently, there is a lack of randomized control trials; these studies will help provide evidence as to whether and how assistive technology affects students’ reading comprehension test scores.

In future studies, measurements of participants’ reading skills such as decoding, working memory, vocabulary, and listening and reading comprehension should be assessed. Text-to-speech technology features including speech speed, text highlighting, and intensity need to be explicitly addressed. With the advent of systemwide user tracking (e.g., length of use, user preferences) of text-to-speech programs, such as Read & Write Gold and Kurzweil 3000, users’ actions can be recorded. This data can be combined with student characteristics like reading level and disability status as well as their reading test outcomes to see how students should be utilizing the software. These data could then be compared to the students’ reading comprehension measures with text-to-speech/read-aloud tools in compensatory settings such as a statewide reading comprehension test. This enables large quantities of data to be collected, but the quality of the data this monitoring provides is questionable. In summary, future research should investigate the following: features of text-to-speech tools, participant reading characteristics and disabilities, study design considerations, and larger scale usage studies.

Conclusion

Text-to-speech/read-aloud presentation positively affects reading comprehension for individuals with reading disabilities, with average weighted effect sizes of $\bar{d} = .35$ ($p < .001$). There is more variability than would be expected due to random error alone. Significant variance was accounted for by design type (e.g., within-subject studies versus between-subjects studies). The overall quantity and quality of studies investigating the effectiveness of text-to-speech and related read-aloud tools on reading comprehension is increasing. However, more studies are needed to explore diverse types of text presentation, for example, different software packages and displays as well as reading disability profiles. These future research studies can then be used to explore how students are interacting with these technologies and their impact on reading comprehension. Taken together, our results indicate that text-to-speech/read-aloud presentation may help students, although the mechanisms remain unknown. Better study designs are required to control for these potential effects.

Authors’ Note

Thank you to the following experts in assistive technology for responding to our request for additional data and/or providing access to unpublished data: Cynthia M. Okolo, Sara M. Flanagan, Emily Bouch, Maria Earman Stetter, Marie Tejero Hughes, Richard E. Ferdig, Tracey Hall, E. A. Draffan, Terry Thompson, Margaret Bausch, Ted S. Hasselbring, Jane Hauser, Melanie Kuhn, Dave Edyburn, and Chuck Hitchcock. Elizabeth L. Tighe is now at Georgia State University, Atlanta, USA.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Support for this project was provided by grant number P50 HD052120 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development.

Notes

1. Studies were deleted for the following reasons: (a) studies had fewer than 3 participants ($n = 2$); (b) studies had poor matching, defined as having a group score at least 2 standard deviations better in pretest ($n = 1$); (c) studies had non-independent samples across studies (e.g., Laitusis, 2010; $n = 3$); (d) studies had not enough information to estimate correlations for the effect sizes ($n = 8$); and (e) studies had single-case study designs of alternating treatments and variations on that design as there is not a standardized effect size available to combine them with between-subjects studies ($n = 5$).

2. Studies’ reports of grade level were inconsistent, and we had to code it according to a few assumptions. Initially, if mean grade was reported, that was used, but if instead a range of grades was reported, the midpoint was used. If an age mean was reported, we used the grade that is typically equivalent to that age in the United States. If an age range was reported, we used the grade that is typically equivalent to the midpoint of the age. If the students were reported to be college students, grade was coded as 13.

3. (a) Raw data that were provided by researchers were used to calculate the effect size and the variance ($n = 1$); (b) correlations estimated from data presented in the publication in either graphical or written form ($n = 13$); (c) instead of using the difference score the posttest score was used to calculate the effect size as the correlation was not available ($n = 2$); (d) effect size was calculated directly from the information provided ($n = 7$); (e) variances were estimated by the published effect size ($n = 5$); (f) average of two measures of reading comprehension ($n = 3$). Citations of these articles are available upon request.

4. Meta-regressions with each of the moderators variable in the between-subjects studies: average grade level, intervention, difficulty, delivery method, decoding needed, partly or fully
oral passages, peer reviewed randomized and baseline (coded only for studies that were not randomized). Similarly, within-subject moderator analysis was run for moderators: average grade level, intervention, difficulty, delivery method, decoding needed, partly or fully oral passages, peer reviewed.

5. Heterogeneity is composed of both random error and true heterogeneity among the studies. When heterogeneity is assumed to be only due to random error, as done in fixed-effect models, under this assumption grade level accounts for a significant amount of variance. Unlike fixed-effect models, random effects models do not assume the same common effect (here the effect of read-aloud tools on reading comprehension), and the heterogeneity present is from the heterogeneity among studies and random error. Under the assumptions of the random-effects model the effect of grade level is no longer seen. The calculation of $Q$ is based on a fixed-effects model. If we had assumed the fixed-effects model was correct, grade would have accounted for significant heterogeneity: heterogeneity explained by the moderator ($QM(1) = 1883.06, p < .01$ with a parameter estimate of $B = -.06, 95\% CI (-.064, -.058)$. The restricted maximum likelihood model calculates $Q$ based on the true effects accounting for heterogeneity (Viechtbauer, 2010). As shown in Table 2 the effect of grade is nearly zero under these less biased assumptions.

References

References marked with an asterisk indicate studies included in the meta-analysis.


