This is the accepted manuscript (post-print version) of the article. The content in the accepted manuscript version is identical to the final published version, although typography and layout may differ.

How to cite this publication
Please cite the final published version:


Publication metadata

**Title:** A Cluster-Randomized Trial Measuring the Effects of a Digital Learning Tool Supporting Decoding and Reading for Meaning in Grade 2.

**Author(s):** Gissel, Stig Toke; Andersen, Simon Calmar.

**Journal:** Journal of Computer Assisted Learning

**DOI/Link:** 10.1111/jcal.12488

**Document version:** Accepted manuscript (post-print)
A Cluster-Randomized Trial Measuring the Effects of a Digital Learning Tool
Supporting Decoding and Reading for Meaning in Grade 2

Abstract
A large-scale, cluster-randomized controlled field trial (Nclassrooms=47; Nstudents=1,013) assessed the impact of a digital text-to-speech reading material that supported eight-year-olds’ decoding and reading comprehension. An active control group used the most prevalent Danish learning material with a research based systematic, explicit phonics approach supporting primarily decoding. The digital tool allows children to read unfamiliar text for meaning. Students are supported in mapping between orthography and phonology by three levels of text-to-speech support and in identifying spelling patterns. The risk of students overusing text-to-speech was countered by postponing access to having words read aloud by directing students towards identifying and training relevant orthographic patterns before activating text-to-speech. Results showed no statistically significant difference in decoding, but treatment improved reading comprehension. The study demonstrates how digital tools can facilitate strengthening students’ decoding skills as efficiently as a traditional phonics-based program whilst students are reading text of relatively high orthographic complexity for meaning.

Keywords: Digital learning tool; reading instruction; text-to-speech; connectionism; randomized controlled trial.

To become good readers, children must develop two general strands of skills: 1) decoding of letters into words, and 2) comprehension of meaning in words, sentences, and texts (Gough & Tunmer, 1986; Zucker, Moody, & McKenna, 2009). The National Reading Panel (NRP) report (2000) found that systematic phonics instruction is the most efficient method of reading instruction for young students. As students become older, progress in comprehension skills may be accelerated if students are exposed to unfamiliar texts with more irregular words than are found in traditional basal readers. Yet, irregular words, which students cannot decode, may block their reading. Digital learning tools that read aloud words that students cannot read independently may be a solution to this dilemma. However, students tend to overuse such text-to-speech technologies, which may impede their
orthographic learning (Van Daal & Reitsma, 1993; Lewin, 2000; Lefever-Davis & Pearman, 2005; Gissel, 2014). An important question therefore is whether digital learning tools can be designed in a way that supports the reading of unfamiliar text while restricting the use of text-to-speech support.

This study reports the results of an effectiveness trial that tested the effects of using a digital tool designed and developed to handle the reading instruction dilemma (Gissel, 2015). The tool supports 2nd grade students in comprehending unfamiliar text independently, embeds code-focused training to create or strengthen orthographic patterns and connections between spellings and pronunciations, and counters student overuse of text-to-speech support to focus student attention on processing the orthographic representation of words. The learning tool was tested against an active control group using traditional, paper-based phonics instruction in order to assess the effectiveness of this type of computer-assisted learning under real classroom conditions.

A key component in children’s reading development is orthographic learning i.e., knowledge about orthographic regularities in the writing system and word-specific knowledge facilitating access of word meaning from print (Castles, Rastle, & Nation, 2018). Orthographic learning is fostered when students perform alphabetic decoding and are exposed to different printed words (Castles, Rastle, & Nation, 2018). As children decode and are exposed to words, they build lexical quality i.e., stored mental representations of words’ form and meaning (Perfetti, 1992, 2007; Perfetti & Hart, 2002).

Cognitive, psychological science has yet to understand reading theoretically and how reading is realized in the brain (Seidenberg, 2012). One of the methods for understanding, which cognitive components make up the reading process and the interactions of these components, is computer simulations of reading processes. In developing the most convincing simulation based model of word recognition, dual-mechanism models (see, for example, Tracy & Morrow, 2006) compete with connectionist models. Dual-mechanism theories assume for the reader to learn correspondences between spelling and sound, two types of knowledge are involved: Regular words are handled with the rules of language and exception words have to be memorized (Coltheart et al., 2001). Hence, following the insights from dual-mechanism models, instructional materials could use texts with highly controlled vocabulary and emphasize direct instruction in phonics rules and rote learning of exception words.

Connectionist models, on the other hand, suggest that the same mechanism is used for processing words regardless of the degree of consistency between spelling and sound. Connectionist models of word reading (Seidenberg & McClelland, 1989; Plaut & McClelland, 1993; Seidenberg,
SUPPORTING DECODING AND COMPREHENSION

2005, 2007) suggest that orthographic learning builds rather inductively through the reader’s exposure to words and text. When a spelling pattern is encountered, the reader processes it by using a series of possible pronunciations built from previous exposure to the spelling pattern. With exposure to print, readers create and strengthen a finite set of units representing the spellings and pronunciations of words. These representations are distributed, i.e., they represent a large set of spelling and sound patterns. All units are connected and have a weight depending on the frequency of previous exposure to the unit.

With orthographic learning and increased lexical quality, children become less reliant on alphabetic decoding and increasingly able to access meaning directly from orthographic processing (Harm & Seidenberg, 2004). Automated decoding is one of the key prerequisites for reading comprehension. If decoding is not automated, cognitive capacity is drained and readers will be less able to focus on comprehending the text (LaBerge & Samuels, 1974; Perfetti, 1985).

Yet, substantial variance has been found in the correlation between reading comprehension and decoding across studies (García & Cain, 2014; Paris & Hamilton, 2009). Bus and van IJzendoorn (1999) and Whitehurst and Lonigan (1998) in their reviews concluded that basic skills (e.g. decoding) and higher-order skills (e.g. comprehension) develop simultaneously and independently rather than sequentially. García and Cain’s (2014) meta-analysis found that the correlation between decoding and reading comprehension reduced dramatically from the age of seven to the age of 16 or older. These results suggest that the focus on decoding, at some point, should be supplemented with an effort to strengthen higher-order comprehension for all students including those struggling with basic skills (Rapp, van den Broek, McMaster, Kendeou, & Espin, 2007).

Connectionist models describe the complex relationship between basic skills and higher order comprehension. In the connectionist models of the full reading process, four processors work together simultaneously: The context processor, the meaning processor, the orthographic processor, and the phonological processor (Seidenberg & McClelland, 1989; Seidenberg, 2012; Adams, 1990). The meaning processor is connected to the three other processors, sending and receiving feedback from both the meaning processor, the orthographic processor, and the phonological processor. When reading printed text, readers use the orthographic processor to process the visual input of the string of letters. Also, and always the phonological processor is immediately activated, and if the word string is pronounceable, feedback is sent to the orthographic processor. The meaning processor can aid the orthographic processor by giving and receiving feedback to and from the
context processor, aiding the recognition of the most likely word. The context processor assembles the meaning of sentences, paragraphs, and text into a coherent message, if possible. In addition, the context processor sends information to the meaning processor forming hypothesis about what the next words are likely to be.

Perfetti and Stafura (2014) expand our understanding of the crucial factors that interact and facilitate students’ reading comprehension: Student knowledge (linguistic, orthographic and general knowledge), the smooth functioning of the aforementioned reading processes (decoding, word identification, meaning making) along with inferencing and comprehension monitoring and general cognitive resources (for example memory). None of these factors can be ignored when predicting reading comprehensions skills (Castles, Rastle, & Nation, 2018).

Word recognition implies that students know the meaning of the target word. There is firm evidence that vocabulary knowledge for understanding most words in the text is crucial for student comprehension (Spencer, Quinn, & Wagner, 2017). Most studies that investigate the effects of direct teaching of vocabulary on reading comprehension are successful in strengthening comprehension of texts containing the target words (Elleman et al., 2009; Wright & Cervetti, 2016). However, these studies typically find minimal effects on generalized comprehension, i.e., when students have to transfer their vocabulary training to texts that do not contain the words taught. In their review, Wright and Cervetti (2016) found that programs involving active comprehension processing were relatively more efficient than alternative methods. These findings suggest, that teaching relevant words to students before reading can help them comprehend the texts and that instructional materials should facilitate that students work actively with making meaning of text.

1.1. Balancing Decoding and Comprehension in Instructional Resources

The NRP report (2000) found that systematic phonics instruction is the most efficient method of reading instruction when compared to whole language programmes and other programmes that do not teach phonics explicitly and systematically. This finding was an upset to the whole language movement that dominated the field of reading instruction and research in the 1990s (Pearson, 2004). Whole language, in contrast to the NRP recommendations, emphasizes incidental phonics instruction based on the teachers’ judgment of individual students’ needs (NRP, 2000; Pearson, 2004) or in some cases disavowal of the value of phonics instruction (Adams, 1991).

As a direct consequence of the NRP report, the Reading First program was implemented in the US as part of the No Child Left Behind Act of 2001. The Reading First program promotes
SUPPORTING DECODING AND COMPREHENSION

instructional practices (e.g. reading materials) validated by scientific research (Gamse, Jacob, Horst, Boulay, & Unlu, 2008). The Reading First impact study (Gamse et al., 2008) found that instructional time spent on the essential components of reading instruction increased significantly in grades one and two, and that the practices in schools and classrooms changed significantly in line with the program. However, the program impact study showed small effect on decoding measures among first-grade students (second grade was not measured) while there was no significant impact on grade 1-3 students’ reading comprehension (Gamse et al., 2008).

