INTERNATIONAL LITERACY ASSOCIATION READING RESEARCH QUARTERLY

Assessing the Effectiveness of Structured Word Inquiry for Students in Grades 3 and 5 With Reading and Spelling Difficulties: A Randomized Controlled Trial

Danielle Colenbrander
Macquarie University, Sydney, New South Wales, Australia; and University of Bristol, UK

Liam Parsons
Essex County Council, UK

Jeffrey S. Bowers
Colin J. Davis
University of Bristol, UK

ABSTRACT

Research syntheses have demonstrated that morphological instruction can improve the literacy skills of poor readers and spellers. However, studies have used a wide variety of training methods. Questions remain about what type of morphological instruction is most effective and under which circumstances. In this study, we conducted a randomized controlled trial evaluating the effectiveness of Structured Word Inquiry for poor readers and spellers. Structured Word Inquiry teaches students the logic of the English spelling system via instruction in morphology and etymology. Students in grades 3 and 5 with poor reading and spelling skills were randomly assigned to receive either Structured Word Inquiry instruction or a comparison instruction method involving robust vocabulary instruction and reciprocal teaching. Instruction was delivered by teaching assistants over the course of a full school year. After training, there were few differences between the groups in terms of literacy or vocabulary skills. However, teaching assistants found Structured Word Inquiry instruction challenging to deliver, which is likely to have impacted the results. Our findings have implications for the nature and content of morphological instruction for poor readers and spellers, and for future attempts to scale up the delivery of morphological interventions.

In the later years of elementary school, students are expected to read and write independently and to learn from what they read. However, a significant minority of students have difficulties with reading and spelling. For example, in the 2019 U.K. Statutory Assessment Tests, 27% of students in grade 6 failed to meet the government's expected standard for reading, and 22% failed to meet the expected standard for spelling, punctuation, and grammar (Department for Education, 2019). In some cases, such students have an existing history of literacy difficulties, and in other cases, students who previously appeared to read adequately may struggle to keep up with the increasingly complex texts presented to them and may begin to fall behind their peers (Chall & Jacobs, 1983; Wanzek, Wexler, Vaughn, & Ciullo, 2010). Whatever the reason, reading and spelling difficulties are likely to have negative consequences for students' educational achievement, occupational choices, and emotional health (Boetsch, Green, & Pennington, 1996; DeWalt, Berkman, Sheridan, Lohr, & Pignone, 2004; McLaughlin, Speirs, & Shenassa, 2014; OECD, 2013). Therefore, it is important to determine which methods of literacy instruction and remediation are most effective for students in the upper elementary years.

INTERNATIONAL LITERACY ASSOCIATION

Reading Research Quarterly, 57(1)
pp. 307–352 | doi:10.1002/rrq.399

The Role of Morphology in Reading and Spelling Development

When students learn to read, they must learn to link the written forms of words to their phonological forms and meaning. At first, this process is gradual and effortful, but with instruction and reading experience, students begin to be able to recognize written words rapidly and efficiently (e.g., Share, 1995, 2011). The more efficiently students can access word meanings from their written forms, the more resources become available for processing higher order meaning. Efficient access to lexical information is enabled by high-quality, well-specified knowledge of a word's orthography (written form), phonology, meaning, and morphology (e.g., Perfetti, 2007; Verhoeven & Perfetti, 2011). High-quality lexical representations are also crucial for accurate spelling (e.g., Andrews, Veldre, & Clarke, 2020; Levesque, Breadmore, & Deacon, 2021). In both reading and spelling, morphological knowledge acts as a binding agent, strengthening relations among orthography, phonology, and meaning (Kirby & Bowers, 2017; Perfetti, 2007). Indeed, the English language is morphophonemic (Venezky, 1999); it contains regular grapheme–phoneme correspondences but tends to preserve morphological and etymological information at the expense of grapheme–phoneme regularity. For example, the spelling of *sign* preserves the historical and semantic relation it shares with other words, such as s*ignature*, *signal*, and *design* (J.S. Bowers & Bowers, 2017).

Typically developing students show implicit sensitivity to basic aspects of morphological structure in written words from the early years of schooling (e.g., Carlisle & Stone, 2005; Deacon & Bryant, 2006; Treiman & Cassar, 1996). There is evidence that both morphological awareness (the ability to manipulate morphological information in spoken language) and morphological decoding and spelling abilities continue to grow throughout the elementary years and beyond, as students are exposed to increasing numbers of morphologically complex words (e.g., Anglin, 1993; Berninger, Abbott, Nagy, & Carlisle, 2010; Carlisle & Kearns, 2017; Singson, Mahony, & Mann, 2000; Tyler & Nagy, 1989). A particularly large increase in exposure to complex derived words appears to occur around grade 3 (Goodwin, Lipsky, & Ahn, 2012; Nagy & Anderson, 1984). Different aspects of morphological knowledge also appear to contribute to reading comprehension in multiple ways. For example, morphological awareness is directly associated with reading comprehension and is also associated with better ability to decode and decipher the meanings of morphologically complex words, which are in turn related to reading comprehension (e.g., Levesque et al., 2021; Levesque, Kieffer, & Deacon, 2017).

Morphological Instruction

Although students can develop implicit knowledge of morphological structure, explicit instruction may provide an additional boost to the development of literacy skills. With explicit morphological knowledge, students may be better able to use cues from the spellings of words to decipher and remember phonology and meaning, and better able to use cues from meaning to deduce and remember correct spellings. This in turn may lead to improved lexical knowledge and more fluent reading, both of which may enable improved reading comprehension (Kirby & Bowers, 2017).

Evidence from research syntheses supports this view. In a systematic review of 22 studies, P.N. Bowers, Kirby, and Deacon (2010) found that morphological instruction was effective for improving sublexical skills (oral morphological awareness, phonological awareness, and nonword decoding) when compared with untreated control groups. The researchers found that the effect size of morphological instruction was larger for poor readers. Goodwin and Ahn (2010) conducted a meta-analysis of data from 17 studies involving students with literacy difficulties and found that morphological interventions were effective for improving phonological awareness, morphological awareness, vocabulary knowledge, spelling skills, and reading comprehension. Effect sizes were small to medium. A further meta-analysis of the effects of morphological intervention on school-age students' literacy outcomes (Goodwin & Ahn, 2013) found that effect sizes were largest for English learners, followed by students with learning disabilities and poor readers and spellers, although these differences did not reach significance.

Beyond these research syntheses, a number of studies with linguistically diverse samples have explored the effectiveness of including morphological instruction in interventions targeting vocabulary, reading comprehension, and academic language (e.g., Goodwin, 2016; Lesaux, Kieffer, Faller, & Kelley, 2010; Lesaux, Kieffer, Kelley, & Harris, 2014; Proctor, Silverman, Harring, Jones, & Hartranft, 2020). These interventions resulted in improvements to students' vocabulary, morphology, reading, and reading comprehension outcomes. The interventions were either equally effective (e.g., Lesaux et al., 2010) or more effective for students whose primary language was not English (e.g., Lesaux et al., 2014). Three of these interventions (Lesaux et al., 2010; Lesaux et al., 2014; Proctor et al., 2020) were multicomponent interventions, and it is therefore difficult to analyze the specific role played by morphology instruction. However, the study by Goodwin (2016) explicitly compared the combination of morphology and comprehension strategy instruction with comprehension strategy instruction alone. Students in the combined morphology and comprehension condition performed better on tasks of vocabulary and generation of morphologically related words, indicating that morphology instruction provided a boost over and above comprehension instruction alone. There were no differences between the groups on other aspects of morphological awareness or on measures of word-reading fluency and reading comprehension, but it is important to note that this was a relatively short intervention, only four sessions in length.

Despite this work, relatively little is known about the effectiveness of different methods of morphological instruction and for whom they are effective. Studies have used a wide variety of different methods of instruction, ranging from explicit instruction in prefixes and suffixes to multicomponent training programs with additional intervention components, such as phonological awareness training or reading practice. Explicit morphology instruction may be particularly crucial for poor readers and spellers, who tend to have lower levels of morphological knowledge than other students of the same age and also tend to have particular difficulty in reading and spelling morphologically complex words (e.g., Breadmore & Carroll, 2016; Carlisle, 1987; Carlisle & Katz, 2006; Treiman & Bourassa, 2000). However, although P.N. Bowers et al. (2010) and Goodwin and Ahn (2013) found larger effect sizes for poor readers, there were fewer studies with poor readers than with typically developing or undifferentiated samples, and the sample sizes of studies with poor readers was generally small, with the majority having fewer than 40 participants.

There is also a need to further explore the effectiveness of morphology-based instruction for diverse samples of students with poor reading and spelling abilities for their age. English-speaking countries typically have diverse populations of learners. In the United Kingdom, over 20% of primary school students speak English as an additional language (EAL[1]; Office for National Statistics, 2020). These students come from a wide range of language backgrounds and have varying degrees of English proficiency (see, e.g., Demie, 2018). Whereas many students who speak EAL have age-appropriate reading and spelling abilities, others do not, and it is important to determine what type of instruction is most effective for enabling them to reach their full potential (e.g., August & Shanahan, 2006, 2010; Demie, 2018; Murphy & Unthiah, 2015). Morphology instruction is a promising method for this purpose (see, e.g., Goodwin & Ahn, 2013; Lesaux et al., 2010, 2014).

The Present Study

In this study, we aimed to test a specific method of morphology-based instruction, Structured Word Inquiry (SWI). SWI is an inquiry-based method that makes sense of spellings by teaching students that spellings are organized around the interrelation of morphology, etymology, and phonology. It has been shown to be effective in improving the vocabulary skills of typically developing students in grades 4 and 5 (P.N. Bowers & Kirby, 2010), spelling in grades 3 and 4 (Devonshire & Fluck, 2010), and decoding in grades 1 and 2 (Devonshire, Morris, & Fluck, 2013). In a recent study of one-on-one instruction with 48 participants (Georgiou, Savage, Dunn, Bowers, & Parrila, 2021), SWI was also shown to improve reading and morphological awareness skills in persistently poor readers, when instruction was delivered by trained research assistants. However, there have been no large-scale studies exploring whether SWI would be effective for improving the literacy skills (reading, spelling, and reading comprehension) and vocabulary knowledge of students with poor reading and spelling skills, and to date, studies of SWI have only included students who are native speakers of English.

In the current study, we recruited students in grades 3 and 5 with poor reading and spelling abilities (both native and EAL speakers) and compared SWI with an alternative intervention program that we called Motivated Reading (MR). In MR, students learned the meanings of new words via robust vocabulary instruction (e.g., Beck, McKeown, & Kucan, 2013), saw their written forms, and learned comprehension strategies through reciprocal teaching (Palincsar & Brown, 1984), but did not learn anything about the morphology or etymology of words.[2] Both robust vocabulary instruction and reciprocal teaching have been shown to improve the reading comprehension of typically developing students and those with specific reading comprehension difficulties (e.g., Beck, Perfetti, & McKeown, 1982; Clarke, Snowling, Truelove, & Hulme, 2010; McKeown, Beck, Omanson, & Perfetti, 1983).

This comparison condition was chosen because we hoped to test the specific effects of learning about word structure, over and above the general effects of small-group instruction and instruction in word meanings in the context of written text. Goodwin (2016) explored a similar question, revealing some benefits for vocabulary knowledge in the condition combining morphology and comprehension instruction, but this study was of short duration. Because SWI instruction, taught over an extended period of time, involves delving deeply into the sublexical structure of English words and identifying relations between words, students may learn general principles about English spelling and meaning. This knowledge may strengthen students' ability to read, spell, and understand words to a greater degree than a method of instruction which focuses on lexical and supralexical information (i.e., learning the meanings of whole words, broader strategies for deducing meaning in context), and therefore, SWI instruction may be more likely than MR to have transfer effects beyond words that are directly taught.

Finally, we were interested in various factors that might influence the effectiveness of SWI and MR. This study is novel in that instruction was delivered by teaching assistants (TAs), rather than expert teachers. In England, TAs represent approximately a quarter of the workforce in elementary schools (Skipp & Hopwood, 2019). Deployment of TAs can be a cost-effective way of improving students' educational attainment, and research has shown that TAs are most effective when delivering structured one-to-one or small-group interventions (Sharples, Webster, & Blatchford, 2018). However, to date, no studies have explored whether SWI can be effectively delivered by TAs. Thus, in our study, we compared the effectiveness of these two methods when administered by paraprofessionals in typical school conditions. Furthermore, we were interested in whether the effectiveness of instruction would differ depending on participants' ages (grade 3 compared with grade 5) and initial levels of reading ability. Finally, we were interested in whether the effectiveness of instruction would differ depending on whether students were native or non-native speakers of English.

Method

Methods are described according to CONSORT guidelines (Boutron, Moher, Altman, Schulz, & Ravaud, 2008a, 2008b; Schulz, Altman, & Moher, 2010). The study was approved by the University of Bristol Ethics Committee, and the trial protocol was preregistered on the Open Science Framework (Colenbrander, Davis, Bowers, & Parsons, 2016). Parents gave informed written consent, and children gave verbal assent.

Participants and Sample Size

We contacted over 160 elementary schools in and around Bristol, England, by letter, email, and telephone. Thirteen schools agreed to participate in the study. Screening assessment was conducted in April and May 2016 to identify suitable participants. All students in grades 2 and 4 at each school (with parental consent) were screened on a 20-item version of the Diagnostic Spelling Test–Nonwords (DiSTn; Kohnen, Colenbrander, Krajenbrink, & Nickels, 2015) and the New Group Reading Test (NGRT; GL Assessment, 2010). Students who met screening criteria and received parental consent participated in the study when they were in grades 3 and 5.

The DiSTn was developed in Australia and is a test of the ability to spell monosyllabic nonsense words that contain regular letter–sound correspondences. We used a shortened version adapted for use with a U.K. sample. The NGRT is a standardized, group-administered test of reading comprehension, in which students silently read sentences and passages and then select answers from five multiple-choice options.

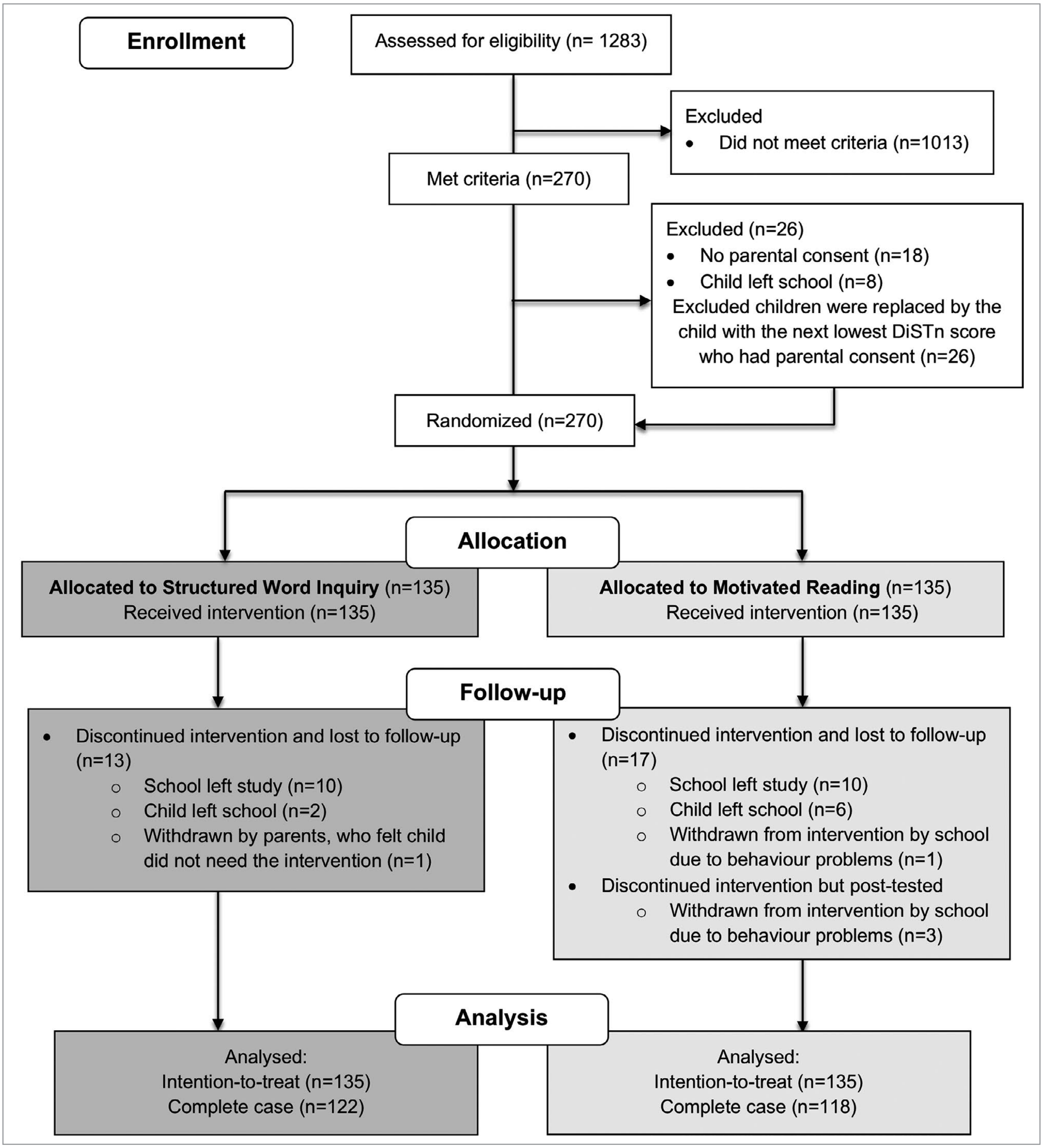
Figure 1 shows the flow of participants through the study. In total, 1,283 students were screened, and those with the 10 lowest DiSTn raw scores from each grade level were invited to participate in the study. When two students had the same score, the one with the lowest NGRT score was invited to participate. If a student did not receive parental consent to participate or left the school before the start of the new academic year, the student with the next lowest spelling score was invited until there were 10 participants in each grade level at each school—in other words, a total of 20 participants per school. There were two exceptions to this. First, we agreed to allow one large elementary school to have 20 participants from each grade level (40 participants in total), as they were in an area with a transient population and were concerned about group sizes becoming too small over the course of the study. Second, one of the participating schools only served students in grades 3–6. Therefore, only grade 4 students were screened, so only 10 students from this school participated. In total, there were 270 participants at the start of the trial, who were randomized into two groups of 135 (see Figure 1).

We used a within-school selection method so there would be two groups of five participants in each grade at each school. Other selection methods could have resulted in uneven numbers at different schools, creating a confound with group size and possibly rendering small-group instruction infeasible at some schools. Mean DiSTn scores for each training group varied across schools. Grade 3 group means varied from 3.80 to 8.50 out of 20 with a median score of 6.40, whereas grade 5 group means varied from 4.50 to 12.40 with a median of 8.65. However, *t*-tests confirmed that DiSTn scores were significantly lower for the selected participants than for students who were not invited to participate (grade 3 participants: mean [*M*] = 6.19, standard deviation [*SD*] = 2.63; grade 3 nonparticipants: $M = 11.80$, $SD = 3.62$, $t < 0.001$; grade 5 participants: $M = 8.28$, $SD = 3.42$; grade 5 nonparticipants: $M = 13.84$, $SD = 3.05$, $t < 0.001$).

Trial Design

This was a crossover trial in which eligible students were randomly assigned to receive either SWI or MR for a full school year and then receive the other intervention in the second year. In this article, we report results from the first year of the study. We decided to compare SWI with MR rather than with a wait-list control because SWI had previously been compared with a business-as-usual control group but not with an alternative training program. Therefore, we wished to conduct a stringent test of SWI over and above the effects of small-group instruction and exposure to whole-word forms and meanings in context.

FIGURE 1
Participant Flow Diagram


Enrollment
Assessed for eligibility (n= 1283)
Excluded
• Did not meet criteria (n=1013)
Met criteria (n=270)
Excluded (n=26)
• No parental consent (n=18)
• Child left school (n=8)
Excluded children were replaced by the child with the next lowest DiSTn score who had parental consent (n=26)
Randomized (n=270)
Allocation
Allocated to Structured Word Inquiry (n=135)
Received intervention (n=135)
Allocated to Motivated Reading (n=135)
Received intervention (n=135)
Follow-up
• Discontinued intervention and lost to follow-up (n=13)
○ School left study (n=10)
○ Child left school (n=2)
○ Withdrawn by parents, who felt child did not need the intervention (n=1)
• Discontinued intervention and lost to follow-up (n=17)
○ School left study (n=10)
○ Child left school (n=6)
○ Withdrawn from intervention by school due to behaviour problems (n=1)
• Discontinued intervention but post-tested
○ Withdrawn from intervention by school due to behaviour problems (n=3)
Analysis
Analysed:
Intention-to-treat (n=135)
Complete case (n=122)
Analysed:
Intention-to-treat (n=135)
Complete case (n=118)

Note. DiSTn = Diagnostic Spelling Test–Nonwords.

MR is based on programs which have led to improved reading comprehension for typical readers and students with specific reading comprehension difficulties (Beck et al., 1982; Clarke et al., 2010; McKeown et al., 1983; Palincsar & Brown, 1984), although the programs' effectiveness for students with word-reading and spelling difficulties has not yet been tested. Students in our study were pretested on the outcome measures in June and July 2016, received training throughout the 2016–2017 school year, and were posttested in June and July 2017.

Randomization

A protocol specifying the composition of the groups was developed by our project adviser, Chris Rogers, using stratified randomization with a block size of 10. Each stratum represented one grade level at each school, except in the case of the large school where two strata were combined into one larger one. This protocol was sent to an independent researcher who randomized students to training groups using the training protocol. This researcher was not involved in the study and had access only to the students' anonymous code numbers and the randomization protocol.

Interventions

Procedure

Each group was scheduled to receive three 20-minute lessons per week for a total of 24 weeks, distributed across the whole school year. Twenty-six TAs, employed by the participating schools, delivered the lessons. TAs in British schools do not need to have formal teacher training and are usually employed to deliver targeted interventions or assist teachers within the classroom. Levels of instructional experience in our sample of TAs varied widely, from approximately two years of experience to approximately 25 years. Within each grade level in each school, the same TA delivered the intervention to both groups. In this way, we counterbalanced instructor effects. To avoid spillover effects, TAs were instructed to ensure that students in one group did not see any of the training materials from the other group and to avoid using language or strategies from one intervention when teaching the other. Regular lesson observations and discussions with TAs showed that these instructions were generally followed.

Lesson Content

For both programs, manuals with detailed lesson plans were provided. Example manuals are available on the Open Science Framework (https://osf.io/bvt5a/?view_only=de49aeafe51e432a8f66bfc18ce40fd8).

Across both SWI and MR programs, we chose a set of words to be directly taught to both groups (see Appendix A). To select these words, we developed a list of frequent and productive bases and derivational prefixes and suffixes. This list was compiled using data on prefix and suffix frequency (Beyersmann, Castles, & Coltheart, 2012; Blevins, 2001), and information on instructional content from P.N. Bowers (2009). From this list, we selected 28 training words and 22 untrained words. These were matched as closely as possible for number of letters and CBBC frequency from the SUBTLEX database (van Heuven, Mandera, Keuleers, & Brysbaert, 2014). We also included a range of different parts of speech.

In SWI, students were guided through an inquiry process to determine the morphological and etymological properties of words. Each week, there were two compulsory lessons in which students learned words and core concepts via inquiry activities, and there was one practice lesson. The aim of the practice lesson was to give TAs an opportunity to follow student inquiries further or to revise and secure core concepts. Trained words were spread across the 24 weeks in a way that fitted with the sequence of morphological and etymological instruction (e.g., in general, morphologically simpler words were taught earlier in the program and words with greater morphological complexity taught later). Once all trained words had been taught, TAs were free to choose other words but were requested not to teach the matched untrained words.

Visual tools such as word matrices (see Figure 2) and flowcharts (see Figure 3) were used to help students understand the links between word structure and meaning. Thus, students learned how to identify the morphological components of words (prefix, base, and suffix) and the difference between etymological and morphological relations (e.g., the words *plea* and *please* share a historical root but do not have the same morphological base; i.e., *please* cannot be broken down to *plea* + *se*). Students learned the difference between free bases (bases that can stand alone as a

FIGURE 2
Word Matrix

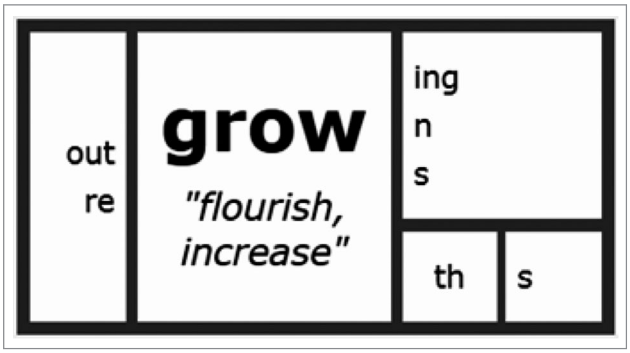
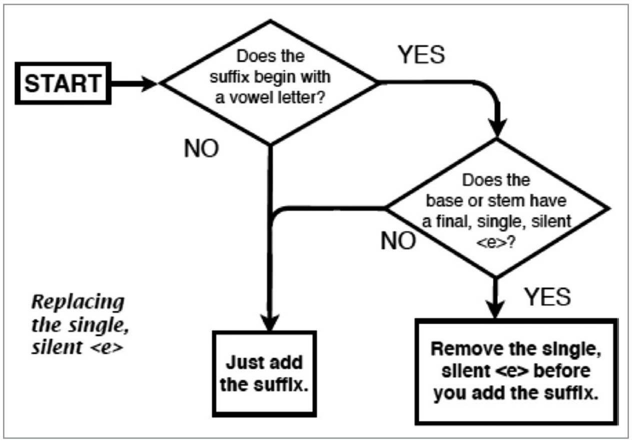


FIGURE 3
Flowchart


START
Does the suffix begin with a vowel letter?
YES
NO
Does the base or stem have a final, single, silent <e>?
NO
YES
Replacing the single, silent <e>
Just add the suffix.
Remove the single, silent <e> before you add the suffix.

word; e.g., *free*, *walk*) and bound bases (bases that carry meaning but need to be connected to another morpheme to form a word; e.g., *struct-* in *structure*, *-sist-* in *assistance*). Students learned to use the Online Etymology Dictionary (https://www.etymonline.com/) to help deduce the base forms of morphologically complex words and built families of etymologically related words using word matrices and word sums. In addition, students learned how morphology influences spelling (e.g., the addition of certain suffixes might trigger spelling changes, such as dropping single word-final *-e*, consonant doubling, and word-final *-y* changing to *-i*), and about orthographic conventions in English (e.g., English words end in *-ve* rather than *-v* unless they are borrowings or contractions).

By contrast, in MR, students received direct instruction in the words' meanings and were exposed to the written forms of the words, but did not learn about the words' morphological structure or etymology. As described earlier, MR is based on two methods of instruction that have previously been shown to be effective for typical readers and students with poor reading comprehension: robust vocabulary instruction (Beck et al., 2013) and reciprocal teaching (Palincsar & Brown, 1984). As noted earlier, we wished to determine whether instruction in morphology and etymology would offer benefits over and above instruction in word meanings and would guide experience with written texts.

Students received two lessons a week of reciprocal teaching. During these lessons, students selected a book or passage to read and were guided to apply reciprocal teaching strategies (i.e., clarification, summarization, prediction, question generation) to the text. Age-appropriate books on interesting topics were donated to the schools for this purpose by Oxford University Press. In the third lesson of each week, students learned two or three vocabulary words via robust vocabulary instruction, which aims to help students develop deep conceptual knowledge of a word by providing multiple encounters with the word in context and involving students in rich discussions about word meanings (Beck et al., 2013). For 14 of the 24 weeks, two of the words taught had to be the trained words, which were also taught in the SWI program. One of the words was chosen by the group or the TA. For the remaining 10 weeks, TAs were free to choose other words to teach. Given the different content and structure of the two training programs, it was not possible to ensure that words were taught in the same order across both programs.

Outcome Measures

Overview

We were interested in whether the two intervention programs would result in different outcomes for students' reading, spelling, and reading comprehension abilities. We were also interested in whether there would be any differences in vocabulary knowledge, as the contents of both training programs have been shown to lead to vocabulary improvements in typical readers (e.g., Beck et al., 1982; P.N. Bowers & Kirby, 2010). Finally, we were interested in whether there would be differences in students' motivation to read. Proponents of SWI have suggested that such instruction may be inherently motivating because it involves actively processing and solving problems related to meaning–spelling connections (P.N. Bowers & Kirby, 2010).

For the reading, spelling, and vocabulary knowledge outcomes, we were additionally interested in whether instruction would result in transfer to words that were not directly trained. One of the potential advantages of morphological instruction is that students may learn how to decipher the meanings and spellings of unfamiliar words, and if this is the case, then we should expect generalization to untrained words after instruction (e.g., Goodwin & Ahn, 2013; Kirby & Bowers, 2017), although it is also the case that some studies of robust vocabulary instruction have found improvements on words that were not directly trained (e.g., Beck et al., 1982; McKeown et al., 1983). Therefore, we included measures of both trained and untrained words for the literacy and vocabulary outcomes.

Our reading comprehension measure (the NGRT) was administered at screening in April/May 2016 when the students were in grades 2 and 4, as were some other group-administered outcome measures (morphological spelling of nonwords and multiple-choice vocabulary, described later). All other outcome measures were administered at pretesting in June/July 2016. Students were posttested on the outcome measures in June/July 2017, when they were in grades 3 (average age 8 years 4 months) and 5 (average age 10 years 4 months).

Assessments were administered by trained research assistants who were blind to group membership. Data entry was also blinded and was carried out subsequently by the same research assistants and by trained undergraduate research apprentices. Tests with verbal responses were audio recorded and scored from these recordings. For all reading and spelling assessments, at least 5% of the data was double-entered, and the percentage agreement between entries was calculated. If percentage agreement was below 80%, data for the entire outcome measure were double-entered. Any discrepancies were resolved by a third blinded rater. For the morphologically complex vocabulary task, students' verbal definitions had to be scored on a 2-point scale. This scoring had a greater subjective element than the other outcome measures, so a more stringent inter-rater reliability procedure was followed (see the Experimenter-Designed Vocabulary Task subsection).

In our preregistration, we specified both primary and secondary outcome measures. Here, for simplicity, we describe and report only our primary outcome measures, but our analytical notes and the full results of all outcome measures are provided in Appendix B.

Reading Outcome Measures

Experimenter-Designed Reading Task

Students read 32 morphologically complex words aloud. Seven of these items were trained words with free bases (e.g., *movement*). Seven other items were trained words with bound bases (e.g., *assistance*); students in both conditions were exposed to the words during instruction, but only students in the SWI condition were taught to identify the relevant bound bases (i.e., *-sist-*, from the Latin *sistere*). Yet another seven items were nonsense words made up of the affixes and bases seen in trained words (e.g., *actable*). These were included as a measure of near transfer, that is, whether students were able to read words containing novel combinations of trained morphemes. A further seven were untrained words with free bases (e.g., *misplace*), and the final four were untrained words with bound bases (e.g., *implosion*). The untrained words were included as a measure of far transfer. There were fewer of the untrained words with bound bases because it was expected that they might be challenging. As noted earlier, the untrained words were roughly matched to the trained words on length and CBBC frequency from the SUBTLEX database (van Heuven et al., 2014). The nonwords were matched on length. Cronbach's alpha was .85 for the trained items, .78 for the nonword items, and .86 for the untrained items. Inter-rater agreement was 94.45%.

Test of Word Reading Efficiency

The Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999) is a standardized task of word- and nonword-reading fluency. In the Sight Word Efficiency subtest, students are asked to read aloud as many words as fast as they can in 45 seconds. This acted as a measure of far transfer to broader word recognition abilities. In the Phonetic Decoding Efficiency subtest, students read monosyllabic nonwords aloud. This was another measure of far transfer and was included to determine whether students' decoding abilities improved as a result of training. Inter-rater agreement was 97.98% for sight word decoding and 87.76% for pseudoword decoding. Because students read different numbers of words depending on their reading abilities, Cronbach's alpha could not be calculated. However, alternate-form internal consistency values reported in the test manual are high, ranging from .86 to .98 (Torgesen et al., 1999).

Spelling Outcome Measures

Experimenter-Designed Spelling Task

This task was the spelling equivalent of the experimenter-designed reading task described earlier. The spelling task contained the same types of items: 14 trained words with free and bound bases, seven nonwords made up of trained bases and affixes to measure near transfer, and 11 untrained words with free and bound bases to measure far transfer. We matched items on the spelling task to items on the reading task for length and SUBTLEX frequency (van Heuven et al., 2014). The trained items contained the same bases as those in the reading task but with different affixes (e.g., *interact* in the reading task and *counteract* in the spelling task). Cronbach's alpha was .89. Inter-rater agreement was relatively low (78.5%) because the handwriting of the youngest participants was particularly poor for this task, which may be related to the fact that the items were quite challenging for these participants (there is evidence of a link between poor handwriting and poor spelling; e.g., Abbott & Berninger, 1993; Sumner, Connelly, & Barrnett, 2014). Therefore, the entire test was double-entered, and an extra independent rater checked all discrepant responses.

Morphological Spelling Test–Nonwords

The Morphological Spelling Test–Nonwords (MoSTn) is a test of knowledge of morphological spelling principles developed by researchers at Macquarie University in Australia (Kohnen, Colenbrander, Caruana, & Barisic, 2016). We administered a 20-item version of the task adapted for the U.K. context. In the task, students heard morphologically complex nonwords in a sentence context (e.g., "This one is quite mog, but that one is much mogger. Spell *mogger*"). This test assesses the ability to apply key morphological spelling rules (dropping final *-e*, changing *-y* to *-i*, consonant doubling). Nonsense words were used to ensure that we were testing knowledge of morphological spelling principles, rather than lexical orthographic knowledge. Because students had to apply morphological knowledge to unfamiliar bases, this test was included as a measure of far transfer. Cronbach's alpha was .81, and inter-rater agreement was 96.75%.

Reading Comprehension

Reading comprehension was measured by the NGRT (GL Assessment, 2010). This task was administered during the screening phase. One form of the test is designed for grades 2–4 and consists of 20 sentence completion and 28 reading comprehension questions, all multiple choice. Another form of the test is designed for grades 5–8 and consists of 20 sentence completion and 32 comprehension questions. This meant that grade 3 students completed pre- and posttesting on the same version of the test and that grade 5 students completed a different version. Therefore, we conducted analysis on test standardized scores. Cronbach's α was .95.