The best available evidence suggests that phonics instruction should begin in kindergarten or in first grade to maximise impact on decoding and comprehension (NRP, 2000; see also Brooks, Torgerson, & Hall, 2008). However, we lack knowledge about the role of reading instruction for students at the age of 7-8 years, when the balance between a focus on decoding and a focus on reading for meaning may shift (Connor, Morrison, & Underwood, 2007).

The challenge of finding ways to teach decoding efficiently whilst students are reading for meaning is not new (Gough, 1981). Traditional phonics training targets decoding of regular words with reading materials containing highly controlled vocabulary and little emphasis on reading for meaning. However, research is inconclusive in determining whether phonics instruction should be extended beyond basic mappings i.e., to include instruction in context-sensitive rules, or if these mappings can be learned through text experience once the student acquires the basic grapheme-phoneme correspondences (Castles, Rastle, & Nation, 2018). Stuart, Masterson, Dixon, and Quinlan (1999) reported that 6-7 year old children could acquire knowledge of sublexical correspondences between print and sound from reading experience. In a comparison between two synthetic phonics programmes, which differed in the extent that multiple letter-sound mappings were taught, no difference was found (Shapiro & Solity, 2016).

Decodable texts are widely used in reading instruction. They ease decoding by containing regular words that reflect the letter-sound correspondences that students have been taught along with repetitions of highly frequent exception words. Drawbacks of decodable texts are that they might be uninteresting for students and that they do not expose students to enough variation in vocabulary (Castles, Rastle, & Nation, 2018).

A wide variety of instructional measures for promoting reading comprehension have been tested for efficiency. The NRP report (2000) and other meta-analysis report large positive effects of explicit teaching of reading strategies (Sencibaugh, 2007; Rosenshine & Meister, 1994; Swanson, 1999). For example, reciprocal teaching (Palinscar & Brown, 1984) where students discuss texts
with each other or in dialogue with the teacher using comprehension strategies such as summarising, predicting, clarifying, and questioning, show large effects on reading comprehension among students that are comprehension disabled (Moore, 2005). However, Okkinga, van Steensel, van Gelderen, van Schooten, Sleegers, and Arends (2018) found that reading strategy interventions often were conducted in controlled settings with small groups of students. When Okkinga et al. summarized the effectiveness of studies performed in whole-classroom settings, very small effects on measures of reading comprehension were found.

Progress in comprehension skills may be accelerated if students learn to read unfamiliar texts that, in comparison to decodable texts, will contain more irregular words, a more varied vocabulary and which lie closer to students’ listening levels (Willingham, 2017; Castles, Rastle, & Nation, 2018). By unfamiliar text, we refer to a text that the student reads independently for meaning without having it read aloud by the teacher beforehand. It is a very different task than reading familiar texts. When students read a familiar text, it allows them to focus on orthographic decoding, because they do not need to focus on comprehending the text.

However, reading familiar text may inadvertently cause students to adopt a reading strategy that relies on context processes at the expense of orthographic and phonological processes (Stanovich, 1980). Interactive models and connectionist models alike assume that word reading ideally is a synthesis of information provided simultaneously from relatively independent processes, that is orthographic, phonological, semantic, and context processes (Stanovich, 1980; Seidenberg & McClelland, 1989; Adams, 1990; Plaut & McClelland, 1993; Seidenberg, 2005, 2007). For the crucial connections between letters and words to be strengthened, the reader must on the one hand implicitly activate alphabetic knowledge in order to analyze all the orthographic features in each word (Adams, 1990; Ehri, 2007). On the other hand, context processing should facilitate comprehension, not word recognition. For students to practice reading comprehension and self-regulate their reading process, that is, to use contextual processing as a constraint vis-à-vis word recognition, they need to work with unfamiliar texts. In addition, strengthening low-level, data-driven word recognition is essential for freeing up capacity for comprehension processes (Stanovich, 1980).

Nonetheless, reading unfamiliar texts may damage comprehension and leave some students frustrated if too many words are too difficult or not part of their vocabulary (Spencer, Quinn, & Wagner, 2017; Betts, 1946). Irregular words, which students cannot decode, may block their reading. This might be a particular challenge in languages such as Danish and English that contain
SUPPORTING DECODING AND COMPREHENSION

many irregular words (Elbro, 2006). Digital learning tools that read aloud words that students cannot decode are a possible solution. However, students tend to overuse text-to-speech technologies and related read-aloud tools (henceforth referred to as TTS), which may impede their orthographic learning (Van Daal & Reitsma, 1993; Lewin, 2000; Lefever-Davis & Pearman, 2005; Gissel, 2014). Consequently, TTS can potentially support the balance between decoding and comprehension-focused instruction if student overuse of TTS support is countered. The design tested in this study attempts to reap the benefits of TTS whilst minimizing overuse.

1.2. Using Digital Learning Tools to Balance Focus on Decoding and Comprehension

TTS technology and related read-aloud tools are technical solutions to the problem of letting students read unfamiliar text, while avoiding the frustration of being unable to read irregular words that are not easily decoded. Read-aloud tools allows readers to click on words they cannot decode and have them read aloud. Thereby, TTS supports beginning readers in reading unfamiliar texts with a difficulty level approximating their listening level, which, in turn, facilitates reading for meaning and thus comprehension. Furthermore, TTS helps readers identify the meaning of words in those instances, where they know the meaning of the word but find it difficult to decode.

According to connectionist models, correspondences between phonology and sub-word letter patterns can be acquired from feedback on whole-word level (Seidenberg & McClelland, 1989), which echoes the findings of Bowey and Muller (2005) and Cunningham, Perry, Stanovich, and Share (2002). Share (1995, 1999, 2004) showed that students can gain lasting word-specific orthographic representations from phonological recoding of new words from one or few exposures to the word. However, if students listen instead of read, the lack of exposure to the orthographic material could reduce any effect on decoding skills (Van Daal & Reitsma, 1993; Lewin, 2000; Lefever-Davis & Pearman, 2005; Gissel, 2014). Hence, to reap the potential benefits of using TTS in reading instruction we need to counter students’ overuse of TTS and direct their attention towards the orthographic representations.

Research on the effect of TTS on decoding and word recognition has largely been dormant for the past decade. Most research on the use of TTS for strengthening decoding skills focused on remedial reading (Karemaker, Pitchford, & O'Malley, 2010a; Olofsson, 1992; Olson & Wise, 1992; Zucker, Moody, & McKenna, 2009). It would seem that more capable readers gain less than poorer readers from using TTS (Balajthy, 2005). However, Karemaker, Pitchford, and O'Malley (2010b) found greater gains in normally developing 5–6-year-old readers’ word recognition and word
SUPPORTING DECODING AND COMPREHENSION

naming from using whole-word multimedia software that reads a story aloud and highlights each word compared to typical whole-class teaching methods. Reitsma (1988) compared three conditions, i.e., guided reading with sustaining feedback, independent reading with whole-word TTS and reading-while-listening. It was found that both guided reading and TTS reading were significantly more effective than both the control condition and reading-while-listening with first-grade students. Surprisingly, the guided reading and independent reading with whole-word TTS were equally effective.

Previous intervention studies of relevance to the present study include Horne’s (2017) intervention. In this study, students used a digital program, where they read texts and answered multiple-choice questions to check comprehension. The program was adaptive regarding the difficulty level of the texts, and when students clicked on a word the program gave students a pronunciation or showed a definition. In the small-scale study targeting primary–age at risk-readers, Horne found significant effects on measures of reading comprehension and reading accuracy.

In an intervention similar to ours, Wise, Ring, and Olson (2000) tested the impact of using a program for remedial teaching with children aged 7-11. Students read texts, answered multiple-choice questions and the program also allowed students to get speech support in two different formats. In one condition, students received explicit phonological training in small groups, and in the other condition students used the digital program more instead of having phonological training. Furthermore, in the digital part of the intervention subjects were divided in two conditions. In one condition, students received TTS support as whole words (i.e. not segmented) and the program prompted the students to try to read the word again before providing the pronunciation. The other group received segmented support for regular words whilst irregular words were presented as whole words. In this condition, students were prompted to first attempt using a visual segmentation to identify the correct word. If unsuccessful, students would have the segments pronounced. Wise et al. found no significant difference between the two segmentation conditions nor any differences on reading comprehension. Even though the group receiving explicit phonological training outperformed the group that did not in phonological skills and untimed word reading, students using the digital program more extensively gained more in time-limited word reading. In a recent meta-analysis, Wood, Moxley, Tighe, and Wagner (2018) found positive effects on reading comprehension from using TTS and related read-aloud tools for students with reading disabilities.

Other research, on the other hand, does not give much cause for general optimism when assessing the impact of a digital tool for reading instruction aimed at ordinary classrooms. A meta-
analysis of educational technology applications on reading outcomes found that randomized experiments had an average treatment effect of only 0.08 (Cheung & Slavin, 2012; see also Archer et al., 2014). However, there are very few large-scale randomized trials. Cheung and Slavin’s (2012) meta-analysis showed that small studies including fewer than 250 students and non-randomized studies had much larger effect sizes than large, randomized controlled trials. Many of these randomized trials were targeted at students with special needs, which reduces the interventions’ relevance in ordinary classrooms with a large diversity in decoding skills. Some of the relevant randomized trials were implemented by researchers, which may be important for assessing the theoretical effect of the interventions, but reduces generalisability to ordinary classrooms, where teachers may implement with less efficiency (Archer et al., 2014). Yet, Cheung and Slavin (2012) in their meta-analysis found an overall standardized effect of 0.28 from “comprehensive interventions” that, similar to ours, had blends of computer assisted instruction and non-computer activities take centre stage in the interventions. They also found that high intensity programs in which the tested technology was used for more than 75 minutes a week showed a marginally higher standardized effect size (0.19) compared to low intensity programs (0.11).

Cheung and Slavin (2012) neither report which areas of reading are targeted in the reviewed studies nor the technological characteristics of the educational technology applications. A meta-analysis of studies using eBooks in education found no randomized experiments that assessed both decoding and comprehension skills (Zucker, Moody, & McKenna, 2009).

Yee and McIntyre (2013) criticized interventionist and educational reading research for focusing on selected aspects of reading and using incomplete reading theories, which in turn yields incomplete research. In a more recent analytical review of digital reading programs for elementary grades (Jamshidifarsani, Garbaya, Lim, Blazevic, & Ritchie, 2018), a more nuanced picture emerged. It was found that 10 out of 47 interventions targeted more than one area of literacy (for example decoding and comprehension). Of single-component interventions, studies targeting solely phonics skills were twice as frequent as studies targeting comprehension. The authors concluded that the reviewed phonics programs typically train phonological-orthographic matching (e.g. Messer & Nash, 2017; Saine, Lerkkanen, Ahonen, Tolvanen, & Lyttinen, 2011) and that comprehension programs typically teach strategies of comprehension (e.g. Sung, Chang, & Huang, 2008; Ponce, López, & Mayer, 2012). We add the observation that component programs typically train the targeted components in sections rather than integrating and strengthening the sub-skills in authentic reading activities (e.g. Chambers, Slavin, Madden, Abrami, Logan, & Gifford, 2011; Kim, Samson,
Fitzgerald, & Hartry, 2010). Studies that target both phonics and comprehension have shown positive outcomes on both decoding and comprehension (Madden & Slavin, 2017; Ecalle, Kleinsz, & Magnan, 2013; Kim, Capotosto, Hartry, & Fitzgerald, 2011; Cassady & Smith, 2005), whereas other studies produced positive results in decoding skills but not in reading comprehension (Abrami, Borohkovski, & Lysenko, 2015; Chambers et al., 2008).

1.3. The Present Study

To examine if a digital learning tool can develop comprehension and word identification skills by supporting a balance between decoding and comprehension focused instruction, the first author developed a digital learning tool (Gissel, 2015). By digitally supporting students in reading unfamiliar text for meaning whilst focusing their attention on orthographic processing and providing mappings to phonology and exposure to relevant orthographic patterns, the design aimed to balance students’ use of the orthographic, phonological, semantic, and context processes. The design utilized TTS technology so that students could click for support on words that they were unable to decode independently. To counter students’ overuse of TTS support the tool allowed the teacher to control a differentiated access to TTS. The overall aim of the design was to scaffold students’ independent reading of unfamiliar text, i.e., reading for meaning, while strengthening decoding skills and supporting orthographic learning. Computer simulations are prototypical of basic educational psychological research performed in a non-school setting (Levin & O’Donnell, 1999). These strands of research focus on developing the theoretical base for understanding reading and not primarily on putting the theories to practical use to solve educational problems (Shultz, 2010). In our study, connectionist models and theory of word reading, orthographic learning and the interplay between elements in the reading process form the main theoretical base for designing the digital program tested (Seidenberg & McClelland, 1989; Plaut & McClelland, 1993; Seidenberg, 2005, 2007). In our interpretation, the instructional consequences of the connectionist model of reading acquisition could be, that in order to train students in making these processes interact and work together, instructional programs should facilitate that students are exposed to as much text as possible. Furthermore, students must be led to focus on orthographic and phonological processing for word recognition, whilst practicing building and monitoring their comprehension of the text using the meaning and context processors.

In the digital learning tool tested here, when the student has tried to decode a word using the orthographic input, the student can use TTS to map between the highly correlated orthographic and
phonological codes. This fine-tuning and revision of the student’s decoding attempt resembles backpropagation, the learning process used in connectionist simulations of reading processes (Seidenberg, 2007). Student overuse of TTS is countered by supportive features that postpone the use of TTS. The tool supported students’ independent reading of text, i.e., self-teaching (Share, 1999, 2004), by challenging students to decode words of varying complexity and difficulty.

Students were furthermore supported in identifying the difficulty level of each word and the type of challenge in decoding each word. Different types of support were provided for the different types of words which facilitates item-based orthographic learning (Harm & Seidenberg, 2004). The supportive features in the tool promote students’ identification of the relevant spelling patterns and in generalising, thereby supporting orthographic learning. The experimental learning material aims to make students decode using larger subunits of words and prompts students to use different strategies and different grain sizes for different categories of words. The word material in the instructional texts was divided into different categories based on the frequency and regularity of the word or its constituent parts. Each category corresponded to different types of support, e.g., syllabification, rhyme analogy, and exposure to sight words, to facilitate generalization of the relevant orthographic units beyond word-specific knowledge (Seidenberg, 2007).

Students reread the text with successively less support until they reached fluency, followed by a classroom discussion about students’ understanding of the text. The research presented here tested this new digital learning tool against an active control group. The control group used a prototypical, research-based instructional approach to reading instruction in deep orthographies. The system used a direct, systematic, synthetic phonics programme with decodable paper-based texts (Borstrøm & Petersen, 2001). Every student was taught the same in a teacher-managed series of steps. Rules for the most frequent and most regular context sensitive spellings were introduced one after another and trained through reading along and rereading of text, where these features were highly frequent. This single profile of texts for beginning readers, that is, texts that facilitate instructional focus on for example phonetically regular words, is typical (Hiebert, 1999). The teacher initially read the text aloud. Hence, students used familiar texts. The control group was active in the sense that teachers were instructed in spending the same amount of time on the reading instruction as the treatment group, namely 5 lessons (5x45 minutes) per week in 13 weeks.

Based on the findings in the NRP report (2000) the instructional approach in the control condition focusing primarily on decoding of regular words in familiar text might be expected to spill-over on improved comprehension skills. However, we hypothesized that the use of the digital
learning tool to balance the focus on decoding with a higher focus on reading unfamiliar text for meaning (1) would improve students’ comprehension skills relative to the control group, while (2) not having significantly negative effects on students’ decoding skills.

The digital learning tool is targeted at students in Danish grade 2 (8-year olds). At this level, many students start to be proficient in decoding, but the dispersion in their reading skills increases. The gender reading gap for fourth graders persists in nearly all countries according to the Progress in International Reading Literacy Study (PIRLS), and parental support is related to higher reading achievement (Mullis, Martin, Foy, & Hooper, 2017). Despite variation between countries, PISA consistently reports students’ socio-economic status, immigrant status and gender to be heavily correlated with performance in reading (OECD, 2019). The tool was therefore developed to be able to differentiate vis-à-vis the reading skills of different children at this age.

We test the combined effect of a digital learning tool and a balanced focus on decoding and comprehension. We do not test, whether any improvements in comprehension skills could also have been obtained by an analogue, paper-based learning material with greater focus on reading for meaning. We just note that the digital tool utilizes affordances in the digital format for interactive supportive features and support tailored to the reading level of the individual child. This may be more difficult in standard classroom instruction with paper-based reading materials and where a teacher’s capacity to differentiate instruction may be limited (Ross, Morrison, & Lowther, 2010). Indeed, an evaluation of differentiated instruction in Denmark showed that it occurred only to a very limited extent (EVA, 2011). There are no Danish digital teaching systems that facilitate state-of-the-art instruction in decoding or comprehension for second grade. In grades 1-3 in Denmark, digital learning materials are used much less than in higher grades (Bundsgaard, Buch, & Fougt, 2017).

2. Method

2.1. Participants and Procedure

Sample size was predetermined based on a power calculation. Assumptions were based on Schochet’s (2008) estimates of intraclass coefficient (0.15) combined with the assumption that pretest scores and other covariates potentially could explain 50% of the variance in the outcomes. A level of significance of 5% and power of 80% were used. Assuming a class size averaging 20 students and a minimal detectable standardised effects size (Cohen’s $d$) of 0.30, 45 classrooms were required. In hindsight, an effect size of 0.30 was probably optimistic, even though Cheung and
SUPPORTING DECODING AND COMPREHENSION

Slavin (2012), as mentioned, in their meta-analysis found an average effect size of .28 of “comprehensive interventions” like ours.

Teachers responsible for language instruction in 2nd grade classrooms in public schools in Denmark were invited to participate in the experiment through various media. They were informed that they would randomly be assigned to one of the two materials, and should test their students’ reading skills at baseline and at the end of the intervention. The publishers of both learning materials provided learning materials for the students. Fifty teachers from 34 schools from different parts of Denmark enrolled 50 classrooms with a total of 1,088 students in the trial. For teachers that signed up, schools and the research centre signed a formal agreement regulating the collection and use of data. The schools obtained consent from the parents for participation in this specific trial, unless schools already had obtained a general consent from their parents for participation in projects of this kind. Three teachers failed to pretest their students and were therefore excluded before randomization (see Appendix Figure B.1 for attrition flow diagram).