Vocabulary

Experimenter-Designed Vocabulary Task

In this task, students provided definitions for words and nonsense words made up of trained bases and affixes. Half the items came from the experimenter-designed reading task and the other half from the experimenter-designed spelling task. As with the other two experimenter-designed tasks, in this task, there were seven trained words with free bases, seven trained words with bound bases, seven nonwords made up of trained bases and affixes (near transfer), seven untrained words with free bases, and four untrained words with bound bases (far transfer). Items could receive a score from 0 to 2 points. Two points were awarded if a student's response demonstrated understanding of the meaning of the word's base element and affixes or if the student gave a correct definition for the word (e.g., for the word *repayment*, 2-point responses included "if you pay someone again" and "a credit card repayment, to pay something off"). One point was awarded if a student demonstrated understanding of either the base or suffix but not both (e.g., for *retold*, "if you tell something"), or if a student used a word correctly in a sentence, but there was no clear evidence that the student knew the meaning of the word (e.g., for dispense, "you dispense something").

The tests were scored by nine research assistants, blind to group membership. The research assistants received half an hour of training in how to score the words and were provided with a scoring rubric containing examples of 0-, 1-, and 2-point responses. All research assistants initially scored 15 of the same tests (approximately 3% of the total sample). Krippendorff's alpha (Krippendorff, 2011) was then calculated using the R package irr (Gamer, Lemon, Fellows, & Singh, 2017; R Core Team, 2019). Krippendorff's alpha was .83, which is considered high. The research assistants then compared and discussed their scores for the 15 tests with other raters until consensus was reached. Once this process was complete, the raters single-entered the remaining tests. If they were unsure how to score an item, they discussed it with other raters until consensus was reached. Cronbach's alpha for this test was .78.

Multiple-Choice Vocabulary Knowledge

We administered a shortened, group-administered version of the British Picture Vocabulary Scale (Dunn, Dunn, & Styles, 2009; Dyson, Best, Solity, & Hulme, 2017) as a measure of far transfer. Students saw four pictures, heard a word, and then had to choose which of the four pictures best matched the word by circling their response on the answer sheet. This task was administered during screening assessment to all students in grades 2 and 4 at the participating schools, and at posttest to training participants only. There were 30 items, but at pretest some grade 2 classes (grade 3 by the time training began) found the task difficult and were only administered 25 items. This did not seem to limit scores; the highest score by a grade 2 student administered the full set of items was 21. Therefore, we included all grade 2 pretest data in the analyses, regardless of whether students were administered 25 or 30 items. Cronbach's alpha was .88.

Motivation to Read

Motivation to read was measured via an adaptation of the Motivation to Read Profile–Revised (Malloy, Marinak, Gambrell, & Mazzoni, 2013). Students answered 20 multiple-choice questions about their self-concept as a reader and how much they liked reading. Responses formed a Likert-type scale (e.g., "Reading a book is something I like to do" had the options never, almost never, sometimes, and often). All data for this assessment were entered by a single research assistant, so no inter-rater agreement was calculated. Cronbach's alpha was .84.

Blinding

Given the nature of the trial, it was impossible to blind training study participants or TAs to the type of training they were receiving or administering. However, testing and data entry were carried out by research assistants who were blind to group membership.

Training Fidelity

TAs attended a four-day training workshop. For the first three days, they received training in SWI, delivered by Peter Bowers. TAs learned about SWI terminology, core concepts of SWI (e.g., "spelling preserves meaning"), and tools of SWI (e.g., word sums, word matrices, flowcharts). TAs saw video examples of instruction and participated in interactive activities allowing them to practice instructional methods. On the final day of the workshop, TAs received training in MR, delivered by the first author. They received instruction in the principles of robust vocabulary and reciprocal teaching methods and participated in interactive activities, such as coming up with example sentences for word meanings, identifying targets for clarification, practicing think-alouds, and role-playing summarization activities. We decided to devote more time to training the TAs in SWI because the MR methods were comparatively familiar to the TAs and required less detailed knowledge of language structure.

Detailed lesson plans were provided, although in SWI, one lesson per week for the first 20 weeks was designated a practice lesson, during which the TA could decide to explore a word of the group's choice or reinforce a previously taught concept, and the final four weeks of instruction were set aside for free inquiry or revision, for which lesson templates were provided, but no content was compulsory. This built-in flexibility was important given the inquiry-based nature of SWI. In MR, content for the first two terms was also delivered from detailed lesson plans (two reciprocal teaching lessons and one vocabulary lesson per week). However, for the second two terms, TAs were given lesson templates, a list of words to teach and a list of texts for students to choose from, but had more freedom to decide when and how to apply the strategies taught in the first two terms. In this way, across both programs, there were approximately 10 weeks in total in which TAs had greater flexibility to choose content.

Each school was visited approximately once a fortnight by either the first or second author or an intern. During this time, the researchers completed fidelity checklists (examples are available at https://osf.io/a4y76/?view_only=c8458f1af979477f8add5d7582c4f15d). Lessons were rated on a number of criteria, such as whether lesson materials were prepared and organized, whether key lesson concepts were covered in the given time, whether TAs used appropriate terminology, and whether students were focused on the task. Ratings were assigned on a scale of 1 (*poor*) to 4 (*excellent*). After each visit, the TAs received targeted feedback and support to address any areas where fidelity was low (e.g., modeling of relevant aspects of instruction, strategies for managing behavior).

At the end of the first year, the TAs were also asked to complete a fidelity questionnaire. On a scale of 1 (*not at all true*) to 4 (*very true*), the TAs rated how confident they were in delivering each training program, how challenging they found lesson delivery, and practical issues such as whether they had enough time to prepare. The TAs also received templates for attendance records so we could keep track of the amount of training each student received.

Results

Numbers Analyzed and Participant Flow

Two hundred seventy students were randomized to training groups and began the training programs in September 2016. One school left the study in January 2017 because they felt the training was not improving students' phonics skills. This resulted in the loss of 20 participants. Eight students were lost to the study because they changed schools. One student was withdrawn from the study by his parents, who felt that he did not need the intervention. A further four students were withdrawn from the training programs at the schools' requests due to severe behavioral problems that were disrupting other students; we were able to posttest three of these students. This meant that 237 students completed the training programs, and 240 students were posttested on the majority of the outcome measures (the sample size for each outcome measure is 240 unless reported otherwise). Of this sample, 49% were female, and 33% were eligible for free school meals (a measure of socioeconomic disadvantage). Forty-one percent of our participants were identified by their schools as EAL speakers, representing a very diverse group of students who ranged from bilingual students born in the United Kingdom to those who had arrived in the United Kingdom within the pprevious two years. These students spoke a range of languages at home, including Somali, Polish, Italian, Punjabi, and Urdu.[3]

Outcomes

Analysis

Our analyses were preregistered on the Open Science Framework, but we made a number of changes to them in response to statistical advice and reviewer comments. First, we initially specified that we would conduct an intention-to-treat analysis on each of our outcome measures, including all participants who had been randomized to a group, even if they later dropped out of the study. However, we could not impute data for the school that withdrew, because there were no posttest data at all from that cluster. For the remaining 250 participants, the majority of our missing data were on our dependent variables, and there is evidence that imputation of dependent variables can lead to noise in the resulting estimates (e.g., von Hippel, 2007). For this reason, we instead ran a complete-case analysis with data from the 240 participants who had been both pre- and posttested. Only 3% of the data was missing from the sample of 250 students, the demographic characteristics of both samples were nearly identical (see Table 1), and the pattern of results was substantially the same across all analytic methods. Full details of the originally planned analyses, including details of missing data and results of all analytic models, are provided in Appendix B.

TABLE 1
Comparisons of Demographic Data for the Complete-Case (CC) and Intention-to-Treat (ITT) Samples

	Structured Word Inquiry				Motivated Reading			
	ITT sample (*N* = 125)		CC sample (*N* = 121)		ITT sample (*N* = 125)		CC sample (*N* = 119)	
Demographic category	%	*n*	%	*n*	%	*n*	%	*n*
Female	52.00	65	52.89	64	44.80	56	42.86	51
Grade 3	48.00	60	47.93	58	48.00	60	47.06	56
Speakers of English as an additional language	42.40	53	42.15	51	40.00	50	40.34	48
Eligible for free school meals	30.40	38	30.58	37	38.40	48	36.97	44

Second, in response to suggestions from reviewers, we reduced the total number of outcome measures by creating composite scores for trained and untrained reading, spelling, and vocabulary measures. We created composite scores by averaging the centered and standardized scores for the relevant assessments.[4] Thus, for the purposes of analysis, the following were our eight outcome measures:

1. *Trained reading (near transfer):* Average of trained word and trained nonword scores from the experimenter-designed reading task
2. *Untrained reading (far transfer):* Average of untrained word scores from the experimenter-designed reading task and the TOWRE Sight Word Efficiency and TOWRE Phonetic Decoding Efficiency subtests
3. *Trained spelling (near transfer):* Average of trained word and trained nonword scores from the experimenter-designed spelling task
4. *Untrained spelling (far transfer):* Average of untrained word scores from the experimenter-designed spelling task and MoSTn scores
5. *Trained vocabulary (near transfer):* Average of trained word and trained nonword scores from the experimenter-designed vocabulary task
6. *Untrained vocabulary (far transfer):* Average of untrained word scores from the experimenter-designed vocabulary task and the multiple-choice vocabulary task
7. *Reading comprehension:* NGRT scores
8. *Motivation to read:* Scores from our adaptation of the Motivation to Read Profile–Revised

For each of these outcomes, we conducted a hierarchical linear regression analysis. The selection of model terms was driven by our research questions. We included fixed effects of group, pretest score, age (grade 3 vs. grade 5), and whether or not each participant spoke EAL. We included a fixed effect of eligibility for free school meals as a covariate.[5] We also included random intercepts for school to account for clustering at the school level. Continuous variables were grand mean centered and standardized. We conducted analyses using the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017) in the R software environment (R Core Team, 2019) and investigated interactions using the emmeans package (Lenth, 2019). (Data and analytic scripts are available at https://osf.io/dx39t/?view_only=522debdd4b9442a1a6fd3ce3dcc2a947.)

Raw score means and standard deviations for each assessment at each testing point are reported in Table 2, and correlations between the outcome measures are reported in Table 3. Full results of all analytic models are shown in Tables 4–11. For all outcome measures, the main effect of pretest was highly statistically significant, such that students with higher pretest scores also tended to have higher posttest scores (see Tables 4–11). Next, we report all other significant main effects and interactions.

Reading

On the trained reading measure, there was a significant interaction between group and pretest score (β = −0.25, standard error [*SE*] = 0.10, *t* = 2.47, *p* = .01). Students who had lower scores at pretest tended to have higher posttest scores in the MR condition than in the SWI condition. For students with higher pretest scores, the opposite effect occurred (see Figure 4). On the untrained reading measure, none of the main effects or interactions was significant.

Spelling

On the trained spelling measure, none of the main effects or interactions reached significance, but two results approached significance: the interaction between group and grade (β = 0.35, *SE* = 0.19, *t* = 1.82, *p* = .07) and the interaction between group and native language (β = −0.34, *SE* = 0.18, *t* = 1.89, *p* = .06). The interaction between group and grade reflected a trend whereby students in the SWI group tended to score more highly than students in the MR group in grade 3 (SWI estimated marginal mean [*EMM*] = −0.08, *SE* = 0.10; MR *EMM* = −0.37, *SE* = 0.10, *p* = .03), but there was no such trend in grade 5 (SWI *EMM* = 0.10, *SE* = 0.10; MR *EMM* = 0.16, *SE* = 0.10, *p* = .64). The interaction between group and native language reflected a trend toward higher scores for EAL speakers in the SWI group as compared with the MR group (SWI *EMM* = 0.03, *SE* = 0.11; MR *EMM* = −0.26, *SE* = 0.11, *p* = .04) but no difference between the groups for the native speakers (SWI *EMM* = −0.005, *SE* = 0.09; MR *EMM* = 0.05, *SE* = 0.09, *p* = .63).

For the untrained spelling measure, the sample size was 209. This was largely due to missing data from the MoSTn. The MoSTn was administered to whole classes during normal class time, and because it was usually administered at the beginning of the session, some students missed the test because they arrived late to class or went directly to other school activities (e.g., music lessons) instead of returning to class after recess. On this measure, the interaction between group and grade was significant (β = 0.41, *SE* = 0.17, *t* = 2.39, *p* = .02). In grade 3, when all other factors were held constant, students in the SWI group scored higher than students in the MR group (SWI *EMM* = 0.03, *SE* = 0.10; MR *EMM* = −0.25, *SE* = 0.09, *p* = .02). The difference between the groups was not significant in grade 5 (SWI *EMM* = 0.03, *SE* = 0.09; MR *EMM* = 0.16, *SE* = 0.09, *p* = .27). There was also a significant main effect of eligibility for free school meals, with students who were eligible scoring lower than students who were not, when all other factors were taken into account (β = −0.20, *SE* = 0.09, *t* = 2.30, *p* = .02). Finally, the

TABLE 2
Descriptive Statistics for all Outcome Measures

	Pretest		Posttest	
Outcome measure	**SWI *M* (*SD*)**	**MR *M* (*SD*)**	**SWI *M* (*SD*)**	**MR *M* (*SD*)**
Morphological reading task				
• Trained words	6.52 (3.28)	7.33 (3.73)	11.25 (2.41)	11.78 (2.23)
• Untrained words	5.81 (3.40)	6.72 (3.48)	5.91 (3.21)	6.54 (2.99)
• Generalization nonwords	3.15 (1.76)	3.70 (1.87)	4.57 (1.48)	4.80 (1.28)
TOWRE Phonetic Decoding Efficiency	20.82 (10.64)	25.70 (12.71)	27.24 (11.00)	31.14 (12.59)
TOWRE Sight Word Efficiency	48.60 (13.64)	52.24 (14.78)	58.82 (10.68)	62.00 (13.10)
Morphological spelling task				
• Trained words	2.19 (1.73)	2.97 (2.56)	4.42 (2.59)	5.13 (3.22)
• Untrained words	1.44 (1.60)	2.31 (2.39)	2.93 (2.24)	3.75 (2.71)
• Generalization nonwords	1.46 (1.24)	1.56 (1.54)	2.67 (1.56)	2.63 (1.69)
Morphological Spelling Test–Nonwords	3.56 (2.87)	4.29 (3.27)	5.90 (3.21)	6.32 (3.48)
Diagnostic Spelling Test–Nonwords	7.12 (3.16)	7.82 (3.53)	8.77 (3.46)	9.80 (3.72)
Morphological vocabulary task				
• Trained words	4.42 (2.26)	4.89 (2.28)	6.43 (3.12)	7.32 (3.56)
• Untrained words	7.60 (2.95)	7.82 (2.76)	9.74 (2.80)	10.11 (2.69)
• Generalization nonwords	2.77 (1.87)	3.05 (1.96)	4.21 (2.37)	4.53 (2.51)
Multiple-choice vocabulary task	15.62 (4.41)	15.79 (4.70)	18.22 (4.08)	18.89 (3.95)
New Group Reading Test (reading comprehension)[a]	92.27 (9.94)	92.98 (10.21)	94.36 (11.22)	94.72 (10.46)
Motivation to read	2.92 (0.37)	2.93 (0.45)	2.98 (0.37)	2.91 (0.42)

Note. MR = Motivated Reading; SWI = Structured Word Inquiry; TOWRE = Test of Word Reading Efficiency.
[a]Standard score.

interaction between group and EAL status approached significance (β = −0.31, *SE* = 0.16, *t* = 1.91, *p* = .06). As with the trained outcome measure, this reflected a nonsignificant trend toward higher scores for EAL speakers in the SWI group relative to the MR group (SWI *EMM* = 0.07, *SE* = 0.10; MR *EMM* = −0.16, *SE* = 0.10, *p* = .07) but no evidence of a difference between the groups for native speakers (SWI *EMM* = −0.01, *SE* = 0.08; MR *EMM* = 0.07, *SE* = 0.08, *p* = .45).

Vocabulary

The sample size for the trained vocabulary measure was 237 (data for three additional students were missing due to behavioral issues during assessment). On this measure, no main effects or interactions were significant, apart from the main effect of pretest. The sample size for the untrained vocabulary measure was 234 (data for six additional students were missing due to absence and behavioral issues during assessment). On the untrained measure, the main effects of grade and EAL were significant. With all other factors held constant, students in grade 5 scored higher on average than students in grade 3 (β = 0.40, *SE* = 0.12, *t* = 3.39, *p* < .001), and native speakers scored higher on average than EAL speakers (β = −0.25, *SE* = 0.10, *t* = 2.43, *p* = .02).

Reading Comprehension

The sample size for this analysis was 238 (data were missing for two additional students, one due to absence and the other due to behavioral issues during testing). No main effects or interactions were significant, apart from the main effect of pretest.

Motivation to Read

The sample size for this analysis was 213. The motivation to read task was administered by TAs during the first and last group lessons and then had to be returned to the researchers. Two TAs lost the posttest data for their groups (20 students), and data for another seven students were missing due to absence on the day or testing or behavioral issues.

TABLE 3
Correlations Between Outcome Measures

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Trained reading pretest	—															
2. Trained reading posttest	.74*	—														
3. Untrained reading pretest	.88*	.75*	—													
4. Untrained reading posttest	.79*	.80*	.89*	—												
5. Trained spelling pretest	.68*	.55*	.69*	.64*	—											
6. Trained spelling posttest	.76*	.66*	.78*	.78*	.72*	—										
7. Untrained spelling pretest	.66*	.52*	.69*	.64*	.79*	.74*	—									
8. Untrained spelling posttest	.73*	.65*	.77*	.78*	.72*	.84*	.79*	—								
9. Trained vocabulary pretest	.34*	.20*	.33*	.32*	.32*	.34*	.35*	.34*	—							
10. Trained vocabulary posttest	.46*	.37*	.47*	.47*	.45*	.53*	.46*	.45*	.45*	—						
11. Untrained vocabulary posttest	.46*	.36*	.46*	.40*	.48*	.44*	.47*	.43*	.56*	.52*	—					
12. Untrained vocabulary pretest	.53*	.43*	.50*	.47*	.52*	.53*	.51*	.51*	.52*	.64*	.76*	—				
13. Reading comprehension pretest	.41*	.36*	.44*	.41*	.39*	.45*	.41*	.47*	.26*	.34*	.39*	.39*	—			
14. Reading comprehension posttest	.41*	.44*	.44*	.47*	.44*	.51*	.46*	.51*	.36*	.40*	.42*	.48*	.69*	—		
15. Motivation to read pretest	.12	.10	.14	.13	.07	.14	.08	.08	.05	.02	−.09	−.09	.07	.09	—	
16. Motivation to read posttest	.00	.01	.05	.08	−.04	.03	−.03	.01	.04	−.03	−.13	−.12	.10	.18*	.60*	—

*p < .05.

There were significant main effects of group (β = −0.40, *SE* = 0.18, *t* = 2.23, *p* = .03) and grade (β = −0.31, *SE* = 0.16, *t* = 2.01, *p* = .045). With all other factors held constant, students in the SWI group tended to rate their motivation to read more highly than students in the MR group, and students in grade 3 tended to rate their motivation to read more highly than students in grade 5. There was a nonsignificant trend toward an interaction between group and grade (β = 0.39, *SE* = 0.22, *t* = 1.78, *p* = .08). This reflected a trend toward higher ratings for the SWI group in grade 3 (SWI *EMM* = 0.22, *SE* = 0.11; MR *EMM* = −0.16, *SE* = 0.12, *p* = .02) but similar ratings for the MR group in grade 5 (SWI *EMM* = −0.09, *SE* = 0.11; MR *EMM* = −0.07, *SE* = 0.11, *p* = .90).

Fidelity and Attendance

Fidelity Checklists

Fidelity checklists were completed by three different raters (the first and second authors and an intern). These consisted of 22 statements about practical issues, instructional features, and student engagement and behavior. Fourteen of the 22 statements were identical across conditions, such as “materials are prepared and organized” and “paces lessons appropriately.” Eight of the statements were different across conditions, relating to specific instructional elements. For example, in MR, one of the specific instructional statements was “prompts students to provide the reasoning behind their responses,” and one of

TABLE 4
Reading Results: Trained Reading

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	0.05	0.11	0.45	.65
Eligibility for free school meals (eligible)	0.001	0.09	0.02	.99
Group (Motivated Reading)	−0.13	0.14	−0.95	.34
Pretest score	0.87	0.07	12.08	<.001***
Grade level (5)	−0.14	0.13	−1.10	.27
EAL (EAL speakers)	0.03	0.12	0.27	.78
Group × Pretest interaction	−0.25	0.10	−2.47	.01*
Group × Grade interaction	0.20	0.19	1.09	.28
Group × EAL interaction	0.04	0.16	0.26	.79
Random effect	**Variance**		***SD***	
Intercept (school)	0.02		0.12	
Residual	0.37		0.61	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	487.24		.56	.58

Note. EAL = English as an additional language.
*p < .05. ***p < .001.

TABLE 5
Reading Results: Untrained Reading

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	0.04	0.07	0.64	.53
Eligibility for free school meals (eligible)	−0.09	0.06	−1.68	.09
Group (Motivated Reading)	−0.06	0.09	−0.62	.54
Pretest score	0.89	0.05	19.52	<.001***
Grade level (5)	−0.08	0.08	−1.09	.31
EAL (EAL speakers)	0.09	0.07	1.16	.25
Group × Pretest interaction	−0.01	0.07	−0.08	.94
Group × Grade interaction	0.14	0.12	1.18	.24
Group × EAL interaction	−0.13	0.10	−1.29	.20
Random effect	**Variance**		***SD***	
Intercept (school)	0.003		0.06	
Residual	0.15		0.39	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	281.65		.81	.81

Note. EAL = English as an additional language.
***p < .001.

the specific SWI statements was “demonstrates appropriate use of word sums.” Statements were rated on a scale of 1 (*poor*) to 4 (*excellent*).

The second author and the intern each received training in how to complete the checklists and then observed two training lessons with the first author, where both raters

TABLE 6
Spelling Results: Trained Spelling

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	−0.04	0.12	−0.30	.75
Eligibility for free school meals (eligible)	−0.11	0.10	−1.14	.25
Group (Motivated Reading)	−0.12	0.15	−0.79	.43
Pretest score	0.52	0.08	6.26	<.001***
Grade level (5)	0.18	0.13	1.39	.17
EAL (EAL speakers)	0.03	0.13	0.23	.82
Group × Pretest interaction	0.11	0.11	1.04	.30
Group × Grade interaction	0.35	0.19	1.82	.07
Group × EAL interaction	−0.34	0.18	−1.89	.06
Random effect	**Variance**		***SD***	
Intercept (school)	0.02		0.12	
Residual	0.45		0.67	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	536.44		.47	.48

Note. EAL = English as an additional language.
***p < .001.

TABLE 7
Spelling Results: Untrained Spelling

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	0.06	0.10	0.59	.56
Eligibility for free school meals (eligible)	−0.20	0.09	−2.30	.02*
Group (Motivated Reading)	−0.12	0.13	−0.97	.34
Pretest score	0.79	0.08	10.43	<.001***
Grade level (5)	0.003	0.11	0.03	.98
EAL (EAL speakers)	0.08	0.12	0.68	.50
Group × Pretest interaction	−0.07	0.10	−0.67	.50
Group × Grade interaction	0.41	0.17	2.39	.02*
Group × EAL interaction	−0.31	0.16	−1.91	.06
Random effect	**Variance**		***SD***	
Intercept (school)	0.01		0.12	
Residual	0.31		0.55	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	393.78		.62	.64

Note. EAL = English as an additional language.
*p < 0.05. ***p < .001.

completed the checklist. Any discrepancies in the ratings were discussed. After that point, most of the sessions were only rated by a single rater. This was for logistical reasons, as there were 50 groups that needed to be observed approximately once a fortnight in 12 different schools. However, to compute inter-rater reliability, three further sessions were

TABLE 8
Vocabulary Results: Trained Vocabulary

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	−0.11	0.13	−0.88	.38
Eligibility for free school meals (eligible)	−0.01	0.11	−0.14	.89
Group (Motivated Reading)	−0.02	0.16	−0.10	.92
Pretest score	0.53	0.09	6.07	<.001***
Grade level (5)	0.18	0.14	1.30	.20
EAL (EAL speakers)	−0.05	0.15	−0.35	.72
Group × Pretest interaction	−0.20	0.13	−1.54	.12
Group × Grade interaction	0.30	0.20	1.48	.14
Group × EAL interaction	−0.14	0.20	−0.71	.48
Random effect	**Variance**		***SD***	
Intercept (school)	0.03		0.17	
Residual	0.54		0.73	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	569.5		.28	.32

Note. EAL = English as an additional language.
$^{***}p < 0.001$.

TABLE 9
Vocabulary Results: Untrained Vocabulary

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	−0.09	0.09	−1.01	.32
Eligibility for free school meals (eligible)	0.01	0.08	0.08	.94
Group (Motivated Reading)	−0.08	0.12	−0.63	.53
Pretest score	0.62	0.07	8.75	<.001***
Grade level (5)	0.40	0.12	3.39	<.001***
EAL (EAL speakers)	−0.25	0.10	−2.43	.02*
Group × Pretest interaction	−0.09	0.10	−0.89	.38
Group × Grade interaction	0.16	0.16	0.99	.33
Group × EAL interaction	0.02	0.15	0.16	.88
Random effect	**Variance**		***SD***	
Intercept (school)	0.004		0.06	
Residual	0.29		0.54	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	418.49		.60	.61

Note. EAL = English as an additional language.
$^{*}p < .05$. $^{***}p < 0.001$.

double-rated by the first and second authors, and four were double-rated by the first author and the intern. Krippendorff's alpha was calculated for these sessions.

Inter-rater reliability was moderate (α = .60), largely due to differences of 1 point in statement ratings (e.g., one rater gave a particular statement a rating of 2, and the other

TABLE 10
Reading Comprehension Results

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	0.08	0.12	0.65	.52
Eligibility for free school meals (eligible)	−0.17	0.10	−1.59	.11
Group (Motivated Reading)	0.06	0.16	0.36	.72
Pretest score	0.64	0.07	8.96	<.001***
Grade level (5)	−0.02	0.14	−0.15	.88
EAL (EAL speakers)	−0.11	0.14	−0.81	.42
Group × Pretest interaction	0.02	0.10	0.22	.83
Group × Grade interaction	0.08	0.19	0.39	.70
Group × EAL interaction	−0.19	0.20	−0.94	.35
Random effect	**Variance**		***SD***	
Intercept (school)	0.01		0.08	
Residual	0.53		0.73	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	566.42		.47	.47

Note. EAL = English as an additional language.
***$p < .001$.

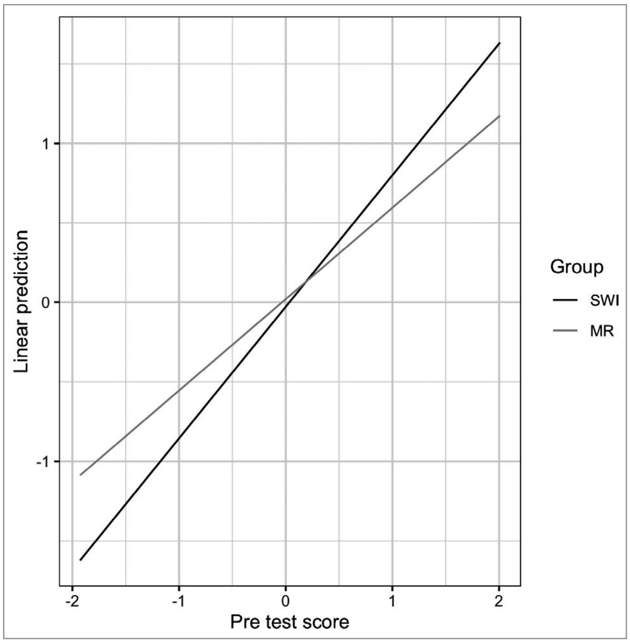
TABLE 11
Motivation to Read Results

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	0.29	0.13	2.37	.02*
Eligibility for free school meals (eligible)	−0.21	0.12	−1.78	.07
Group (Motivated Reading)	−0.40	0.18	−2.23	.03*
Pretest score	0.62	0.09	7.16	<.001***
Grade level (5)	−0.31	0.16	−2.01	.045*
EAL (EAL speakers)	0.06	0.16	0.37	.71
Group × Pretest interaction	−0.01	0.12	−0.08	.93
Group × Grade interaction	0.39	0.22	1.78	.08
Group × EAL interaction	0.06	0.22	0.25	.80
Model fit	**Multiple R^2**		**Adjusted R^2**	
	.41		.38	

Note. EAL = English as an additional language. Results are from a regression model without random effects, as the fit for the model with random intercept for school was singular.
*$p < .05$. ***$p < .001$.

gave it a rating of 3). Because inter-rater reliability was only moderate, we did not compute an average fidelity score for each individual group or each TA. However, across groups, the difference in the proportion of lessons rated by each rater was not significant: The first author rated 28% of the SWI lessons and 35% of the MR lessons, $\chi^2(1) = 0.66$, $p = .42$; the second author rated 43% of the SWI lessons and 38% of the MR lessons, $\chi^2(1) = 0.30$, $p = .59$; and the intern rated 29% of the SWI lessons and 27% of the MR lessons, $\chi^2(1) = 0.06$, $p = .81$. Therefore, we were able to

FIGURE 4
Interaction Between Group and Pretest Score on the Trained Reading Measure



Note. MR = Motivated Reading; SWI = Structured Word Inquiry.

calculate whether there was a significant difference in mean fidelity ratings. Sixty-two lessons were rated for the SWI group (*M* rating = 3.28, *SD* = 0.53) and 53 for the MR group (*M* rating = 3.43, *SD* = 0.47). We conducted a mixed-effects regression on these ratings with a fixed effect of group and a random effect for TA. This demonstrated that there was a nonsignificant trend for higher mean fidelity ratings for the MR group as compared with the SWI group (β = −0.16, *SE* = 0.09, *t* = 1.83, *p* = .07). Full model details are provided in Table B40 in Appendix B.

We explored this trend further by looking at responses to individual checklist items. A series of *t*-tests revealed that there were no significant differences between groups on the statements that were identical across conditions (see Table 12). The statements that were different across conditions could not be directly compared, but the mean for these criteria was lower for the SWI group than for the MR group (3.18 and 3.42, respectively; see Table 13), suggesting that the trend toward lower fidelity in the SWI group was due to difficulties in delivering the specifics of the intervention (e.g., using word sums and word matrices).

TA Fidelity Questionnaires

Nineteen of the 26 TAs (67%) completed the fidelity questionnaires. The mean ratings for each item are shown in Table 14, along with *p*-values from paired-samples *t*-tests. Overall, the TAs felt significantly less confident in delivering the SWI lessons and found the SWI lessons more challenging to deliver.

We attempted to collect attendance data from TAs, but only 42% of the data were returned (in some cases, TAs forgot to collect the data, and in other cases, they collected the data but lost the attendance sheets). Therefore, to obtain a rough measure of attendance, as part of the fidelity questionnaires, we asked the TAs to estimate what

TABLE 12
Mean Fidelity Ratings for Checklist Statements Common to Both Conditions

Statement	Motivated Reading *M*	Structured Word Inquiry *M*	*p*
Materials are prepared and organised	3.73	3.68	.63
Lesson begins promptly	3.54	3.49	.77
All children can see and hear information	3.79	3.81	.78
Concepts are covered in allotted time	3.27	3.39	.42
Lesson is paced appropriately	3.25	3.39	.35
High levels of specific praise are used	3.34	3.37	.82
All students given a chance to respond	3.48	3.42	.61
Awards Word Detective points appropriately	3.33	3.43	.53
Correct terminology used	3.50	3.48	.90
Appropriate prompting used	3.39	3.30	.47
Children are focused on task	3.36	3.15	.19
Children answer relevant questions	3.63	3.57	.66
Children ask relevant questions	3.32	3.31	.93
Children use correct terminology	3.25	3.13	.40

TABLE 13
Mean Fidelity Ratings for Statements Specific to Each Intervention

Statement	*M* rating
Motivated Reading	
Uses modelling and think alouds appropriately	3.34
Encourages children to apply background knowledge	3.59
Prompts children to provide reasoning behind responses	3.30
Avoids discussion of word structure	3.61
Encourages dialogue between children	3.11
Children provide reasoning behind responses	3.16
Children discuss learnt materials in terms of their own experiences and background knowledge	3.55
Children demonstrate interest in the words or texts	3.70
M	3.42
Structured Word Inquiry	
Uses spelling out appropriately (e.g., pronounces affixes as a unit, says "is rewritten as," pauses at morpheme boundaries, announces suffixing changes)	3.10
Demonstrates appropriate use of word matrices	3.39
Demonstrates appropriate use of word sums	3.42
Treats mistakes as learning opportunities	3.24
Follows student-generated lines of inquiry	3.44
Children use spelling out appropriately	2.79
Children use word sums appropriately	3.08
Children use word matrices appropriately	2.98
M	3.18

percentage of the lessons they were able to deliver. At four of the schools, TAs reported that they were able to deliver fewer than 75% of the lessons, with one TA estimating that she was able to deliver only 30%. This TA was at a school that had been placed under special measures during the course of the study, meaning that the school was subject to regular short-notice inspections. The remaining eight schools reported delivering more than 75% of the lessons.

Indeed, TAs reported a number of barriers to implementation that were corroborated by the observations of the research team. At a small number of the schools, there was a reduced level of support from senior management or classroom teachers. This led to situations in which some TAs did not have regularly time-tabled lessons, access to laptop or desktop computers, or regular spaces in which to conduct the lessons. At four of the schools, TAs were observed delivering lessons in noisy or crowded locations such as hallways. The implications of this are discussed in the next section.

Discussion

In this study, we set out to investigate whether SWI would be more effective than a comparison instruction, MR, for improving the reading, spelling, vocabulary, reading comprehension, and motivation to read of students with reading and spelling difficulties. In the SWI group, students received instruction in morphology and etymology, with a focus on discovering the logic of the English spelling system. In the MR group, students learned new vocabulary words and strategies for deducing meaning in the context of written texts but did not learn about word structure. Thus, we hoped to tease apart the effects of SWI over and above the effects of small-group instruction and increased exposure to word meanings and written text, and we were interested in whether SWI could be effectively delivered by paraprofessionals in typical school environments.