The teachers did not receive compensation for their efforts. The publishers of both learning materials provided learning materials for the students. The teachers did not receive training prior to the intervention, but were instructed to familiarise themselves with the teacher’s manual. Teachers in both experimental conditions were instructed to use the learning materials for a minimum of five lessons (5 x 45 minutes) per week. Teachers were instructed to adhere as closely as possible to the instructions in the teacher’s manuals for both learning materials. The intervention period was 13 weeks, from September 21–December 18, 2015 (one week was a holiday). The posttests were administered no later than, and as close to December 21, 2015, as possible.

Students were pretested three weeks before the start of the intervention allowing for teachers to submit test results online. Following the recommendation of making small, but not pairwise, strata (Imbens, 2011), the 47 classrooms were rank ordered based on their average pretest scores and divided into strata of six classrooms each (one stratum was randomly selected to include only five classrooms). Within each stratum, three schools were randomly assigned to the treatment or active control condition described in the next two sections. Teachers were informed about their assignment after the pretest and shortly before the beginning of the intervention.

2.2. Treatment: Digital TTS-based Learning Material

The supportive features in the digital learning tool tested in this study allowed students to read unfamiliar text, lessened pressure on controlling vocabulary and allowed for instructional use of
texts of greater complexity both verbally and regarding contents, that is, texts closer to the students’ listening levels. The tool thereby was developed to increase the instruction focus on reading comprehension and reading for meaning, without abandoning the focus on supporting students’ decoding skills.

The digital learning tool was targeted at students in Danish grade 2 (8-year-olds). At this level, many students start to be proficient in decoding, i.e. moving towards the consolidated alphabetic phase (Ehri, 2007), but the dispersion in their reading skills increases. For instance, the gender reading gap for fourth graders persists in nearly all countries according to PIRLS (Mullis, Martin, Foy, & Hooper, 2017). The tool was therefore developed to be able to differentiate according to the reading skills of different children at this age.

The working sequence using the digital learning material was the same for every text in the material. Before reading the unfamiliar text, the teacher had a brief discussion with the class about the relevant genre, strategies for reading the text, and a couple of relevant concepts relating to the text. Next, students would read the text using the supportive features in the learning material. In Video 1 the learning sequence and functionalities of the digital learning tool are shown. Teachers learned how to use the system through watching four short instructional videos.

[Insert Video 1 about here]

**Video 1. Video describing the learning sequence and functionalities of the learning tool.**

When the student had tried to decode a word using the orthographic input, the student could use TTS to map between the orthographic and phonological codes. This fine-tuning and revision of the student’s decoding attempt resembles backpropagation, the learning process used in connectionist simulations of reading processes (Seidenberg, 2007). Student overuse of TTS was countered by supportive features that postponed the use of TTS while supporting students in identifying the relevant spelling patterns and in generalising, thereby supporting orthographic learning and strengthening decoding skills.
The learning tool supported differentiation at three levels (C–A). Only the poorest readers started reading the text at level C (Figure 1). The rest of the students in the class started at level B and progressed to level A, reading the same text on all three levels.

In version C, students listened to the text being read aloud by TTS. Each word was highlighted as the text was read aloud in order to give the student exposure to the orthographic material. This process allowed the student to focus his or her limited attention (LaBerge & Samuels, 1974) mainly on comprehending the text. After listening to a page, the student must click on and thereby identify specific written words to check if the student was paying attention to the orthographic representation. The student then moved on to version B of the same text, which was used by all students in the class.

Version B supplied the instructional text with features that support students’ independent reading of the text with TTS technology. The student read until the need for support for decoding arose. Hence, if the student could correctly and independently identify a word, the student did not receive code-focused instruction relating to the word and its constituent parts. Identifying the word was assumed to give students the exposure necessary to strengthen connections between units. Furthermore, the student would have to use input from semantic and context processing to self-monitor correct word identification.

When the student encountered a word that the student could not read independently, the digital learning material provided the student with the target and/or support to help establish the relevant connections between orthography and phonology.
For each word in the text, the difficulty and the importance of the word vis-à-vis orthographic learning determined how much and which kind of support the student would get. The word material in the instructional texts in version B was divided into five different categories with different types of support (e.g., syllabification, rhyme analogy) based on the frequency and regularity of the word or its constituent parts. Each of the five categories had a colour code, and the student had to look below the word to a coloured box to see which category the word belonged to. Appendix A and Video 1 provide a detailed explanation of the colour-code system.

To counter overuse of TTS, the learning material limited or postponed student use of TTS to make students process the orthographic material before activating TTS. For example, if a word was repeated on a page in the text, only the first occurrence of the word could be read aloud by TTS, and words belonging to a rime family could not be read aloud by TTS, only a cue word with the same rime could be read aloud. The most frequent (and highly irregular words) students had to learn as sight words.

In deep orthographies, such as Danish and English, regularities are found at different grain sizes (Brown & Deavers, 1999; Ziegler & Goswami, 2005; Seidenberg, 2007). Hence, instructional methods should develop multiple connections between phonology and orthography instead of only single connections as the whole word and phonics methods do (Hart, Berninger, & Abbott, 1997). The tool tested in our study facilitated this heterogeneous grain focus. Support in the learning material pertained to and helped the student identify the most relevant units in every word to obtain optimal conditions for generalising from an incidence of an orthographic pattern. The supports made use of syllabification and rhyme analogy. For example, rime units allow for maximum generalisation (Harm, McCandliss, & Seidenberg, 2003) because these units show more regularity than are found at the level of individual graphemes and phonemes (Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995). McKenna, Reinking, and Bradley (2003) found that a feature in their talking book design that provided phonics analogies (saying “hat like cat”) was too distractive for the students and made it difficult for them to follow the story. The proposed analogy feature in our study is designed to be less intrusive because only the analogy word is pronounced when clicked.

Furthermore, the colour codes supported students in determining word difficulty, which has been shown to be relatively difficult for both normal and disabled readers (Van Daal & Reitsma, 1993).
During the activities using version B, the teacher observed the students’ interactions with the learning material and reminded the students to attempt decoding every word and to use the support whenever they need it. After reading the text, the student was presented with a multiple-choice task to check comprehension. Next, the teacher managed a discussion of various orthographic features and words. Instead of teaching students phonics rules or features that they had to apply deductively, the discussion could be based on students’ experiences with trying to decode the text.

In version A of the text, the support faded further. There was no visible support for the student, but when clicked, a few of the most difficult words could be read aloud using TTS. The student was led to try to read the text without support and only use support for a few, difficult words. Finally, the teacher managed a discussion with the class about the meaning of the text.

For some words, students could access a spoken explanation or an image representing noun words to aid comprehension. Apart from these supportive features, the collective comprehension work was mostly carried out by the teacher in accordance with the research on effective teaching practices (NRP, 2000; Duke, Pearson, Strachan, & Billman, 2011). The supportive features for decoding and comprehending the texts served to widen the range of matching text difficulty level with student decoding skills, making it possible for the whole class to be able to independently read the same texts. This, in turn, allowed for teacher and peer feedback, classroom discussions, vocabulary teaching, and comprehension work.

Furthermore, as the texts was unfamiliar for students reading only versions B and C, students practiced attacking unknown words using low-level stimulus analysis and self-regulating their reading process using contextual processing as a constraint with respect to word recognition. Students were encouraged to comprehend or gradually shift their attention from decoding to comprehension (Samuels, 1994) as they attain fluency working with the text.

No researchers were involved in the implementation of the programme and teachers did not attend any courses. Whereas previous research suggests that this low degree of control with the implementation process may reduce, if not eliminate, any positive effects of computer-assisted instruction (Archer et al., 2014), the practical relevance of digital learning tools will be vastly increased if they can be used by teachers with a minimum amount of researcher involvement.

2.3. Active Control Group: Standard Phonics-Focused Learning Material

The control learning material, “Den første læsning” (‘The First Reading’; Borstrom & Petersen, 2001) was a paper-based material, which follow the research-based recommendations of focusing
on decoding (NRP, 2000) by prescribing and facilitating a direct, systematic, synthetic phonics programme.

The instructional texts in control group material were decodable, that is, they had strictly controlled vocabulary in which regular words were favoured and where ‘rules’ (Borstrøm & Petersen, 2001) for the most frequent and most regular context sensitive spellings were introduced one after another and trained through reading and rereading of text where these features were highly frequent. The few irregular words were repeated manifold to stimulate rote learning, i.e., handling through the lexical route (Coltheart, 2005).

The teacher must thoroughly prepare students for the encounter with the texts. The procedure for teaching with control group material presented in the teacher’s manual starts with a classroom dialogue about the title and illustrations of each new chapter of basal text. Then, the teacher demonstrates a phonetic analysis of all the new words in the text in dialogue with the class. Next, the teacher reads the text aloud, and the students are supposed to read along and pay attention to the written text. Classroom dialogue about the text strengthens student comprehension of the text. Finally, students reread the text in pairs, discuss new aspects of the text and read it one more time as homework.

The material was the most prevalent analogue reading book system in Denmark in grades K–3 (Bundsgaard, Buch, & Fougt, 2017). Only five of the teachers assigned to the control group worked at a school in which they did not have access to the control learning material before participating in the experiment. The teachers in the control group used the learning material for second grade.

2.4 Fidelity

After the intervention period, a survey assessed implementation fidelity (for the full questionnaire, see Appendix C). All but one teacher reported that they used the learning material they were assigned to, and all their students had access to the learning material. The only teacher that did not use the assigned learning material used the correct material (Den første læsning), but for grade 1 instead of for grade 2. Of the 47 teachers who signed up for the trial, 4 dropped out before the fidelity survey was administered. 42 of 43 (98%) teachers participated in the fidelity survey. Appendix figure B.1 describes the participant flow in the trial.
were asked to retrospectively assess how many lessons they conducted with their assigned learning material. The mean number of lessons for the 21 teachers in the experiment group who answered was 53 (range 40–64), whereas the 19 respondents from the control group averaged 55 lessons in the intervention period (range 45–62) \(t(38) = -0.9911, p = 0.3279\).

Recordings of student activity in the digital learning tool can be used to supplement teachers’ self-reported fidelity. Figure 2 shows the number of minutes of activity for each day during the intervention period. The figure reflects a weekly pattern of little activity on Saturdays and Sundays and that the fourth week was a holiday. More importantly, though, there is a clear trend across the intervention period showing that it took about four weeks (beside the holiday) to reach maximum activity. After that, activity faded to less than a third in the final week compared to the maximum activity week. The total number of minutes of activity is also somewhat smaller than what would have been expected based on teachers’ self-reports. In other words, the recordings in the digital learning tool indicate that teachers did not implement the program with as high fidelity as indicated by their self-reports. In the discussion, we consider the trade-offs involved in this low level of monitoring of teacher activity in the classroom. Closer monitoring might have improved fidelity and thereby maybe produced even bigger effects of the program, but stricter implementation requirements might also have created researcher effects (e.g. Hawthorne effects) or compromised teachers’ teaching domain.

---

2 For data protection reasons we are not able to link the individual student activity to reading test data or individual classrooms. Therefore, we cannot make further analyses of the effect of the tool for students or classrooms that used the tool the most.
2.5. Measures

2.5.1 Tests for Word Identification Skills

A meta-analysis has shown that the size of the positive correlation between decoding and reading comprehension depends on how decoding is measured. Pseudoword reading has lower correlations with reading comprehension than measures of accuracy of single word reading with items presented in a list. Whereas both approaches to measuring decoding have their merits, pseudoword reading may underestimate the relationship with reading comprehension (García & Cain, 2014) – and since the ultimate goal is reading comprehension, not decoding by itself, there are reasons to use measures of decoding that are closer to aspects of decoding that relate to comprehension.

For this study we used the Danish standardized reading test “OS120”, in which students read single words on a list and must match the word to one of four pictures (Nielsen, Kreiner, Poulsen, & Søegård, 1989). The score is based on the number of correct items and the test has 120 items. To avoid the ceiling effect, the time limit for OS120 was set to two minutes. Besides word
identification, the test therefore contains an element of word comprehension, which probably makes its relationship with reading comprehension stronger than decoding measures using pseudoword reading. Even though students could guess rather than read the words, on average students better at decoding should be both faster and better at finding the correct picture than students just guessing. The test is commonly used in Denmark and not specifically designed for the intervention of this study. The test was administered by the teachers.

2.5.2 Tests for Reading Comprehension

For measuring reading comprehension, a Danish translation of four measures from easyCBM® for grade 2 was used; two were used in the pretest (Farm Life and Helping Grandma) and two as the posttest (Rainy Day Running and Feeding the Birds). The reasons for using this material were twofold. First, it has been tested for reliability (Lai, Irvin, Alonzo, Park, & Tindal, 2012). Second, we wanted to test students’ comprehension of text that does not have controlled vocabulary with a preponderance of orthographically regular words. Translating age appropriate texts from the same test system into Danish was a way to ensure that story structure and length were comparable and age appropriate, but without controlled vocabulary. The four measures were all short narratives, which is the genre used for all texts in the control learning material and for most texts in the experiment learning material. Furthermore, each easyCBM® test contains 12 multiple choice items with two subscales: seven questions assessing literal comprehension, and five questions assessing inferential comprehension. Following Lai et al. (2012), missing responses on any question were coded as wrong – unless if the student did not take the test so all responses were missing.

To counter the tendency of floor effects on the measures of reading comprehension in the early grades (Slavin, Lake, Cheung, & Davis, 2009), teachers were asked to choose between three ways of administering the tests to the students basing their choice on their knowledge of their students’ reading skills. If possible, students had to read the text independently. If the teachers found this to be impossible for the student, the students were allowed to use TTS technology. If, for practical reasons, the students did not have access to TTS technology, the teacher would have to read the text and tasks aloud to the student. In the online survey gathering the pre- and posttest scores, teachers had to report which method of administration they had chosen. Only 29 students (3 %) had the text read aloud in the posttest. Robustness analyses test whether controlling for teacher help in both the pretest and the posttest affect the estimated effect of the digital learning tool. The results show that effects are not driven by students receiving teacher help (see Appendix B, Tables
B.2 and B.3). Leaving the 29 students that had the text read aloud out of the analyses does not change the effect estimates.

The alpha reliability coefficients for the test scores used as the outcome for reading comprehension were as follows: literal comprehension .94, inferential comprehension .88, and combined .96.

2.5.3 Correlation between word identification and comprehension measures

Pairwise correlations between the word identification and reading comprehension measures for pre- and post tests are presented in Table 1 for both treatment and control group. The correlations in the pre-test are .33 for control group and .34 for the treatment group. In existing research, the magnitude of variance in reading comprehension explained by word reading ranges from $R^2 = .05$ to $R^2 = .80$ for 8-year-olds (García & Cain, 2014, p. 76).

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Word identification, pre test</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) Word identification, post test</td>
<td>.74**/.82**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3) Reading comprehension, pre test</td>
<td>.33**/.34**</td>
<td>.32**/.33**</td>
<td></td>
</tr>
<tr>
<td>(4) Reading comprehension, post test</td>
<td>.26**/.25**</td>
<td>.24**/.30**</td>
<td>.42**/.37**</td>
</tr>
</tbody>
</table>

Note: Correlations to the left of the slash are for control group, to the right of the slash for treatment group. ** $p < 0.01$.

Test scores from both word identification and reading comprehension tests were standardised to have mean of 0 and standard deviation of 1. Thereby effect estimates are equal to Cohen’s $d$.

2.6. Statistical Methods

Posttest scores were regressed on the treatment indicator in multilevel regression model, which essentially compares the means of the treatment and control group. The multilevel model accounts for the nested structure of our data (students within classrooms). To increase the precision of our effect estimates, we included pretest scores and strata indicator variables as covariates. Missing indicators are included for missing data on the covariates.

Multilevel models and ordinary least squares (OLS) with clustered standard errors produce similar results (Green & Vavreck, 2008). Nevertheless, to assess the robustness of the results, we
compared the multilevel model to an OLS model. At the same time, we compared models that sequentially included (1) only the treatment indicator, (2) pretest scores, (3) strata fixed effects, (4) pretest scores and strata fixed effects, (5) pretest, strata and teacher support in pretest taking, i.e. the aforementioned three ways of administering the tests to the students and (6) pretest, strata, pretest-support and teacher support in posttest taking. Treatment effects turn out to be very consistent across all these alternative model specifications (see Appendix tables B.2. and B.3.)

2.7. Baseline Balance

Table 2 shows that the randomization procedure produced a fine balance between the treatment and control groups. We see only relatively small differences in means between the groups. There were, however, significantly more missing on the pretest in the treatment group, and the mother’s educational attainment was significantly higher in the control group at the 0.1 level. As mentioned, robustness analyses (Appendix B, tables B.2 and B.3) show that results are robust to a number of model specifications including using missing pretest scores as covariates in the regressions.

Table 2. Balancing of student and parent characteristics.

<table>
<thead>
<tr>
<th>Variables</th>
<th>(1) Control</th>
<th>(2) Treatment</th>
<th>Difference (1)-(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Students</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest word identification score</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>Missing pretest score, word identification</td>
<td>0.01</td>
<td>0.03</td>
<td>-0.02</td>
</tr>
<tr>
<td>Pretest comprehension</td>
<td>0.22</td>
<td>0.24</td>
<td>-0.02</td>
</tr>
<tr>
<td>Pretest comprehension missing</td>
<td>0.06</td>
<td>0.07</td>
<td>-0.01</td>
</tr>
<tr>
<td>Pretest comprehension score, literal items</td>
<td>0.23</td>
<td>0.24</td>
<td>-0.01</td>
</tr>
<tr>
<td>Pretest comprehension all missing, literal items</td>
<td>0.06</td>
<td>0.07</td>
<td>-0.00</td>
</tr>
<tr>
<td>Pretest comprehension score, inferential items</td>
<td>0.30</td>
<td>0.31</td>
<td>-0.02</td>
</tr>
<tr>
<td>Pretest comprehension all missing, inferential items</td>
<td>0.13</td>
<td>0.12</td>
<td>0.00</td>
</tr>
<tr>
<td>Student gender</td>
<td>0.50</td>
<td>0.52</td>
<td>-0.01</td>
</tr>
<tr>
<td><strong>Parent background variables</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mother’s educational attainment</td>
<td>0.49</td>
<td>0.44</td>
<td>0.06</td>
</tr>
<tr>
<td>Immigrant/ descendant of immigrant</td>
<td>0.07</td>
<td>0.07</td>
<td>0.00</td>
</tr>
<tr>
<td>Income</td>
<td>252379</td>
<td>238068</td>
<td>14311</td>
</tr>
<tr>
<td>Income</td>
<td>152861</td>
<td>163960</td>
<td></td>
</tr>
<tr>
<td>Number of children in the family</td>
<td>2.26</td>
<td>2.32</td>
<td>-0.06</td>
</tr>
<tr>
<td>N schools</td>
<td>23</td>
<td>24</td>
<td>47</td>
</tr>
<tr>
<td>N students</td>
<td>477</td>
<td>536</td>
<td>1013</td>
</tr>
</tbody>
</table>

Notes. t-tests are corrected for clustering at the classroom level. ***p < 0.01, **p < 0.05 and *p < 0.1. Income is in Danish kroner.
2.8. Attrition

Sample attrition at the cluster level occurred in four instances. Two teachers in the intervention group and two teachers from the control group left the study for various reasons (new job, sickness, etc.). Sample attrition at the student level occurred because students changed schools, were taken out of the trial by the teachers, or were absent for testing. In total, 888 students completed the word identification posttest, and 847 students completed the comprehension posttest. Appendix Figure B.1 shows the flow of participants from enrollment to analysis. Attrition rates in the word identification posttest (12%) and comprehension posttest (16%) were not significantly different between the intervention and control groups (see Appendix table B.1).