There was little evidence of an overall difference in effectiveness across the two programs. However, some patterns emerged that have implications for instruction and future research. Starting with the reading results, it is noteworthy that levels of improvement differed depending on students' initial levels of reading ability. On the

TABLE 14
Results of the Teacher Assistant Fidelity Survey (*n* = 19)

Statement	Motivated Reading *M*	Structured Word Inquiry *M*	*p*
I enjoyed teaching the training program.	3.37	3.16	.33
Lessons were challenging to deliver.	1.84	3.11	<.001***
I felt confident teaching the lessons.	3.53	2.47	<.001***
Training allowed me to deliver lessons effectively.	3.53	2.84	.006*
I felt that I had enough preparation time.	2.79	2.16	.048*
It was difficult to fit the lessons into available time.	2.26	2.84	.045*
Children were generally focused/on task.	3.16	3.00	.45

*p < .05. ***p < .001.

trained reading measure (reflecting students' ability to read trained words and nonwords made up of trained bases and affixes), students with relatively weaker reading abilities tended to benefit more from MR, whereas children with relatively stronger reading abilities tended to benefit more from SWI. A key reason for this may be the fact that TAs found SWI instruction much more challenging to deliver than MR. Indeed, many of the concepts taught in SWI were completely new to the TAs at the beginning of the study. This may have reduced TAs' capacity to tailor their feedback to students' ability levels, which may have disproportionately affected the weakest readers. The weakest readers in the study had very poor reading abilities and extremely limited knowledge of relevant linguistic concepts; for example, some did not know the difference between a consonant and a vowel. Therefore, it is also possible that the SWI lessons were pitched at too high a level for these students, who may have needed more basic and explicit instruction before they could benefit from the SWI instruction that we provided.

It is also worth noting that an interesting pattern was observable on the raw scores of the experimenter-designed reading assessments. Although both groups showed similarly sized raw score gains on their ability to read trained words and nonwords made up of trained morphemes, neither group showed any evidence of improvement on the untrained words (see Table 1). This supports the view that improvements were due to training (rather than maturation). It also suggests that students in the MR group may have been able to use implicit knowledge of the morphological structure of the trained words to correctly read some of the nonwords. However, the raw scores should be interpreted with caution because they do not take into account any of the covariates or interactions.

Results on the spelling tasks showed a different pattern. On the untrained spelling measure, grade 3 students in the SWI group scored higher than students in the MR group, but there was no evidence of a difference between the groups in grade 5. This trend was also present on the trained spelling measure, although it did not reach statistical significance. This pattern may be related to the fact that instruction in prefixes, suffixes, and morphological spelling rules is mandated as part of the U.K. grade 3 Spelling and Grammar curriculum, and some of the prefixes and suffixes listed appeared in our trained words (e.g., *dis-*, *mis-*, *-ly*, *-ous*). Students in the grade 3 SWI group may therefore have received a double dose of instruction in these affixes as compared with their counterparts in the MR group.

Given this possibility, it is surprising that the trend appears stronger on the untrained measure. This result may be due to differences in implicit morphological knowledge. The students in grade 5, despite being poor readers and spellers, had more experience with both spoken and written language than the students in grade 3 did, and therefore likely also had better implicit morphology knowledge. This possibility is supported by the fact that students in grade 5 had higher pretest scores on the morphologically complex spelling pretests. Thus, the added boost to morphological knowledge in the SWI condition may have been comparatively small in grade 5 but comparatively large in grade 3. This would have been less apparent on the trained items (because students in the MR condition also received instruction on these words) and more apparent on the untrained items, where the ability to apply morphological knowledge to unfamiliar words was crucial. Further research is required to explore the nature and causes of age-related differences in response to morphology instruction.

On both the trained and untrained spelling measures, there was also a statistically nonsignificant trend toward greater improvements in the SWI group by students who spoke EAL. The only other instance of different outcomes for native and EAL speakers occurred on the untrained vocabulary measure, where the EAL speakers made smaller gains on average. It is possible that EAL speakers may have had fewer opportunities for incidental learning of the untrained words than their monolingual English speaking peers. Importantly, native speakers and EAL speakers did not differ in their ability to learn the trained vocabulary words. However, all results comparing EAL and native speakers should be interpreted in light of the fact that the EAL speakers in our study had a wide range of different levels of English proficiency. For reasons beyond our control, we were unable to collect data on students' language proficiency, but future studies should aim to distinguish whether the effectiveness of SWI differs for students with different levels of experience with English.

There were no other significant results on the vocabulary measures, and there was no indication of any group differences or other significant findings on the reading comprehension task. On the motivation to read task, students' ratings increased to a greater extent in the SWI group than in the MR group, particularly in grade 3. Therefore, SWI instruction may have been more effective for increasing students' motivation to read. However, the size of the increase in ratings was very small (see Table 1). This was a self-report measure administered by the TAs during lesson time, so the measure may have lacked sensitivity, and it is difficult to determine the extent to which students' answers were motivated by the fact that they believed their instructor might see their responses. We chose to use a TA-administered self-report measure largely for practical reasons; further work using more in-depth measures of motivation to read is required.

Summary and Conclusions

On the majority of measures, we did not find evidence that SWI was superior to an alternative training condition for poor readers and spellers. We hypothesized that SWI instruction might be more likely than MR instruction to result in transfer to untrained words. There was some evidence that this may have been the case for the spelling results of students in grade 3, but there was no evidence of such transfer on the reading or vocabulary measures. These findings should be interpreted in light of the fact that our study was the first to explore the effectiveness of SWI when delivered by paraprofessionals. Although TAs were provided with lesson manuals, three days of intensive training in SWI (compared with one day of MR training), and regular coaching during the course of the study, they found SWI instruction very challenging to deliver. Data from lesson observations indicated that the TAs did not always implement specific elements of SWI instruction effectively (e.g., using word sums and word matrices). Our findings suggest that delivery of SWI instruction requires deep knowledge of morphological structure, which takes time to acquire. SWI instruction may be more effective if it is delivered by instructors with greater levels of linguistic knowledge and expertise.

There were also other implementation challenges. TAs at different schools received different amounts of support from senior management and classroom teachers. In some schools, TAs were provided with regular teaching spaces, and lessons were centrally time-tabled. In other schools, TAs had to find their own teaching spaces and negotiate with classroom teachers over time-tabling. This was often as a result of factors affecting the whole school, such as staffing changes due to reduced funding. The lack of dedicated space and time often resulted in missed lessons, suboptimal teaching spaces (e.g., noisy hallways), and lost intervention and assessment materials.

These findings speak to the importance of considering the way TAs are deployed to work with students with poor reading and spelling. Research has suggested that TAs are most effective when delivering structured intervention in one-on-one or small-group settings (Sharples et al., 2018). However, if TAs do not receive sufficient practical support from senior management to deliver this intervention, and if there is little communication among TAs, class teachers, and school management about the intervention, it is less likely to succeed (e.g., Sharples et al., 2018). Therefore, one of the important lessons from this intervention study is that there are challenges in rolling out SWI at scale using paraprofessional instructors. A priority for future research is to determine whether SWI instruction can be delivered by TAs if they receive additional training. It is also worth exploring whether the effectiveness of SWI would improve if class teachers were more closely involved in the planning and implementation of SWI instruction and if SWI activities were more integrated with classroom activities and instructional content.

Myriad studies of morphological instruction have been conducted, but few have attempted to answer specific questions about which methods of morphological instruction work for poor readers and spellers and whether these methods are effective under realistic school conditions. Our findings draw attention to the fact that we cannot assume that findings from controlled studies delivered by expert educators will generalize to other settings and other instructors. Future studies need to go beyond laboratory conditions to explore which methods work, for whom, and under what circumstances.

NOTES

This research project was funded by a Nuffield Foundation grant (EDU/42179) to Colin J. Davis and Jeffrey S. Bowers. Bowers is the brother of Peter Bowers, who runs the company WordWorks where he uses SWI to work with students, teachers, and schools. We thank the TAs, other staff, and students at participating schools; Paula Clarke for assistance with development of the MR training program; Bowers for training TAs in SWI during the workshop; the members of our advisory committee (Kate Nation, Chris Rogers, and Gemma Davies) for their invaluable advice; Samantha McCormick for assistance with randomization; Bobby Stuyvesant, Peter Humburg, and Serje Robidoux for statistical advice; Oxford University Press for generously donating books; Aardman animations for the use of the Morph character on our study materials; and our research assistants and volunteer interns, particularly Queenie Hon, Shawna Murphy, Cristiano Altamura, Alwena Luff, Laura Threadgold, Eleanor Underhill, and Rebecca Jackson.

[1] We use the term *English as an additional language* because it is used by England's Department for Education to refer to any students who speak a language other than English at home.

[2] In our initial funding application, the intention was to compare SWI with a phonics condition. However, when planning the interventions, we consulted with a phonics expert, who had concerns about delivering small-group phonics intervention to mixed-ability groups of students in grades 3 and 5, and advised that best practice at that age would be to tailor intervention to the needs of individual students. Unfortunately, this was not feasible within our research design, in which students were randomly assigned to receive intervention in small groups (i.e., neither individual instruction nor ability grouping was possible). Therefore, we developed the MR condition to ensure a stringent test of the effectiveness of SWI.

[3] We intended to collect data on students' home languages and years of exposure to English, but at the time, there was a political controversy whereby some parents objected to the collection of data on their children's country of birth in the national school census. Therefore, it was not appropriate to attempt to collect similar data for this research study. The information we have on home language and length of time speaking English is general information reported by TAs.

[4] We decided to use composite scores rather than a latent variable analysis primarily for theoretical reasons; we wished to preserve the distinction between near- and far-transfer measures for each outcome type (e.g., reading, spelling, vocabulary).

[5] We included gender as a covariate in our initial analysis models but removed it for the revised analysis at the request of a reviewer, as it did not relate to our research questions, and its inclusion had little influence on the results. See Appendix B for results of our initial analytic models.

[6] Missing data could not be estimated for these participants because estimation of missing data in hierarchical data structures requires at least some data on the relevant variables from within each cluster.

REFERENCES

Abbott, R.D., & Berninger, V.W. (1993). Structural equation modeling of relationships among developmental skills and writing skills in primary- and intermediate-grade writers. *Journal of Educational Psychology*, *85*(3), 478–508. https://doi.org/10.1037/0022-0663.85.3.478

Andrews, S., Lo, S., & Xia, V. (2017). Individual differences in automatic semantic priming. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(5), 1025–1039. https://doi.org/10.1037/xhp0000372

Andrews, S., Veldre, A., & Clarke, I.E. (2020). Measuring lexical quality: The role of spelling. *Behavior Research Methods*, *52*, 2257–2282. https://doi.org/10.3758/s13428-020-01387-3

Anglin, J.M. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development*, *58*(10). https://doi.org/10.2307/1166112

August, D., & Shanahan, T. (Eds.). (2006). *Developing literacy in second-language learners: Lessons from the report of the National Literacy Panel on Language-Minority Children and Youth*. Mahwah, NJ: Erlbaum.

August, D., & Shanahan, T. (2010). Response to a review and update on developing literacy in second- language learners: Report of the National Literacy Panel on Language Minority Children and Youth. *Journal of Literacy Research*, *42*(3), 341–348. https://doi.org/10.1080/1086296X.2010.503745

Beck, I., McKeown, M., & Kucan, L. (2013). *Bringing words to life: Robust vocabulary instruction* (2nd ed.). New York, NY: Guilford.

Beck, I., Perfetti, C.A., & McKeown, M. (1982). Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology*, *74*(4), 506–521. https://doi.org/10.1037/0022-0663.74.4.506

Berninger, V.W., Abbott, R.D., Nagy, W., & Carlisle, J. (2010). Growth in phonological, orthographic, and morphological awareness in grades 1–6. *Journal of Psycholinguistic Research*, *39*(2), 141–163. https://doi.org/10.1007/s10936-009-9130-6

Beyersmann, E., Castles, A., & Coltheart, M. (2012). Morphological processing during visual word recognition in developing readers: Evidence from masked priming. *Quarterly Journal of Experimental Psychology*, *65*(7), 1306–1326. https://doi.org/10.1080/17470218.2012.656661

Blevins, W. (2001). *Teaching phonics and word study in the intermediate grades*. New York, NY: Scholastic.

Boetsch, E.A., Green, P.A., & Pennington, B.F. (1996). Psychosocial correlates of dyslexia across the life span. *Development and Psychopathology*, *8*(3), 539–562. https://doi.org/10.1017/S0954579400007264

Boutron, I., Moher, D., Altman, D.G., Schulz, K.F., & Ravaud, P. (2008a). Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: Explanation and elaboration. *Annals of Internal Medicine*, *148*(4), 295–309. https://doi.org/10.7326/0003-4819-148-4-200802190-00008

Boutron, I., Moher, D., Altman, D.G., Schulz, K.F., & Ravaud, P. (2008b). Methods and processes of the CONSORT Group: Example of an extension for trials assessing nonpharmacologic treatments. *Annals of Internal Medicine*, *148*(4), W60–W66. https://doi.org/10.7326/0003-4819-148-4-200802190-00008-w1

Bowers, J.S., & Bowers, P.N. (2017). Beyond phonics: The case for teaching children the logic of the English spelling system. *Educational Psychologist*, *52*(2), 124–141. https://doi.org/10.1080/00461520.2017.1288571

Bowers, P.N. (2009). *Teaching how the written word works: Using morphological problem-solving to develop students' language skills and engagement with the written word*. Wolfe Island, ON, Canada: WordWorks Literacy Centre.

Bowers, P.N., & Kirby, J.R. (2010). Effects of morphological instruction on vocabulary acquisition. *Reading and Writing*, *23*, 515–537. https://doi.org/10.1007/s11145-009-9172-z

Bowers, P.N., Kirby, J.R., & Deacon, S.H. (2010). The effects of morphological instruction on literacy skills: A systematic review of the literature. *Review of Educational Research*, *80*(2), 144–179. https://doi.org/10.3102/0034654309359353

Breadmore, H.L., & Carroll, J.M. (2016). Morphological spelling in spite of phonological deficits: Evidence from children with dyslexia and otitis media. *Applied Psycholinguistics*, *37*(6), 1439–1460. https://doi.org/10.1017/S0142716416000072

Carlisle, J.F. (1987). The use of morphological knowledge in spelling derived forms by learning-disabled and normal students. *Annals of Dyslexia*, *37*(1), 90–108. https://doi.org/10.1007/BF02648061

Carlisle, J.F., & Katz, L.A. (2006). Effects of word and morpheme familiarity on reading of derived words. *Reading and Writing*, *19*, 669–693. https://doi.org/10.1007/s11145-005-5766-2

Carlisle, J.F., & Kearns, D.M. (2017). Learning to read morphologically complex words. In K. Cain, D.L. Compton, & R.K. Parrila (Eds.), *Theories of reading development* (pp. 191–214). Philadelphia, PA: John Benjamins.

Carlisle, J.F., & Stone, C.A. (2005). Exploring the role of morphemes in word reading. *Reading Research Quarterly*, *40*(4), 428–449. https://doi.org/10.1598/RRQ.40.4.3

Castles, A., Coltheart, M., Larsen, L., Jones, P., Saunders, S.J., & McArthur, G. (2009). Assessing the basic components of reading: A revision of the Castles and Coltheart Test with new norms. *Australian Journal of Learning Difficulties*, *14*(1), 67–88. https://doi.org/10.1080/19404150902783435

Chall, J.S., & Jacobs, V.A. (1983). Writing and reading in the elementary grades: Developmental trends among low SES children. *Language Arts*, *60*(5), 617–626.

Choirat, C., Honaker, J., Imai, K., King, G., & Lau, O. (2020). *Zelig: Everyone's statistical software* (Version 5.1.7) [Computer software]. Retrieved from https://zeligproject.org/

Clarke, P.J., Snowling, M.J., Truelove, E., & Hulme, C. (2010). Ameliorating children's reading-comprehension difficulties: A randomized controlled trial. *Psychological Science*, *21*(8), 1106–1116. https://doi.org/10.1177/0956797610375449

Colenbrander, D., Davis, C., Bowers, J.S., & Parsons, L. (2016). *MORPH Project: Preregistration template from AsPredicted.org*. Retrieved from https://osf.io/vpb6f

Collins, L.M., Schafer, J.L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330–351. https://doi.org/10.1037/1082-989X.6.4.330

Deacon, S.H., & Bryant, P. (2006). Getting to the root: Young writers' sensitivity to the role of root morphemes in the spelling of inflected and derived words. *Journal of Child Language*, *33*(2), 401–417. https://doi.org/10.1017/S0305000906007409

Demie, F. (2018). English as an additional language and attainment in primary schools in England. *Journal of Multilingual and Multicultural Development*, *39*(3), 210–223. https://doi.org/10.1080/01434632.2017.1348508

Department for Education. (2019). *National curriculum assessments at key stage 2 in England, 2019 (provisional)*. London, UK: Author. Retrieved from https://www.gov.uk/government/statistics/national-curriculum-assessments-key-stage-2-2019-provisional

Devonshire, V., & Fluck, M. (2010). Spelling development: Fine-tuning strategy-use and capitalising on the connections between words. *Learning and Instruction*, *20*(5), 361–371. https://doi.org/10.1016/j.learninstruc.2009.02.025

Devonshire, V., Morris, P., & Fluck, M. (2013). Spelling and reading development: The effect of teaching children multiple levels of representation in their orthography. *Learning and Instruction*, *25*, 85–94. https://doi.org/10.1016/j.learninstruc.2012.11.007

DeWalt, D.A., Berkman, N.D., Sheridan, S., Lohr, K.N., & Pignone, M.P. (2004). Literacy and health outcomes: A systematic review of the literature. *Journal of General Internal Medicine*, *19*(12), 1228–1239. https://doi.org/10.1111/j.1525-1497.2004.40153.x

Dunn, L.M., Dunn, D.M., & Styles, B. (2009). *British Picture Vocabulary Scale* (3rd ed.). London, UK: GL Assessment.