3. Results

Table 3 shows the results of regressing each of the reading outcome measures on the treatment indicator variable. To increase precision in the estimates, pre-test variable, indicator of missing pre-test as well as strata indicator variables were included in the regression models. Table entries are the regression coefficients with clustered standard errors in parentheses. Coefficients for the strata indicator variables are not shown in the table. Column 1 shows that the effect of the treatment on word identification was not statistically significant (Cohen’s d = -0.13, p=0.18). These results suggest that students in the treatment group have had almost the same progression in word identification skills as the control group.

Table 3 column 2 shows the results on measures of reading comprehension. Using the experiment learning material had a substantial, positive, and significant effect on reading comprehension (effect size (Cohen’s d) 0.19, p=0.009), as expected. The treatment effect is about twice the average size found in randomized experiments of educational technology applications (Cheung & Slavin, 2012). Robustness analyses presented in appendix tables B.2 and B.3 show that results are similar when using OLS regression models.

<table>
<thead>
<tr>
<th>Variables</th>
<th>(1) Word identification</th>
<th>(2) Comprehension</th>
<th>(3) Comprehension Literal</th>
<th>(4) Comprehension Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>-0.13</td>
<td>0.19***</td>
<td>0.18*</td>
<td>0.17*</td>
</tr>
<tr>
<td></td>
<td>(0.10)</td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>Pretest</td>
<td>0.81***</td>
<td>0.36***</td>
<td>0.36***</td>
<td>0.24***</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.04)</td>
</tr>
</tbody>
</table>
We ran separate models for the two reading comprehension subscales: literal comprehension and inferential comprehension (Table 3, columns 3 and 4). A positive and significant result was found for the treatment group on both literal comprehension (effect size 0.18, p=0.01), and inferential comprehension (effect size 0.17, p=0.02).

The study was not designed to detect heterogeneous effects on subgroups. However, it is important to explore potential imbalances in impact on subgroups to guide future designs. We ran separate models to check for the influence of the following background variables: student gender, mother’s educational level, and country of origin. Table 4 shows a positive and significant effect on comprehension for students in the treatment group whose mother had a low level of education (effect size 0.20, p=0.02), and for students of western origin (effect size 0.22, p=0.003).
Table 4. Heterogeneous intervention effects on student word identification skills and reading comprehension by gender and country of origin, and parental education level.

<table>
<thead>
<tr>
<th>Outcome variables and observations</th>
<th>(1) Girls</th>
<th>(2) Boys</th>
<th>(3) Western</th>
<th>(4) Non-Western</th>
<th>(5) Mother low education</th>
<th>(6) Mother high education</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word identification</td>
<td>-0.10</td>
<td>-0.16</td>
<td>-0.13</td>
<td>-0.29</td>
<td>-0.14</td>
<td>-0.09</td>
</tr>
<tr>
<td></td>
<td>(0.11)</td>
<td>(0.09)</td>
<td>(0.10)</td>
<td>(0.17)</td>
<td>(0.09)</td>
<td>(0.13)</td>
</tr>
<tr>
<td>Observations</td>
<td>436</td>
<td>448</td>
<td>826</td>
<td>58</td>
<td>460</td>
<td>402</td>
</tr>
<tr>
<td>Comprehension</td>
<td>0.19</td>
<td>0.18</td>
<td>0.22**</td>
<td>-0.00</td>
<td>0.20*</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>(0.10)</td>
<td>(0.09)</td>
<td>(0.07)</td>
<td>(0.23)</td>
<td>(0.09)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>Observations</td>
<td>411</td>
<td>432</td>
<td>786</td>
<td>57</td>
<td>435</td>
<td>386</td>
</tr>
</tbody>
</table>

Multilevel model. Standard errors in parentheses. Pretest and strata indicator variables included. High education is defined as short, middle, BA, long higher education or researcher.

\* \(p < 0.05\), \** \(p < 0.01\)

In contrast, for non-western immigrants the estimated effect on comprehension was close to zero (effect size -0.0002, \(p=0.999\)), and we found a negative result for word identification that was close to being significant (effect size -0.29, \(p=0.08\)). Subdividing the sample reduces the statistical power, and we had only 58 non-western immigrants. These subgroup analyses should therefore be interpreted with caution. However, the negative result for non-Western immigrants raises the concern that for students learning the language of instruction, the traditional decoding-focused approach may be better. We return to this question in the discussion.

4. Discussion

This randomized controlled trial testing a computer-based approach to support students in moving into the consolidated alphabetic phase (Ehri, 2007) indicates that the digital support is not significantly worse in terms of word identification, whereas it can give students an advantage in reading comprehension when compared to the active control group. This applies to both literal and inferential reading comprehension. These results indicate, that the tested approach does succeed in balancing the instructional focus on decoding and comprehension.

Reading comprehension is a complex entity that depends on the smooth functioning and interaction of a series of cognitive processes. Compton, Miller, Elleman, and Steacy (2014) call for research that explores interventions for students with reading deficits using connectionist theory to find ways to “mimic and promote the inductive learning mechanisms that characterize typical
reading development” (p. 55). In line with connectionist theory, students were in the present intervention supported in an inductive learning process that aimed to collate and strengthen both phonological, orthographic, meaning and context processing. We targeted whole class teaching. Furthermore, the tested program is designed to be the primary learning resource for reading instruction in second grade, not a supplement.

The digital program tested facilitated that students were supported in making meaning of unfamiliar text whereby they practiced word reading, received whole word feedback on words they found difficult to decode and alphabetic training in a meaningful context, whilst monitoring their comprehension. Furthermore, students were exposed to more text of greater orthographic complexity and more varied vocabulary than in a conventional analogue basal reader. Support was differentiated in three levels potentially allowing every student to reach fluency. Lastly, the teacher played an important role in both preparing students for reading the texts (vocabulary training and comprehension strategies), monitoring students’ interaction with the program, and leading classroom dialogues about the texts.

Testing a digital learning tool that supports the complex task of balancing decoding and comprehension instruction raises issues concerning construct validity, i.e. determining the higher constructs of which the intervention is representative (Shadish, Cook, & Campbell, 2002). An intervention that measures the effects of using a learning material design that operationalizes a host of instructional measures and theoretical issues (diSessa & Cobb, 2004), for example parts of connectionist theory, use of TTS, teacher managed classroom dialogue including modeling of decoding, independent but guided silent reading and repeated reading, complicates unambiguous categorization.

The word identification skill level in the treatment group was lower after the intervention, but, as hypothesized, not statistically significantly lower than the control group, even though the traditional learning material used by the control group facilitates a great deal of decoding training. The intervention taps into the debate about whether direct phonics instruction should be extended beyond the basic grapheme-phoneme correspondences (Castles, Rastle, & Nation, 2018). Based on connectionist theory, we propose that students’ successful reading of a word and, if the student is unable to identify the correct word, that exposure to the most relevant orthographic pattern or a mapping to the word’s pronunciation will suffice.

The tested program is an operationalization and interpretation of connectionist theory. We do not claim that connectionist models are more accurate than dual route models in describing how
SUPPORTING DECODING AND COMPREHENSION

reading and reading acquisition occurs. However, the proposed instructional design is designed to support and strengthen the relevant cognitive processes and interaction between them assuming that connectionist theory is accurate. We think that the connectionist approach has the best chance of success when students have learned the basic grapheme-phoneme correspondences. At that point, facilitation of self-teaching (Cunningham, Perry, Stanovich, & Share, 2002) by maximizing the impact of student exposure to words and orthographic patterns combined with opportunities for extensive text experience seems a feasible approach. Also, exposure to texts without controlled vocabulary can expand students’ vocabulary knowledge, which is a necessary component in comprehension (Spencer, Quinn, & Wagner, 2017).

Because the experiment learning material does not have controlled vocabulary and less teacher led direct instruction, we hypothesize that students using the experiment learning material will become more capable of decoding unfamiliar text and generalize decoding skills beyond the word material found in the instructional texts as well as being more capable of decoding irregular, unfamiliar words. On the other hand, we expect students using traditional phonics-oriented learning materials based on the dual-route models to be more capable of decoding familiar words (i.e. words learned by rote through the teacher’s introduction and repetition in the instructional texts) and regular words, in which each letter has the standard pronunciation. The reading tests used in this study do not specifically measure the decoding of regular and irregular words, so future studies are needed to test this hypothesis.

We believe that our test design is rigorous. The appropriate use of clustered standard errors for cluster-randomized trials (e.g. Green & Vavreck, 2008) sets the study apart from most studies included in the reviews by NRP (2000) and Brooks, Torgerson, and Hall (2008). The use of an active control group teaching with state of the art phonics-based material means that we compare the experiment learning material to the best available research-based treatment on decoding measures. In many other studies the control condition is treatment as usual, which can often be a rather heterogeneous set of practices. This use of an active control group adds significance to the fact that the progression in word identification skills for students in the treatment group was not significantly worse than for the students in the control group.

A large-scale trial comparing the effects of using state-of-the-art analogue textbooks with using a digital learning material cannot be expected to show great advantage on the part of the digital material. In a systematic review of the effect of four different approaches to improving reading success of elementary graders, Slavin, Lake, Cheung, and Davis (2009) found a mean effect...
size (0.12) for reading curricula, i.e., textbooks for initial, non-remedial reading instruction, with an average effect size of 0.23 for decoding measures and 0.09 for comprehension. Only 13 studies of information technology for beginning reading met the standards in the review and showed a mean effect size of 0.09. No studies evaluated teaching in which computer programs were the sole learning material. This could be a symptom of the challenges of making teachers conduct all or most of their teaching with digital learning resources for no less than 12 weeks in an experiment, which is part of the inclusion criteria in Slavin et al. (2009).

4.1 Limitations and Future Research

The study presented here tested a digital learning tool with a balanced focus on decoding and reading for meaning against an active control group that was instructed to spend the same amount of instruction using a standard, paper-based teaching material focusing on decoding. In neither condition did teachers receive training prior to the intervention, and no researchers monitored teacher activity in the classroom during the intervention. These features of the design ensure a high level of external validity in the sense that it would be realistic to obtain the same effect sizes in a subsequent implementation of the intervention. The effect size of around 0.18 found in this study is high compared to both large-scale, randomized trials of computer-assisted instruction (Cheung & Slavin, 2012), and other interventions that target whole classrooms (Lipsey et al., 2012). Notably, interventions applying digital technology tend to show smaller effects when researchers are not involved in the delivery of the intervention (Archer et al., 2014).

However, this design with a high degree of external validity also means that the study cannot answer questions that are more specific. First, would the effect sizes have been larger if implementation fidelity had been higher? Data from the digital learning tool show that student activity was lower than indicated by teacher-reported fidelity – especially towards the end of the intervention period. An important question would therefore be how to support teachers in maintaining a high use of the digital learning tool – without compromising teachers’ authority to decide how they will spend the time in the classroom.

Second, the intervention tested the combination of a digital learning tool with a balanced focused on decoding and comprehension against a traditional paper-based material with a primary focus on decoding. Therefore, we do not know whether we could obtain the same effects by a paper-based tool with balanced focus on decoding and comprehension. The reason for developing the digital tool was exactly that it is difficult to combine the support on decoding with reading
SUPPORTING DECODING AND COMPREHENSION

unfamiliar texts with orthographically irregular words, because students around the age of 8 years are at very different reading levels and need very different support.

Third, despite the relatively large size of this randomized trial, it was not powered to detect heterogeneous subgroup effects. The explorative subgroup analyses indicated, however, that non-Western immigrant students did not benefit from the intervention. It therefore remains an open question whether the weakest decoders got sufficient support and code-focused instruction from interacting with the digital learning material. Previous research has shown that the students who need TTS support the most are the least strategic users (Anderson-Inman & Horney, 2007). Students either over- or underuse the support (McKenna, 1998; Gissel, 2014). In the present intervention, overuse is sought to be countered by the supportive mechanisms. Underuse of the support is less of a problem compared to overuse. From a connectionist perspective, it is productive when students keep trying to decode words instantly and less significant that they are provided with the correct target on every trial; representations that allow generalisation hold more value than word-specific knowledge (Seidenberg, 2007). We would expect a greater gain in word identification skills for the more skilled decoders. In the experiment learning material, students practice reading within their zone of proximal development and spend little time receiving collective code-focused instruction that might be irrelevant for them. Instead, they get supported exposure to more text, more different words, and to more irregular and difficult words than they would with the control material. We also note that we cannot rule out small negative effects on word identification skills. Future research should therefore continue to test for any negative side-effects on decoding skills.

5. Conclusion

Connectionist theory of reading has rarely been put to practical use to solve educational problems (Shultz, 2010), and doing so requires a great deal of interpretation because the learning conditions of a student in second grade will differ significantly from those in computer simulations (Seidenberg, 2007). Using insights from computational modeling of orthographic learning in designing a learning material that facilitates a new instructional approach is a way of putting theory to work and producing both practically applicable and scientifically rigorous results (Levin & O’Donnell, 1999; Kaestle, 1993).

The positive and significant effect on comprehension from using the experiment learning material is promising. The supportive features allow students to read unfamiliar text, lessen pressure
on controlling vocabulary and allow for instructional use of texts of greater complexity both verbally and regarding contents, that is, texts closer to the students’ listening levels.
SUPPORTING DECODING AND COMPREHENSION

6. References


SUPPORTING DECODING AND COMPREHENSION


https://doi.org/10.1002/9780470757642.ch1


https://doi.org/10.1016/s0022-0965(02)00008-5

https://doi.org/10.1207/s15327809jls1301_4

https://doi.org/10.1598/0829.03

https://doi.org/10.1016/j.chb.2013.01.041

https://doi.org/10.1002/9780470757642.ch8
SUPPORTING DECODING AND COMPREHENSION


SUPPORTING DECODING AND COMPREHENSION


SUPPORTING DECODING AND COMPREHENSION


*Journal of Educational and Behavioral Statistics, 33*(1), 62-87. 
[http://dx.doi.org/10.3102/1076998607302714](http://dx.doi.org/10.3102/1076998607302714)


[https://doi.org/10.1093/oxfordhb/9780198568971.013.0014](https://doi.org/10.1093/oxfordhb/9780198568971.013.0014)

[https://doi.org/10.1037/0033-295x.96.4.523](https://doi.org/10.1037/0033-295x.96.4.523)


[https://doi.org/10.1002/pam.10129](https://doi.org/10.1002/pam.10129)

[https://doi.org/10.1111/bjep.12097](https://doi.org/10.1111/bjep.12097)


[https://doi.org/10.1006/jecp.1998.2481](https://doi.org/10.1006/jecp.1998.2481)


APPENDIX A – Supplementary text

Explanation of the Color-Coding System in the Digital Learning Tool

Figure A.1 Example of the colour-codes used in the digital learning tool.

Figure A.1 provides an example of the colour coding system in the digital learning tool. Only the first occurrence of a word on a page can be read aloud using TTS. On the second occurrence, a grey coloured box signifies that the student either has already read the word independently or has attempted to read it on a previous occasion, perhaps using a supportive feature.

Words marked with a green box cannot be read aloud immediately. This category consists of the most frequently used words in the Danish language. These words are for the most part short but highly irregular. The low correlation to the phonological code in these words means that a connection from orthography to semantics has to be strengthened from thorough student exposure to the visual representation of the word. Hence, if the student cannot read a green word independently he has to remember how it is spelled, pick out the word in a pop up box containing
the target as well as other words in the green category. When clicked, the target is read aloud by TTS. That gives the student at least two exposures to the orthographic representation and a pronunciation allowing her to identify the meaning of the word. The teacher can fade this support by selecting which green words the individual student can read independently and therefore does not need support for.

![Image of text with colored boxes]

Figure A.2. Example of Yellow Boxes.

Words with a yellow box are words from onset rime families. The student cannot have the target word read aloud using TTS. Instead, the student gets a cue through a written keyword that rhymes with the target word (see Figure A.2). The student has to transfer the rhyme from the analogous word to the target. The student can have the cue word read aloud with TTS, if necessary. By using rime analogy and substitution of one or two consonants, the student decodes the target word and gets three exposures to the rhyme.

Multisyllabic words that do not fall into one of the other categories and that consist of regular and/or frequent word parts are marked with blue boxes. The student must initially try to read the word independently and, if unsuccessful, try to read the word using the syllables and/or
SUPPORTING DECODING AND COMPREHENSION

morphemes. The student thus identifies and gets exposure to the relevant word parts. As a last option, students can access TTS by clicking once more.

Finally, words marked by a red box either are infrequent words or consist of infrequent word parts and/or word parts with irregular neighbours. The student can access TTS directly with these words. This serves to control student frustration and lessens demands to control vocabulary in the texts. Also in this category are a few words in which each letter represents its most common sound, which should be easy for the student to read. Therefore, no further support is necessary for these words. This mix of hard and easy words might encourage students to try to read words in the red category independently before activating TTS.
SUPPORTING DECODING AND COMPREHENSION

APPENDIX B. Supplementary Analyses

Figure B.1. Attrition flow diagram.
### Table B.1. Test for attrition balance.

<table>
<thead>
<tr>
<th></th>
<th>(1) Control</th>
<th>(2) Treatment</th>
<th>Difference (1)-(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Decoding posttest missing</td>
<td>0.12</td>
<td>0.32</td>
<td>0.13</td>
</tr>
<tr>
<td>Comprehension posttest all missing</td>
<td>0.17</td>
<td>0.38</td>
<td>0.16</td>
</tr>
<tr>
<td>Posttest comprehension all missing, literal items</td>
<td>0.17</td>
<td>0.38</td>
<td>0.16</td>
</tr>
<tr>
<td>Posttest comprehension all missing, inferential items</td>
<td>0.18</td>
<td>0.39</td>
<td>0.16</td>
</tr>
</tbody>
</table>

| N schools   | 23 | 24 | 47 |
| N students  | 477| 536| 1013|

Notes. *t*-tests are corrected for clustering at the classroom level. ***p < 0.01, **p < 0.05 and *p < 0.1.
### Table B.2. Treatment effects on decoding. Robustness analyses.