Dyson, H., Best, W., Solity, J., & Hulme, C. (2017). Training mispronunciation correction and word meanings improves children's ability to learn to read words. *Scientific Studies of Reading*, *21*(5), 392–407. https://doi.org/10.1080/10888438.2017.1315424

Enders, C.K. (2017). Multiple imputation as a flexible tool for missing data handling in clinical research. *Behaviour Research and Therapy*, *98*, 4–18. https://doi.org/10.1016/j.brat.2016.11.008

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2017). irr: Various coefficients of interrater reliability and agreement (Version 0.84) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/irr/index.html

Georgiou, G.K., Savage, R., Dunn, K., Bowers, P., & Parrila, R. (2021). Examining the effects of Structured Word Inquiry on the reading and spelling skills of persistently poor grade 3 readers. *Journal of Research in Reading*, *44*(1), 131–153. https://doi.org/10.1111/1467-9817.12325

GL Assessment. (2010). *The New Group Reading Test (NGRT)*. London, UK: Author.

Goodwin, A.P. (2016). Effectiveness of word solving: Integrating morphological problem-solving within comprehension instruction for middle school students. *Reading and Writing*, *29*, 91–116. https://doi.org/10.1007/s11145-015-9581-0

Goodwin, A.P., & Ahn, S. (2010). A meta-analysis of morphological interventions: Effects on literacy achievement of children with literacy difficulties. *Annals of Dyslexia*, *60*(2), 183–208. https://doi.org/10.1007/s11881-010-0041-x

Goodwin, A.P., & Ahn, S. (2013). A meta-analysis of morphological interventions in English: Effects on literacy outcomes for school-aged children. *Scientific Studies of Reading*, *17*(4), 257–285. https://doi.org/10.1080/10888438.2012.689791

Goodwin, A.P., Lipsky, M., & Ahn, S. (2012). Word detectives: Using units of meaning to support literacy. *The Reading Teacher*, *65*(7), 461–470. https://doi.org/10.1002/TRTR.01069

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, *45*(7). https://doi.org/10.18637/jss.v045.i07

Kirby, J.R., & Bowers, P.N. (2017). Morphological instruction and literacy: Binding phonological, orthographic, and semantic features of words. In K. Cain, D.L. Compton, & R.K. Parrila (Eds.), *Theories of reading development* (pp. 437–462). Philadelphia, PA: John Benjamins.

Knowles, J.E., & Frederick, C. (2019). merTools: Tools for analyzing mixed effect regression models (R package version 0.5.0) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/merTools/index.html

Kohnen, S., Colenbrander, D., Caruana, N., & Barisic, K. (2016). *The Morphological Spelling Test–Nonwords*. Unpublished test, Macquarie University, Australia.

Kohnen, S., Colenbrander, D., Krajenbrink, T., & Nickels, L. (2015). Assessment of lexical and non-lexical spelling in students in grades 1–7. *Australian Journal of Learning Difficulties*, *20*(1), 15–38. https://doi.org/10.1080/19404158.2015.1023209

Krippendorff, K. (2011). *Computing Krippendorff's alpha-reliability* (Working paper). Retrieved from https://repository.upenn.edu/asc_papers/43/

Kuznetsova, A., Brockhoff, P.B., & Christensen, R.H.B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lenth, R. (2019). emmeans: Estimated marginal means, aka least-squares means (Version 1.4.2) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/emmeans/index.html

Lesaux, N.K., Kieffer, M.J., Faller, S.E., & Kelley, J.G. (2010). The effectiveness and ease of implementation of an academic vocabulary intervention for linguistically diverse students in urban middle schools. *Reading Research Quarterly*, *45*(2), 196–228. https://doi.org/10.1598/RRQ.45.2.3

Lesaux, N.K., Kieffer, M.J., Kelley, J.G., & Harris, J.R. (2014). Effects of academic vocabulary instruction for linguistically diverse adolescents: Evidence from a randomized field trial. *American Educational Research Journal*, *51*(6), 1159–1194. https://doi.org/10.3102/0002831214532165

Levesque, K.C., Breadmore, H.L., & Deacon, S.H. (2021). How morphology impacts reading and spelling: Advancing the role of morphology in models of literacy development. *Journal of Research in Reading*, *44*(1), 10–26. https://doi.org/10.1111/1467-9817.12313

Levesque, K.C., Kieffer, M.J., & Deacon, S.H. (2017). Morphological awareness and reading comprehension: Examining mediating factors. *Journal of Experimental Child Psychology*, *160*, 1–20. https://doi.org/10.1016/j.jecp.2017.02.015

Luke, S.G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Malloy, J.A., Marinak, B.A., Gambrell, L.B., & Mazzoni, S.A. (2013). Assessing motivation to read: The Motivation to Read Profile-Revised. *The Reading Teacher*, *67*(4), 273–282. https://doi.org/10.1002/trtr.1215

McKeown, M., Beck, I., Omanson, R., & Perfetti, C.A. (1983). The effects of long-term vocabulary instruction on reading comprehension: A replication. *Journal of Literacy Research*, *15*(1), 3–18. https://doi.org/10.1080/10862968309547474

McLaughlin, M.J., Speirs, K.E., & Shenassa, E.D. (2014). Reading disability and adult attained education and income: Evidence from a 30-year longitudinal study of a population-based sample. *Journal of Learning Disabilities*, *47*(4), 374–386. https://doi.org/10.1177/0022219412458323

Murphy, V.A., & Unthiah, A. (2015). *A systematic review of intervention research examining English language and literacy development in children with English as an additional language (EAL)*. London, UK: Education Endowment Foundation.

Nagy, W.E., & Anderson, R.C. (1984). How many words are there in printed school English? *Reading Research Quarterly*, *19*(3), 304–330. https://doi.org/10.2307/747823

OECD. (2013). *OECD skills outlook 2013: First results from the Survey of Adult Skills* (Rev. ed.). Paris, France: Author.

Office for National Statistics. (2020). *Academic year 2019/2020: Schools, pupils and their characteristics*. Retrieved from https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics#

Palincsar, A.S., & Brown, A.L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, *1*(2), 117–175. https://doi.org/10.1207/s1532690xci0102_1

Perfetti, C.A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357–383. https://doi.org/10.1080/10888430701530730

Proctor, C.P., Silverman, R.D., Harring, J.R., Jones, R.L., & Hartranft, A.M. (2020). Teaching bilingual learners: Effects of a language-based reading intervention on academic language and reading comprehension in grades 4 and 5. *Reading Research Quarterly*, *55*(1), 95–122. https://doi.org/10.1002/rrq.258

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Schulz, K.F., Altman, D.G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMJ*, *340*, Article c332. https://doi.org/10.1136/bmj.c332

Share, D.L. (1995). Phonological recoding and self-teaching: *Sine qua non* of reading acquisition. *Cognition*, *55*(2), 151–218. https://doi.org/10.1016/0010-0277(94)00645-2

Share, D.L. (2011). On the role of phonology in reading acquisition: The self-teaching hypothesis. In S.A. Brady, D. Braze, & C.A. Fowler (Eds.), *Explaining individual differences in reading: Theory and evidence* (pp. 45–68). New York, NY: Psychology.

Sharples, J., Webster, R., & Blatchford, P. (2018). *Making best use of teaching assistants: Guidance report*. London, UK: Education Endowment Foundation.

Singson, M., Mahony, D., & Mann, V. (2000). The relation between reading ability and morphological skills: Evidence from derivational suffixes. *Reading and Writing*, *12*, 219–252. https://doi.org/10.1023/A:1008196330239

Skipp, A., & Hopwood, V. (2019). *Deployment of teaching assistants in schools: Research report*. London, UK: Department for Education.

Sumner, E., Connelly, V., & Barnett, A.L. (2014). The influence of spelling ability on handwriting production: Children with and without dyslexia. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1441–1447. https://doi.org/10.1037/a0035785

Torgesen, J.K., Wagner, R.K., & Rashotte, C.A. (1999). *Test of Word Reading Efficiency*. Austin, TX: Pro-Ed.

Treiman, R., & Bourassa, D. (2000). The development of spelling skill. *Topics in Language Disorders*, *20*(3), 1–18. https://doi.org/10.1097/00011363-200020030-00004

Treiman, R., & Cassar, M. (1996). Effects of morphology on children's spelling of final consonant clusters. *Journal of Experimental Child Psychology*, *63*(1), 141–170. https://doi.org/10.1006/jecp.1996.0045

Tyler, A., & Nagy, W.E. (1989). The acquisition of English derivational morphology. *Journal of Memory and Language*, *28*(6), 649–667. https://doi.org/10.1016/0749-596X(89)90002-8

van Heuven, W.J.B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Venezky, R. (1999). *The American way of spelling: The structure and origins of American English orthography*. New York, NY: Guilford.

Verhoeven, L., & Perfetti, C.A. (2011). Morphological processing in reading acquisition: A cross-linguistic perspective. *Applied Psycholinguistics*, *32*(3), 457–466. https://doi.org/10.1017/S0142716411000154

von Hippel, P.T. (2007). Regression with missing Ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology*, *37*(1), 83–117. https://doi.org/10.1111/j.1467-9531.2007.00180.x

Wanzek, J., Wexler, J., Vaughn, S., & Ciullo, S. (2010). Reading interventions for struggling readers in the upper elementary grades: A synthesis of 20 years of research. *Reading and Writing*, *23*, 889–912. https://doi.org/10.1007/s11145-009-9179-5

Submitted May 8, 2020
Final revision received February 16, 2021
Accepted February 22, 2021

DANIELLE COLENBRANDER (corresponding author) is a research fellow in the Department of Cognitive Science and the Macquarie University Centre for Reading at Macquarie University, Sydney, Australia, and an honorary research associate in the School of Psychological Science at the University of Bristol, UK; email danielle.colenbrander@mq.edu.au.

LIAM PARSONS is an educational psychologist at the Essex County Council, UK; email liam.parsons@essex.gov.uk.

JEFFREY S. BOWERS is a professor in the School of Psychological Science at the University of Bristol, UK; email j.bowers@bristol.ac.uk.

COLIN J. DAVIS is a professor and the Chair in Cognitive Psychology in the School of Psychological Science at the University of Bristol, UK; email pscjd@bristol.ac.uk.

APPENDIX A

Trained and Untrained Words

Trained words				Untrained words			
Word	Length	Frequency	Part of speech	Word	Length	Frequency	Part of speech
abrupt	6	3.04	Adjective	*agreement*	9	4.61	Noun
assistance	10	4.01	Noun	*deflection*	10	3.37	Noun
commit	6	4.12	Verb	*deployment*	10	3.33	Noun
counteract	10	2.83	Verb	*destroyer*	9	3.04	Noun
decline	7	4.14	Noun	*developer*	9	3.83	Noun
destructive	11	3.58	Adjective	*dislike*	7	3.52	Verb
diction	7	2.65	Noun	*distinction*	11	3.73	Noun
dispense	8	2.86	Verb	*implosion*	9	2.29	Noun
displeasure	11	2.56	Noun	*joyful*	6	3.27	Adjective
equitable	9	3.10	Adjective	*laziness*	8	2.84	Noun
happily	7	4.11	Adverb	*misplace*	8	1.65	Verb
happiness	9	4.14	Noun	*misuse*	6	3.28	Noun

(*continued*)

Trained and Untrained Words (*continued*)

Trained words				Untrained words			
Word	Length	Frequency	Part of speech	Word	Length	Frequency	Part of speech
helpful	7	4.21	Adjective	*percussion*	10	3.36	Noun
helpless	8	3.49	Adjective	*prepay*	6	1.39	Verb
inference	9	2.51	Noun	*prettiness*	10	2.19	Noun
interact	8	3.52	Verb	*protective*	10	3.83	Adjective
misfit	6	2.47	Noun	*repayment*	9	3.08	Noun
moveable	8	2.58	Adjective	*retold*	6	2.26	Verb
movement	8	4.71	Noun	*substation*	10	2.31	Noun
preheat	7	1.95	Verb	*unstoppable*	11	3.59	Adjective
projection	10	3.26	Noun	*unthinkable*	11	3.44	Adjective
recurrence	10	2.45	Noun	*useless*	7	4.10	Adjective
reheat	6	2.60	Verb				
retract	7	2.73	Verb				
sensory	7	3.16	Adjective				
unfit	5	3.29	Adjective				
unify	5	2.63	Verb				
unpleasant	10	3.78	Adjective				
Mean	7.93	3.23		Mean	8.73	3.11	
Range	5–11	1.95–4.71		Range	6–11	1.39–4.61	

Note. Length, $t(48) = 0.59$, $p = .11$, and frequency, $t(48) = 1.62$, $p = .56$, were not significantly different across word types.

APPENDIX B

Notes on Analysis

In our preregistration document, we specified that we would conduct an intention-to-treat analysis on each of our outcome measures (primary and secondary), including all 250 participants who had been randomized to a group, even if they later dropped out of the study. However, the majority of our missing data were on our dependent variables, and the imputation of data from dependent variables can be problematic, as it may add noise to the estimates (e.g., von Hippel, 2007). We also stated in our preregistration that we would conduct a sensitivity analysis using data from participants who attended at least 75% of the training sessions, to explore the effectiveness of training when delivered at the intended dosage. However, TAs returned attendance data for only 42% of the lessons.

Therefore, the analyses we initially planned were not feasible, so we conducted a complete-case analysis including all participants who had been both pre- and posttested (regardless of the level of attendance). We reduced the total number of analytic models by creating composite scores for trained and untrained reading, spelling, and vocabulary. Additionally, although we prespecified that we would include gender as a covariate in the models, a reviewer requested that we run the analyses without this covariate because it did not seem to contribute to the models and did not relate directly to our research questions. The results of this adapted analysis appear in the main body of this article. However, for the sake of transparency and completeness, we report the results of our initially planned intention-to-treat analysis here, as well as a complete-case analysis for each outcome measure with all students who completed the training programs. The results of all analyses are substantially similar. The code for these analyses is available on the Open Science Framework (https://osf.

io/dx39t/?view_only=522debdd4b9442a1a6fd3ce3dcc2a947).

Intention-to-Treat Analysis

We used multiple imputation to impute missing values for the intention-to-treat analysis. Although we initially specified that we would conduct full information maximum likelihood analyses to account for missing data, multiple imputation is more flexible when it comes to dealing with complex analytical scenarios, such as when a regression analysis has a mixture of categorical and continuous variables (Enders, 2017). Multiple imputation is more straightforward to implement in the R software environment than full information maximum likelihood, and there is evidence to suggest that the two methods tend to result in similar estimates and standard errors (Collins, Schafer, & Kam, 2001). Therefore, we used multiple imputation for the final intention-to-treat analysis.

We conducted our intention-to-treat analysis using the Amelia II (Honaker, King, & Blackwell, 2011) and merTools packages (Knowles & Frederick, 2019) in the R software environment (R Core Team, 2019). Data were centered and scaled before imputation. The merTools package does not return *p*-values for multilevel analyses, and there is also currently no straightforward way to pool imputation results for post hoc testing, so we conducted post hoc testing only in the complete-case analysis.

Missing Data

During the course of the study, 33 participants withdrew. First, an entire school withdrew from the study after one term, resulting in the loss of 20 participants. This did not bias the randomization, as randomization was stratified within schools. Therefore, all analyses are conducted on the remaining sample of 250 participants.[6]

Of these 250 participants, four were removed from the study at the request of their schools due to challenging behavior. Although these students did not take part in training, we were able to posttest three of them on at least some of the posttest measures (all were from the MR group). One student was removed from the study by his parents, who felt that he did not need the intervention. A further eight students changed schools over the course of the year, and we were not able to posttest them at their new schools. A small amount of additional pre- and posttest data was missing due to absences on the day of testing or accidental data loss. There were no missing data for any of the demographic variables (school year, gender, EAL, or eligibility for free school meals). In total, we collected posttest data for 240 participants, and 3% of the data was missing. The percentage of data missing for each outcome variable (by group) is shown in Table B1, and demographic data for the total sample of 250 participants (the intention-to-treat sample) and the attrited sample of 240 participants (the complete-case sample) are shown in Table 1 in the main body of this article.

TABLE B1
Proportion of Missing Data for Each Variable

	Percentage missing: Structured Word Inquiry		Percentage missing: Motivated Reading	
Variable or outcome measure	Pretest	Posttest	Pretest	Posttest
Experimenter-designed reading aloud	0	3.2	0	4.8
Experimenter-designed spelling	0	3.2	0	4.8
Experimenter-designed vocabulary	0	4.0	0.8	5.6
TOWRE Phonetic Decoding Efficiency	0	2.4	0	4.8
TOWRE Sight Word Efficiency	1.6	3.2	0.8	4.8
CC2–Nonword Reading	0.8	5.6	0	6.4
CC2–Irregular Word Reading	0.8	3.2	0	4.0
Diagnostic Spelling Test–Nonwords	0	3.2	0	5.6
Diagnostic Spelling Test–Irregular Words	0.8	4.8	0	5.6
Morphological Spelling Test–Nonwords	5.6	10.4	9.6	10.4
Multiple-choice vocabulary	0	3.2	0.8	6.4
New Group Reading Test	0.8	3.2	0	6.4
Motivation to read	1.6	12.0	0.8	15.2

Note. CC2 = Castles and Coltheart Test 2; TOWRE = Test of Word Reading Efficiency.

Results of Analytic Models

Morphological Reading Task

TABLE B2
Trained Words: Complete-Case Analysis

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	−0.04	0.13	−0.32	.75
Gender (female)	−0.06	0.09	−0.67	.50
Eligibility for free school meals (eligible)	0.04	0.10	0.41	.68
Group (Motivated Reading)	−0.12	0.16	−0.74	.46
Pretest score	0.83	0.08	10.79	<.001***
Grade level (5)	−0.01	0.14	−0.06	.95
EAL (EAL speakers)	0.06	0.13	0.44	.66
Group × Pretest interaction	−0.25	0.11	−2.42	.02*
Group × Grade interaction	0.25	0.20	1.23	.22
Group × EAL interaction	0.06	0.18	0.35	.72
Random effect	**Variance**		***SD***	
Intercept (school)	0.03		0.16	
Residual	0.46		0.68	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	547.36		.52	.54

Note. EAL = English as an additional language.
*p < .05. ***p < .001.

TABLE B3
Trained Words: Intention-to-Treat Analysis[a]

Fixed effect	Estimate	*SE*	*t*
Intercept	−0.05	0.13	−0.36
Gender (female)	−0.07	0.09	−0.83
Eligibility for free school meals (eligible)	0.04	0.10	0.40
Group (Motivated Reading)	−0.12	0.16	−0.76
Pretest score	0.81	0.07	10.90*
Grade level (5)	0.03	0.14	0.19
EAL (EAL speakers)	0.03	0.13	0.23
Group × Pretest interaction	−0.24	0.10	−2.29*
Group × Grade interaction	0.03	0.14	0.19
Group × EAL interaction	0.10	0.18	0.56
Random effect	***SD***		
Intercept (school)	0.16		
Residual	0.68		
Model fit			
Akaike information criterion	571.0		

Note. EAL = English as an additional language.
[a]We used the merTools package in R (Knowles & Frederick, 2019) for the intention-to-treat analyses. In this package, *p*-values are not provided. See the Intention-to-Treat Analysis subsection in this appendix. As an approximate measure of statistical significance, *t*-values greater than 2 can be considered statistically significant (e.g., Andrews, Lo, & Xia, 2017; Luke, 2017).
*p < .05.

Morphological Reading Task

TABLE B4
Nonwords: Complete-Case Analysis

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	0.05	0.14	0.35	.73
Gender (female)	0.04	0.10	0.35	.72
Eligibility for free school meals (eligible)	−0.06	0.11	−0.55	.59
Group (Motivated Reading)	−0.13	0.18	−0.73	.47
Pretest score	0.73	0.09	8.48	<.001***
Grade level (5)	−0.10	0.16	−0.64	.52
EAL (EAL speakers)	0.04	0.15	0.24	.81
Group × Pretest interaction	−0.23	0.12	−1.91	.06
Group × Grade interaction	0.19	0.24	0.82	.41
Group × EAL interaction	−0.02	0.21	−0.09	.93
Random effect	**Variance**		***SD***	
Intercept (school)	0.02		0.14	
Residual	0.62		0.79	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	610.83		.37	.39

Note. EAL = English as an additional language.
$^{***}p < .001$.

TABLE B5
Nonwords: Intention-to-Treat Analysis

Fixed effect	Estimate	*SE*	*t*
Intercept	0.04	0.14	0.28
Gender (female)	0.03	0.10	0.28
Eligibility for free school meals (eligible)	−0.05	0.11	−0.48
Group (Motivated Reading)	−0.12	0.18	−0.67
Pretest score	0.74	0.09	8.62*
Grade level (5)	−0.07	0.16	−0.45
EAL (EAL speakers)	0.02	0.15	0.11
Group × Pretest interaction	−0.24	0.12	−1.98
Group × Grade interaction	0.18	0.24	0.75
Group × EAL interaction	0.00	0.21	0.00
Random effect	***SD***		
Intercept (school)	0.14		
Residual	0.79		
Model fit			
Akaike information criterion	635.6		

Note. EAL = English as an additional language.
$^{*}p < .05$.

Morphological Reading Task

TABLE B6
Untrained Words: Complete-Case Analysis

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	−0.11	0.12	−0.93	.35
Gender (female)	−0.19	0.09	−2.17	.03*
Eligibility for free school meals (eligible)	−0.16	0.09	−1.78	.08
Group (Motivated Reading)	0.08	0.15	0.53	.59
Pretest score	0.71	0.07	10.31	<.001***
Grade level (5)	0.26	0.13	2.02	.04*
EAL (EAL speakers)	0.26	0.13	2.09	.04*
Group × Pretest interaction	−0.06	0.10	−0.65	.52
Group × Grade interaction	0.07	0.19	0.36	.72
Group × EAL interaction	−0.29	0.17	−1.66	.10
Random effect	**Variance**		***SD***	
Intercept (school)	0.01		0.10	
Residual	0.42		0.65	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	521.02		.58	.59

Note. EAL = English as an additional language.
$^{*}p < .05$. $^{***}p < .001$.

TABLE B7
Untrained Words: Intention-to-Treat Analysis

Fixed effect	Estimate	*SE*	*t*
Intercept	−0.10	0.12	−0.86
Gender (female)	−0.22	0.08	−2.62*
Eligibility for free school meals (eligible)	−0.17	0.09	−1.84
Group (Motivated Reading)	0.06	0.15	0.38
Pretest score	0.71	0.07	10.56*
Grade level (5)	0.30	0.13	2.30*
EAL (EAL speakers)	0.27	0.13	2.13*
Group × Pretest interaction	−0.08	0.10	−0.88
Group × Grade interaction	0.07	0.19	0.35
Group × EAL interaction	−0.27	0.17	−1.59
Random effect	***SD***		
Intercept (school)	0.10		
Residual	0.65		
Model fit			
Akaike information criterion	543.1		

Note. EAL = English as an additional language.
$^{*}p < .05$.

TOWRE Phonetic Decoding Efficiency Task

TABLE B8
Complete-Case Analysis

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	0.21	0.09	2.26	.03*
Gender (female)	−0.15	0.07	−2.19	.03*
Eligibility for free school meals (eligible)	−0.05	0.08	−0.73	.47
Group (Motivated Reading)	−0.18	0.12	−1.50	.13
Pretest score	0.85	0.05	15.63	<.001***
Grade level (5)	−0.17	0.10	−1.68	.09
EAL (EAL speakers)	0.01	0.10	0.08	.93
Group × Pretest interaction	0.02	0.08	0.29	.78
Group × Grade interaction	0.23	0.15	1.49	.14
Group × EAL interaction	−0.08	0.14	−0.57	.57
Random effect	**Variance**		***SD***	
Intercept (school)	0.01		0.08	
Residual	0.28		0.53	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	428.34		.72	.73

Note. EAL = English as an additional language.
**p* < .05. ****p* < .001.

TABLE B9
Intention-to-Treat Analysis

Fixed effect	Estimate	*SE*	*t*
Intercept	0.22	0.09	2.39*
Gender (female)	−0.17	0.07	−2.50*
Eligibility for free school meals (eligible)	−0.07	0.08	−0.92
Group (Motivated Reading)	−0.20	0.12	−1.72
Pretest score	0.84	0.05	15.70*
Grade level (5)	−0.15	0.10	−1.48
EAL (EAL speakers)	−0.001	0.10	−0.01
Group × Pretest interaction	0.02	0.08	0.21
Group × Grade interaction	0.23	0.15	1.53
Group × EAL interaction	−0.05	0.14	−0.36
Random effect	***SD***		
Intercept (school)	0.09		
Residual	0.53		
Model fit			
Akaike information criterion	448.5		

Note. EAL = English as an additional language.
*p < .05.

TOWRE Sight Word Efficiency Task

TABLE B10
Complete-Case Analysis

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	0.05	0.09	0.55	.58
Gender (female)	−0.10	0.07	−1.47	.14
Eligibility for free school meals (eligible)	−0.05	0.08	−0.65	.52
Group (Motivated Reading)	−0.05	0.12	−0.43	.67
Pretest score	0.80	0.06	14.24	<.001***
Grade level (5)	−0.02	0.11	−0.17	.86
EAL (EAL speakers)	0.02	0.10	0.23	.82
Group × Pretest interaction	0.07	0.08	0.80	.42
Group × Grade interaction	0.10	0.17	0.60	.55
Group × EAL interaction	−0.04	0.14	−0.30	.77
Model fit	**Multiple R^2**		**Adjusted R^2**	
	.72		.71	

Note. EAL = English as an additional language. Results are from a regression model without random effects, as the fit for the model with random intercept for school was singular.
$^{***}p < .001$.

TABLE B11
Intention-to-Treat Analysis

Fixed effect	Estimate	*SE*	*t*
Intercept	0.05	0.09	0.49
Gender (female)	−0.12	0.07	−1.70
Eligibility for free school meals (eligible)	−0.03	0.08	−0.45
Group (Motivated Reading)	−0.04	0.12	−0.34
Pretest score	0.79	0.06	14.20*
Grade level (5)	0.02	0.11	0.22
EAL (EAL speakers)	0.01	0.10	0.06
Group × Pretest interaction	0.06	0.08	0.77
Group × Grade interaction	0.08	0.16	0.47
Group × EAL interaction	−0.03	0.14	−0.21
Random effect	***SD***		
Intercept (school)	0.04		
Residual	0.55		
Model fit			
Akaike information criterion	455.6		

Note. EAL = English as an additional language.
$^{*}p < .05$.

Castles and Coltheart Test 2—Nonword Reading (Castles et al., 2009)

TABLE B12
Complete-Case Analysis

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	−0.10	0.11	−0.97	.33
Gender (female)	−0.09	0.08	−1.04	.30
Eligibility for free school meals (eligible)	−0.07	0.09	−0.78	.44
Group (Motivated Reading)	0.04	0.14	0.32	.75
Pretest score	0.77	0.07	11.82	<.001***
Grade level (5)	0.16	0.12	1.33	.19
EAL (EAL speakers)	0.08	0.12	0.66	.51
Group × Pretest interaction	−0.03	0.09	−0.38	.70
Group × Grade interaction	0.09	0.18	0.50	.62
Group × EAL interaction	−0.14	0.17	−0.84	.40
Model fit	**Multiple R^2**		**Adjusted R^2**	
	.61		.59	

Note. EAL = English as an additional language. Results are from a regression model without random effects, as the fit for the model with random intercept for school was singular.
***p < .001.

TABLE B13
Intention-to-Treat Analysis

Fixed effect	Estimate	*SE*	*t*
Intercept	−0.08	0.11	−0.73
Gender (female)	−0.12	0.09	−1.41
Eligibility for free school meals (eligible)	−0.09	0.09	−0.95
Group (Motivated Reading)	0.03	0.14	0.23
Pretest score	0.76	0.06	11.99*
Grade level (5)	0.18	0.12	1.50
EAL (EAL speakers)	0.08	0.12	0.66
Group × Pretest interaction	−0.03	0.09	−0.38
Group × Grade interaction	0.08	0.18	0.43
Group × EAL interaction	−0.14	0.17	−0.80

Note. EAL = English as an additional language. Results are from a regression model without random effects, as the fit for the model with a random intercept for school was singular. The single-level regression analysis was conducted using the Zelig package in R (Choirat, Honaker, Imai, King, & Lau, 2020).
*p < .05.

Castles and Coltheart Test 2—Irregular Word Reading (Castles et al., 2009)

TABLE B14
Complete-Case Analysis

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	0.13	0.12	1.11	.27
Gender (female)	0.02	0.08	0.21	.84
Eligibility for free school meals (eligible)	−0.11	0.09	−1.29	.20
Group (Motivated Reading)	−0.11	0.14	−0.75	.45
Pretest score	0.77	0.07	11.41	<.001***
Grade level (5)	−0.05	0.13	−0.37	.71
EAL (EAL speakers)	−0.11	0.12	−0.93	.36
Group × Pretest interaction	−0.01	0.10	−0.05	.96
Group × Grade interaction	0.18	0.19	0.97	.33
Group × EAL interaction	−0.10	0.16	−0.61	.54
Random effect	**Variance**		***SD***	
Intercept (school)	0.02		0.15	
Residual	0.38		0.62	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	504.13		.60	.62

Note. EAL = English as an additional language.
$^{***}p < .001$.

TABLE B15
Intention-to-Treat Analysis

Fixed effect	Estimate	*SE*	*t*
Intercept	0.13	0.12	1.07
Gender (female)	0.01	0.08	0.15
Eligibility for free school meals (eligible)	−0.11	0.09	−1.22
Group (Motivated Reading)	−0.10	0.14	−0.74
Pretest score	0.76	0.07	11.57*
Grade level (5)	−0.04	0.13	−0.28
EAL (EAL speakers)	−0.13	0.12	−1.06
Group × Pretest interaction	0.01	0.09	0.06
Group × Grade interaction	0.17	0.19	0.94
Group × EAL interaction	−0.09	0.16	−0.56
Random effect	***SD***		
Intercept (school)	0.15		
Residual	0.62		
Model fit			
Akaike information criterion	523.0		

Note. EAL = English as an additional language.
$^{*}p < .05$.

Morphological Spelling Task

TABLE B16
Trained Words: Complete-Case Analysis

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	−0.01	0.13	−0.05	.96
Gender (female)	0.01	0.09	0.07	.95
Eligibility for free school meals (eligible)	−0.14	0.10	−1.37	.17
Group (Motivated Reading)	−0.17	0.16	−1.07	.29
Pretest score	0.61	0.09	7.03	<.001***
Grade level (5)	0.25	0.14	1.80	.07
EAL (EAL speakers)	−0.13	0.14	−0.91	.36
Group × Pretest interaction	−0.03	0.11	−0.27	.78
Group × Grade interaction	0.38	0.20	1.88	.06
Group × EAL interaction	−0.19	0.19	−1.01	.31
Random effect	**Variance**		***SD***	
Intercept (school)	0.01		0.11	
Residual	0.52		0.72	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	568.43		.48	.50

Note. EAL = English as an additional language.
$^{***}p$ < .001.

TABLE B17
Trained Words: Intention-to-Treat Analysis

Fixed effect	Estimate	*SE*	*t*
Intercept	−0.03	0.13	−0.20
Gender (female)	0.01	0.09	0.07
Eligibility for free school meals (eligible)	−0.15	0.10	−1.48
Group (Motivated Reading)	−0.17	0.16	−1.12
Pretest score	0.57	0.08	7.03*
Grade level (5)	0.26	0.14	1.92
EAL (EAL speakers)	−0.14	0.14	−1.04
Group × Pretest interaction	0.01	0.10	0.09
Group × Grade interaction	0.38	0.20	1.92
Group × EAL interaction	−0.16	0.19	−0.86
Random effect	***SD***		
Intercept (school)	0.13		
Residual	0.72		
Model fit			
Akaike information criterion	590.8		

Note. EAL = English as an additional language.
$^{*}p$ < .05.

Morphological Spelling Task

TABLE B18
Nonwords: Complete-Case Analysis

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	−0.11	0.15	−0.74	.46
Gender (female)	0.01	0.11	0.06	.95
Eligibility for free school meals (eligible)	−0.07	0.12	−0.61	.54
Group (Motivated Reading)	−0.10	0.19	−0.56	.58
Pretest score	0.30	0.09	3.48	<.001***
Grade level (5)	0.22	0.17	1.32	.19
EAL (EAL speakers)	0.16	0.16	0.99	.32
Group × Pretest interaction	0.20	0.12	1.69	.09
Group × Grade interaction	0.36	0.24	1.55	.12
Group × EAL interaction	−0.45	0.22	−2.02	.04*
Random effect	**Variance**		***SD***	
Intercept (school)	0.02		0.13	
Residual	0.71		0.85	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	576.77		.46	.49

Note. SE = Standard Error. FSM = Free school meals. EAL = English as an additional language.
$^{*}p < .05$. $^{***}p < .001$.

TABLE B19
Nonwords: Intention-to-Treat Analysis

Fixed effect	Estimate	*SE*	*t*
Intercept	−0.12	0.15	−0.79
Gender (female)	−0.01	0.11	−0.12
Eligibility for free school meals (eligible)	−0.08	0.12	−0.66
Group (Motivated Reading)	−0.14	0.18	−0.76
Pretest score	0.31	0.09	3.59*
Grade level (5)	0.22	0.16	1.34
EAL (EAL speakers)	0.17	0.16	1.06
Group × Pretest interaction	0.17	0.12	1.42
Group × Grade interaction	0.39	0.23	1.67
Group × EAL interaction	−0.42	0.22	−1.92
Random effect	***SD***		
Intercept (school)	0.13		
Residual	0.84		
Model fit			
Akaike information criterion	668.9		

Note. EAL = English as an additional language.
$^{*}p < .05$.

Morphological Spelling Task

TABLE B20
Untrained Words: Complete-Case Analysis

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	−0.06	0.14	−0.43	.67
Gender (female)	0.02	0.10	0.21	.83
Eligibility for free school meals (eligible)	−0.12	0.10	−1.10	.27
Group (Motivated Reading)	−0.12	0.16	−0.76	.45
Pretest score	0.64	0.08	7.78	<.001***
Grade level (5)	0.12	0.14	0.84	.4
EAL (EAL speakers)	0.10	0.14	0.71	.48
Group × Pretest interaction	−0.06	0.11	−0.61	.54
Group × Grade interaction	0.45	0.20	2.23	.03*
Group × EAL interaction	−0.32	0.19	−1.64	.10
Random effect	**Variance**		***SD***	
Intercept (school)	0.03		0.18	
Residual	0.52		0.72	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	576.77		.46	.49

Note. EAL = English as an additional language.
*p < .05. ***p < .001.