<table>
<thead>
<tr>
<th></th>
<th>OLS regression</th>
<th></th>
<th>Multilevel</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1a</td>
<td>2a</td>
<td>3a</td>
<td>4a</td>
<td>5a</td>
<td>6a</td>
<td>1b</td>
<td>2b</td>
</tr>
<tr>
<td>Treatment</td>
<td>-0.15</td>
<td>-0.10</td>
<td>-0.10</td>
<td>-0.11</td>
<td>-0.10</td>
<td>-0.08</td>
<td>-0.15</td>
<td>-0.13</td>
</tr>
<tr>
<td></td>
<td>(0.14)</td>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.09)</td>
<td>(0.10)</td>
<td>(0.09)</td>
<td>(0.14)</td>
<td>(0.11)</td>
</tr>
<tr>
<td>Decoding pretest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.81***</td>
<td>0.81***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>Missing indicator pretest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.22</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.22)</td>
<td>(0.25)</td>
</tr>
<tr>
<td>Teacher help missing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ref.</td>
<td>ref.</td>
</tr>
<tr>
<td>Teacher help pretest=1</td>
<td>-0.20</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.28</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>(0.24)</td>
<td>(0.19)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.15)</td>
<td>(0.18)</td>
</tr>
<tr>
<td>Teacher help pretest=2</td>
<td>0.00</td>
<td>-0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.21*</td>
<td>-0.20*</td>
</tr>
<tr>
<td></td>
<td>(0.24)</td>
<td>(0.22)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.10)</td>
<td>(0.10)</td>
</tr>
<tr>
<td>Teacher help pretest=3</td>
<td>0.09</td>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.06</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>(0.15)</td>
<td>(0.14)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.10)</td>
<td>(0.10)</td>
</tr>
<tr>
<td>Teacher help missing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ref.</td>
<td></td>
</tr>
<tr>
<td>Teacher help posttest=1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.54</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.30)</td>
<td></td>
</tr>
<tr>
<td>Teacher help posttest=2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.28</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.26)</td>
<td></td>
</tr>
</tbody>
</table>
## SUPPORTING DECODING AND COMPREHENSION

<table>
<thead>
<tr>
<th>Teacher help posttest=3</th>
<th>0.20</th>
<th>0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(0.21)</td>
<td>(0.08)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.08</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>(0.12)</td>
<td>(0.10)</td>
</tr>
<tr>
<td></td>
<td>-0.33</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>(0.27)</td>
<td>(0.28)</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>(0.23)</td>
<td>(0.23)</td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>(0.34)</td>
<td>(0.34)</td>
</tr>
<tr>
<td></td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>(0.10)</td>
<td>(0.08)</td>
</tr>
<tr>
<td></td>
<td>-0.29</td>
<td>0.30*</td>
</tr>
<tr>
<td></td>
<td>(0.15)</td>
<td>(0.17)</td>
</tr>
<tr>
<td></td>
<td>0.35*</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>(0.19)</td>
<td>(0.19)</td>
</tr>
</tbody>
</table>

### Random

<table>
<thead>
<tr>
<th>Classroom</th>
<th>0.42***</th>
<th>0.33***</th>
<th>0.26***</th>
<th>0.29***</th>
<th>0.31***</th>
<th>0.32***</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>Student</td>
<td>0.90***</td>
<td>0.54***</td>
<td>0.90***</td>
<td>0.54***</td>
<td>0.53***</td>
<td>0.53***</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observations</th>
<th>888</th>
<th>888</th>
<th>888</th>
<th>888</th>
<th>888</th>
<th>888</th>
<th>888</th>
<th>888</th>
<th>888</th>
<th>888</th>
<th>888</th>
<th>888</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.01</td>
<td>0.6</td>
<td>0.12</td>
<td>0.62</td>
<td>0.63</td>
<td>0.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta R^2$</td>
<td>0.59</td>
<td>0.11</td>
<td>0.61</td>
<td>0.62</td>
<td>0.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Standard errors clustered at classroom level in parentheses. Strata indicator variables included in models 3-6.

Teacher help in pretest and posttest: 1 indicates no help, 2 use of TTS, 3 teacher reading text aloud.

* $p<0.05$, ** $p<0.01$, *** $p<0.001$"
<table>
<thead>
<tr>
<th></th>
<th>OLS regression</th>
<th>Multilevel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1a</td>
<td>2a</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.18</td>
<td>0.17*</td>
</tr>
<tr>
<td></td>
<td>(0.13)</td>
<td>(0.08)</td>
</tr>
<tr>
<td>Comprehension pretest</td>
<td>0.38***</td>
<td>0.37***</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>Missing indicator pretest</td>
<td>-0.90***</td>
<td>-0.85***</td>
</tr>
<tr>
<td></td>
<td>(0.23)</td>
<td>(0.23)</td>
</tr>
<tr>
<td>Teacher help missing</td>
<td>ref.</td>
<td>ref.</td>
</tr>
<tr>
<td>Teacher help pretest=1</td>
<td>-0.22</td>
<td>-0.12</td>
</tr>
<tr>
<td></td>
<td>(0.15)</td>
<td>(0.18)</td>
</tr>
<tr>
<td>Teacher help pretest=2</td>
<td>-0.21</td>
<td>-0.28</td>
</tr>
<tr>
<td></td>
<td>(0.11)</td>
<td>(0.14)</td>
</tr>
<tr>
<td>Teacher help pretest=3</td>
<td>-0.28*</td>
<td>-0.14</td>
</tr>
<tr>
<td></td>
<td>(0.13)</td>
<td>(0.13)</td>
</tr>
<tr>
<td>Teacher help missing</td>
<td>ref.</td>
<td>ref.</td>
</tr>
<tr>
<td>Teacher help posttest=1</td>
<td>0.30*</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>(0.14)</td>
<td>(0.14)</td>
</tr>
<tr>
<td>Teacher help posttest=2</td>
<td>0.65***</td>
<td>0.65***</td>
</tr>
<tr>
<td></td>
<td>(0.15)</td>
<td>(0.15)</td>
</tr>
</tbody>
</table>
## SUPPORTING DECODING AND COMPREHENSION

<table>
<thead>
<tr>
<th>Teacher help</th>
<th>posttest=3</th>
<th>0.25*</th>
<th>0.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.10</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td></td>
<td>(0.10)</td>
</tr>
<tr>
<td>Random</td>
<td></td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>Classroom</td>
<td>0.34***</td>
<td>0.34</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td></td>
<td>(0.05)</td>
</tr>
<tr>
<td>Student</td>
<td>0.93**</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>0.02</td>
<td></td>
<td>(0.02)</td>
</tr>
<tr>
<td>Observations</td>
<td>847</td>
<td>847</td>
<td>847</td>
</tr>
<tr>
<td>R²</td>
<td>0.01</td>
<td>0.19</td>
<td>0.18</td>
</tr>
<tr>
<td>ΔR²</td>
<td>0.18</td>
<td>0.04</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Standard errors clustered at classroom level in parentheses. Strata indicator variables included in models 3-6.

Teacher help in pretest and posttest: 1 indicates no help, 2 use of TTS, 3 teacher reading text aloud.

* p<0.05, ** p<0.01, *** p<0.001"
Appendix C. Questionnaire - Fidelity of intervention.

[Translated from Danish]

In the trial in which you have participated you were assigned a learning material by random. Furthermore, we wrote to you that optimally you should use the learning material for five lessons a week (in average) and that you should adhere to the teacher’s guide.

In trials like this, a lot can go wrong. There can be many reasons why you did not follow our guidelines to the full. Therefore, it is of the utmost importance that you respond to this survey to help us clarify how the teaching with your assigned learning material actually took place over the course of the experiment.

Thank you for your cooperation.

[Researchers’ names withheld]

Did you use the learning material you were assigned in the randomization?

Yes/No

[If no] Which learning material did you use instead?

Did you teach the class using the assigned learning material for an average of five lessons (45 minutes) a week in the experimental period?

Yes/No

In the 12 weeks (not counting the fall holiday), you could have taught the students for 60 lessons with the assigned learning material. How many lessons do you assess that you achieved teaching with the assigned learning material?

Write number
What were (apart from the fall holiday) the reasons why you did not achieve teaching around five lessons a week with the learning material?

To what extent did you follow the directions in the teacher’s guide?

Not at all/To some extent/To a great extent/Completely

Describe as accurately as possible how it took place when you taught with the learning material. How much time passed before you got started, did students work alone or in pairs? When did the students work at the PCs/with independent reading in the textbook, how did they read, when did you present something or manage a dialogue, about what did you present/manage dialogue etc.?

Were there students in the class that you chose not to teach with the assigned learning material?

Yes/No

[If yes] State the students’ name and the reason why you chose to offer these students another solution?

Did every single student have access to the learning material? Did every student have access to *Læsebogen* and *Arbejdsbogen* for *Den første læsning* 2 – or did every single student have access to a device with the digital learning material?

Yes/No

[If no] How many copies/devices were available in class?
How was the class taught in grade 1?

(You can select more than one)

- A phonologically oriented learning material (such as *Den første læsning*, *Nisserne i Ådal*)
- A material that mixes phonological with other approaches (such as *Tid til dansk*, *Freja og Eskild*, *Bogslottet*)
- Writing students’ own texts using invented spelling
- Independent reading of books with suitable LIX/LET-grade for the student’s level
- Other

Thank you for your cooperation.