TABLE B21
Untrained Words: Intention-to-Treat Analysis

Fixed effect	Estimate	*SE*	*t*
Intercept	−0.09	0.13	−0.65
Gender (female)	0.03	0.09	0.27
Eligibility for free school meals (eligible)	−0.10	0.10	−0.95
Group (Motivated Reading)	−0.13	0.16	−0.83
Pretest score	0.62	0.08	7.78*
Grade level (5)	0.11	0.14	0.82
EAL (EAL speakers)	0.11	0.14	0.76
Group × Pretest interaction	−0.05	0.10	−0.46
Group × Grade interaction	0.48	0.20	2.42*
Group × EAL interaction	−0.30	0.19	−1.60
Random effect	***SD***		
Intercept (school)	0.18		
Residual	0.72		
Model fit			
Akaike information criterion	599.3		

Note. EAL = English as an additional language.
*p < .05.

MoSTn

TABLE B22
Complete-Case Analysis

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	0.03	0.13	0.22	.82
Gender (female)	0.09	0.10	0.91	.36
Eligibility for free school meals (eligible)	−0.27	0.11	−2.46	.01*
Group (Motivated Reading)	−0.10	0.16	−0.62	.53
Pretest score	0.58	0.08	7.50	<.001***
Grade level (5)	0.13	0.14	0.89	.38
EAL (EAL speakers)	−0.09	0.15	−0.63	.53
Group × Pretest interaction	−0.02	0.11	−0.20	.84
Group × Grade interaction	0.40	0.21	1.90	.06
Group × EAL interaction	−0.24	0.21	−1.14	.25
Random effect	**Variance**		***SD***	
Intercept (school)	0.01		0.10	
Residual	0.51		0.71	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	494.73		.49	.50

Note. EAL = English as an additional language.
***p < .001.

TABLE B23
Intention-to-Treat Analysis

Fixed effect	Estimate	*SE*	*t*
Intercept	0.01	0.13	0.10
Gender (female)	0.09	0.10	0.89
Eligibility for free school meals (eligible)	−0.34	0.10	−3.28*
Group (Motivated Reading)	−0.06	0.16	−0.37
Pretest score	0.55	0.07	7.43*
Grade level (5)	0.23	0.14	1.67
EAL (EAL speakers)	−0.10	0.14	−0.73
Group × Pretest interaction	0.02	0.10	0.15
Group × Grade interaction	0.29	0.20	1.45
Group × EAL interaction	−0.31	0.20	−1.60
Random effect	***SD***		
Intercept (school)	0.12		
Residual	0.73		
Model fit			
Akaike information criterion	597.5		

Note. EAL = English as an additional language.
*p < .05.

Diagnostic Spelling Test—Nonwords

TABLE B24
Complete-Case Analysis

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	−0.17	0.16	−1.06	.29
Gender (female)	0.07	0.11	0.69	.49
Eligibility for free school meals (eligible)	−0.23	0.12	−1.95	.05
Group (Motivated Reading)	0.04	0.18	0.20	.84
Pretest score	0.44	0.09	5.07	<.001***
Grade level (5)	0.21	0.15	1.38	.17
EAL (EAL speakers)	0.001	0.17	0.01	.995
Group × Pretest interaction	−0.20	0.11	−1.78	.08
Group × Grade interaction	0.49	0.22	2.23	.03*
Group × EAL interaction	−0.34	0.22	−1.55	.12
Random effect	**Variance**		***SD***	
Intercept (school)	0.09		0.30	
Residual	0.66		0.81	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	631.77		.26	.35

Note. EAL = English as an additional language.
*p < .05. ***p < .001.

TABLE B25
Intention-to-Treat Analysis

Fixed effect	Estimate	*SE*	*t*
Intercept	−0.20	0.16	−1.24
Gender (female)	0.08	0.11	0.71
Eligibility for free school meals (eligible)	−0.26	0.12	−2.23*
Group (Motivated Reading)	0.06	0.17	0.35
Pretest score	0.43	0.09	5.07*
Grade level (5)	0.25	0.15	1.66
EAL (EAL speakers)	0.02	0.16	0.15
Group × Pretest interaction	−0.18	0.11	−1.57
Group × Grade interaction	0.47	0.22	2.17*
Group × EAL interaction	−0.39	0.21	−1.80
Random effect	***SD***		
Intercept (school)	0.31		
Residual	0.81		
Model fit			
Akaike information criterion	660.0		

Note. EAL = English as an additional language.
*p < .05.

Diagnostic Spelling Test—Irregular Words

TABLE B26
Complete-Case Analysis

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	−0.15	0.10	−1.46	.15
Gender (female)	0.03	0.07	0.44	.66
Eligibility for free school meals (eligible)	−0.20	0.08	−2.60	.01*
Group (Motivated Reading)	0.19	0.13	1.48	.14
Pretest score	0.78	0.06	12.71	<.001***
Grade level (5)	0.19	0.11	1.77	.08
EAL (EAL speakers)	0.20	0.11	1.80	.07
Group × Pretest interaction	0.03	0.09	0.36	.72
Group × Grade interaction	−0.19	0.17	−1.15	.25
Group × EAL interaction	−0.28	0.15	−1.90	.06
Random effect	**Variance**		***SD***	
Intercept (school)	0.02		0.14	
Residual	0.29		0.54	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	436.03		.69	.71

Note. EAL = English as an additional language.
*p < .05. ***p < .001.

TABLE B27
Intention-to-Treat Analysis

Fixed effect	Estimate	*SE*	*t*
Intercept	−0.17	0.11	−1.64
Gender (female)	0.03	0.07	0.35
Eligibility for free school meals (eligible)	−0.23	0.08	−2.96*
Group (Motivated Reading)	0.22	0.13	1.74
Pretest score	0.75	0.06	12.46*
Grade level (5)	0.26	0.11	2.36*
EAL (EAL speakers)	0.24	0.11	2.20*
Group × Pretest interaction	0.06	0.09	0.72
Group × Grade interaction	−0.25	0.17	−1.50
Group × EAL interaction	−0.33	0.15	−2.25*
Random effect	***SD***		
Intercept (school)	0.15		
Residual	0.55		
Model fit			
Akaike information criterion	467.6		

Note. EAL = English as an additional language.
*p < .05.

Morphological Vocabulary Task

TABLE B28
Trained Words: Complete-Case Analysis

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	−0.13	0.15	−0.85	.40
Gender (female)	−0.19	0.11	−1.69	.09
Eligibility for free school meals (eligible)	−0.09	0.12	−0.75	.45
Group (Motivated Reading)	0.06	0.18	0.30	.76
Pretest score	0.43	0.08	5.12	<.001***
Grade level (5)	0.41	0.16	2.58	.01*
EAL (EAL speakers)	−0.1	0.17	−0.61	.54
Group × Pretest interaction	−0.2	0.12	−1.61	.11
Group × Grade interaction	0.38	0.23	1.64	.10
Group × EAL interaction	−0.27	0.23	−1.18	.24
Random effect	**Variance**		***SD***	
Intercept (school)	0.02		0.13	
Residual	0.71		0.85	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	653.98		.29	.31

Note. EAL = English as an additional language.
*p < .05. ***p < .001.

TABLE B29
Trained Words: Intention-to-Treat Analysis

Fixed effect	Estimate	*SE*	*t*
Intercept	−0.13	0.15	−0.87
Gender (female)	−0.17	0.11	−1.54
Eligibility for free school meals (eligible)	−0.11	0.12	−0.95
Group (Motivated Reading)	0.08	0.18	0.46
Pretest score	0.39	0.08	4.77*
Grade level (5)	0.41	0.16	2.57*
EAL (EAL speakers)	−0.13	0.17	−0.80
Group × Pretest interaction	−0.16	0.12	−1.37
Group × Grade interaction	0.36	0.23	1.58
Group × EAL interaction	−0.27	0.23	−1.18
Random effect	***SD***		
Intercept (school)	0.15		
Residual	0.85		
Model fit			
Akaike information criterion	672.2		

Note. EAL = English as an additional language.
*p < .05.

Morphological Vocabulary Task

TABLE B30
Nonwords: Complete-Case Analysis

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	−0.07	0.17	−0.44	.66
Gender (female)	−0.05	0.12	−0.43	.67
Eligibility for free school meals (eligible)	0.09	0.13	0.65	.52
Group (Motivated Reading)	−0.05	0.20	−0.23	.82
Pretest score	0.37	0.09	4.15	<.001***
Grade level (5)	0.10	0.18	0.57	.57
EAL (EAL speakers)	−0.01	0.18	−0.04	.97
Group × Pretest interaction	−0.08	0.13	−0.63	.53
Group × Grade interaction	0.16	0.25	0.65	.52
Group × EAL interaction	−0.02	0.25	−0.09	.93
Random effect	**Variance**		***SD***	
Intercept (school)	0.03		0.17	
Residual	0.87		0.93	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	681.18		.13	.16

Note. EAL = English as an additional language.
***p < .001.

TABLE B31
Nonwords: Intention-to-Treat Analysis

Fixed effect	Estimate	*SE*	*t*
Intercept	−0.08	0.17	−0.49
Gender (female)	−0.04	0.12	−0.32
Eligibility for free school meals (eligible)	0.06	0.13	0.46
Group (Motivated Reading)	−0.02	0.20	−0.10
Pretest score	0.34	0.09	4.02*
Grade level (5)	0.11	0.17	0.66
EAL (EAL speakers)	−0.03	0.18	−0.18
Group × Pretest interaction	−0.06	0.13	−0.44
Group × Grade interaction	0.15	0.25	0.59
Group × EAL interaction	−0.03	0.25	−0.11
Random effect	***SD***		
Intercept (school)	0.19		
Residual	0.93		
Model fit			
Akaike information criterion	715.2		

Note. EAL = English as an additional language.
*p < .05.

Morphological Vocabulary Task

TABLE B32
Untrained Words: Complete-Case Analysis

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	0.06	0.16	0.35	.73
Gender (female)	−0.08	0.11	−0.77	.44
Eligibility for free school meals (eligible)	−0.05	0.12	−0.40	.69
Group (Motivated Reading)	0.01	0.18	0.04	.97
Pretest score	0.65	0.09	7.34	<.001***
Grade level (5)	−0.09	0.17	−0.50	.62
EAL (EAL speakers)	−0.01	0.16	−0.07	.94
Group × Pretest interaction	−0.22	0.13	−1.77	.08
Group × Grade interaction	0.21	0.24	0.85	.40
Group × EAL interaction	−0.13	0.22	−0.58	.56
Random effect	**Variance**		***SD***	
Intercept (school)	0.05		0.23	
Residual	0.65		0.81	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	621.28		.30	.36

Note. EAL = English as an additional language.
***p < .001.

TABLE B33
Untrained Words: Intention-to-Treat Analysis

Fixed effect	Estimate	*SE*	*t*
Intercept	0.06	0.16	0.37
Gender (female)	−0.06	0.11	−0.60
Eligibility for free school meals (eligible)	−0.05	0.12	−0.41
Group (Motivated Reading)	−0.02	0.18	−0.12
Pretest score	0.62	0.09	7.21*
Grade level (5)	−0.09	0.17	−0.51
EAL (EAL speakers)	−0.06	0.16	−0.37
Group × Pretest interaction	−0.21	0.12	−1.67
Group × Grade interaction	0.22	0.24	0.94
Group × EAL interaction	−0.08	0.22	−0.35
Random effect	***SD***		
Intercept (school)	0.23		
Residual	0.81		
Model fit			
Akaike information criterion	657.1		

Note. EAL = English as an additional language.
*p < .05.

Multiple-Choice Vocabulary Task

TABLE B34
Complete-Case Analysis

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	−0.41	0.12	−3.46	<.001***
Gender (female)	−0.10	0.09	−1.03	.30
Eligibility for free school meals (eligible)	−0.04	0.10	−0.45	.65
Group (Motivated Reading)	0.32	0.16	2.05	.04*
Pretest score	0.37	0.08	4.33	<.001***
Grade level (5)	0.96	0.15	6.30	<.001***
EAL (EAL speakers)	−0.16	0.14	−1.14	.26
Group × Pretest interaction	0.08	0.11	0.73	.47
Group × Grade interaction	−0.36	0.22	−1.66	.10
Group × EAL interaction	−0.18	0.20	−0.88	.38
Model fit	**Multiple R^2**		**Adjusted R^2**	
	.52		.50	

Note. EAL = English as an additional language. Results are from a regression model without random effects, as the fit for the model with random intercept for school was singular.
$^{*}p < .05$. $^{***}p < .001$.

TABLE B35
Intention-to-Treat Analysis

Fixed effect	Estimate	*SE*	*t*
Intercept	−0.45	0.12	−3.74*
Gender (female)	−0.09	0.09	−0.99
Eligibility for free school meals (eligible)	−0.06	0.10	−0.57
Group (Motivated Reading)	0.35	0.16	2.27*
Pretest score	0.32	0.09	3.68*
Grade level (5)	1.03	0.15	6.69*
EAL (EAL speakers)	−0.16	0.14	−1.09
Group × Pretest interaction	0.13	0.12	1.15
Group × Grade interaction	−0.43	0.22	−1.95
Group × EAL interaction	−0.21	0.20	−1.07

Note. EAL = English as an additional language. Results are from a regression model without random effects, as the fit for the model with a random intercept for school was singular. The single-level regression analysis was conducted using the Zelig package in R (Choirat, Honaker, Imai, King, & Lau, 2020).
$^{*}p < .05$.

NGRT

TABLE B36
Complete-Case Analysis

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	0.00	0.13	0.03	.98
Gender (female)	0.18	0.10	1.90	.06
Eligibility for free school meals (eligible)	−0.18	0.10	−1.70	.09
Group (Motivated Reading)	0.07	0.16	0.42	.67
Pretest score	0.63	0.07	8.89	<.001***
Grade level (5)	−0.04	0.14	−0.33	.74
EAL (EAL speakers)	−0.13	0.14	−0.90	.37
Group × Pretest interaction	0.02	0.10	0.21	.83
Group × Grade interaction	0.10	0.19	0.50	.62
Group × EAL interaction	−0.19	0.20	−0.98	.33
Random effect	**Variance**		***SD***	
Intercept (school)	0.01		0.09	
Residual	0.53		0.73	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	567.71		.47	.48

Note. EAL = English as an additional language.
***$p < .001$.

TABLE B37
Intention-to-Treat Analysis

Fixed effect	Estimate	*SE*	*t*
Intercept	0.02	0.13	0.14
Gender (female)	0.17	0.10	1.75
Eligibility for free school meals (eligible)	−0.20	0.10	−1.98
Group (Motivated Reading)	0.09	0.16	0.54
Pretest score	0.61	0.07	8.38*
Grade level (5)	−0.02	0.14	−0.15
EAL (EAL speakers)	−0.12	0.14	−0.89
Group × Pretest interaction	0.06	0.10	0.55
Group × Grade interaction	0.05	0.19	0.27
Group × EAL interaction	−0.20	0.20	−1.01

Note. EAL = English as an additional language. Results are from a regression model without random effects, as the fit for the model with a random intercept for school was singular. The single-level regression analysis was conducted using the Zelig package in R (Choirat, Honaker, Imai, King, & Lau, 2020).
*$p < .05$.

Motivation to Read

TABLE B38
Complete-Case Analysis

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	0.22	0.14	1.59	.11
Gender (female)	0.16	0.11	1.47	.14
Eligibility for free school meals (eligible)	−0.22	0.12	−1.87	.06
Group (Motivated Reading)	−0.39	0.18	−2.20	.03*
Pretest score	0.60	0.09	6.80	<.001***
Grade level (5)	−0.34	0.15	−2.18	.03*
EAL (EAL speakers)	0.06	0.16	0.39	.70
Group × Pretest interaction	0.01	0.12	0.13	.90
Group × Grade interaction	0.41	0.22	1.88	.06
Group × EAL interaction	0.05	0.22	0.22	.83
Model fit	**Multiple R^2**		**Adjusted R^2**	
	.41		.38	

Note. EAL = English as an additional language. Results are from a regression model without random effects, as the fit for the model with random intercept for school was singular.
**p* < .05. ****p* < .001.

TABLE B39
Intention-to-Treat Analysis

Fixed effect	Estimate	*SE*	*t*
Intercept	0.20	0.14	1.41
Gender (female)	0.20	0.11	1.80
Eligibility for free school meals (eligible)	−0.17	0.12	−1.42
Group (Motivated Reading)	−0.39	0.18	−2.18*
Pretest score	0.59	0.09	6.80*
Grade level (5)	−0.35	0.16	−2.25*
EAL (EAL speakers)	0.01	0.15	0.08
Group × Pretest interaction	0.02	0.12	0.14
Group × Grade interaction	0.36	0.22	1.63
Group × EAL interaction	0.19	0.22	0.87

Note. EAL = English as an additional language. Results are from a regression model without random effects, as the fit for the model with a random intercept for school was singular. The single-level regression analysis was conducted using the Zelig package in R (Choirat, Honaker, Imai, King, & Lau, 2020).
$^*p$ < .05.

Fidelity Analysis

TABLE B40
Results of Hierarchical Regression Analysis Comparing Group Fidelity Ratings

Fixed effect	Estimate	*SE*	*t*	*p*
Intercept	3.44	0.08	43.40	.00*
Group (Structured Word Inquiry)	−0.16	0.09	−1.83	.07
Random effect	**Variance**		***SD***	
Intercept (school)	0.06		0.24	
Residual	0.19		0.44	
Model fit	**Akaike information criterion**		**Pseudo-R^2 (fixed)**	**Pseudo-R^2 (total)**
	171.71		.02	.25

*$p < .05$.