Lexical Quality Matters: Effects of Word Knowledge Instruction on the Language and
Literacy Skills of Third- and Fourth-Grade Poor Readers

Ellen Irén Brinchmann, Hanne Næss Hjetland and Solveig-Alma Halaas Lyster

University of Oslo, Oslo, Norway

Abstract

The purpose of this study was to explore the hypothesis that teaching children knowledge of word forms and meanings supports the development of decoding and linguistic comprehension, which are fundamental components of reading comprehension. We examined this hypothesis by investigating the effects of a comprehensive word knowledge intervention on the language and literacy skills of poor readers. The participants included 118 monolingual third- and fourth-grade students from twelve Norwegian elementary schools. A quasi-experimental approach was employed with children in the treatment and control groups matched on grade and reading comprehension level. The intervention was delivered by teachers in small groups for 60 minutes three times per week over a period of ten weeks. At the end of the intervention, the treatment group showed significantly greater gains than the control group on researcher-created and transfer measures of language and a transfer measure of reading comprehension. There were no statistically significant effects of the intervention on two measures of decoding. The results support the hypothesis that comprehensive word knowledge instruction is effective in improving language abilities underpinning reading comprehension. The utility of this approach for improving decoding abilities remains unclear.

Lexical Quality Matters: Effects of Word Knowledge Instruction on the Language and
Literacy Skills of Third- and Fourth-Grade Poor Readers

The Simple View of Reading has proven to be one of the most influential frameworks for understanding reading comprehension in general and reading disabilities in particular. According to this model, reading comprehension is the product of two distinct components: decoding and linguistic comprehension. Decoding refers to the ability to recognize written words by using rules of letter-sound correspondences, whereas linguistic comprehension is broadly understood as “the process by which, given lexical (i.e., word) information, sentences and discourses are interpreted” (Gough & Tunmer, 1986, p. 7). Because the two components are considered distinct, the model predicts that poor readers may have profiles that differ across the dimensions of decoding and linguistic comprehension. Reading disabilities may thus emerge in three different ways: from deficits in decoding, deficits in linguistic comprehension or both.

In the years since Gough and Tunmer first presented the Simple View of Reading, several studies have confirmed the existence of the hypothesized subgroups of reading disabilities (e.g., Catts, Hogan, & Fey, 2003). There has been particular interest in children with deficits specific to either decoding or linguistic comprehension. An extensive body of research has shown that children with deficits in letter knowledge and phoneme awareness demonstrate difficulties with decoding despite normal levels of linguistic comprehension (see Vellutino, Fletcher, Snowling, & Scanlon, 2004, for a review). There is also accumulating evidence that deficits in semantic and grammatical understanding may cause reading disabilities in children who otherwise display adequate decoding proficiency (Catts, Adlof, & Weismer, 2006; Nation, Cocksey, Taylor, & Bishop, 2010).

Because the proximal cause of reading disabilities varies across subgroups, different approaches may be required to prevent and remediate difficulties in children with various

reader profiles (Kirby & Savage, 2008; Snowling & Hulme, 2011). However, although some poor readers have ability profiles with marked dissociations between decoding and linguistic comprehension, the boundaries between these two components are often not clear cut (Snowling & Hulme, 2012). Several studies have shown that oral language difficulties beyond phonology are prevalent in children with poor decoding skills (Catts, Fey, Zhang, & Tomblin, 1999; Snowling, Muter, & Carroll, 2007) and that children with oral language disorders frequently demonstrate difficulties in decoding (Botting, Simkin, & Conti-Ramsden, 2006; McArthur, Hogben, Edwards, Heath, & Mengler, 2000). In other words, although the existence of readers who have deficits limited to either decoding or linguistic comprehension is evident, there are also some indications of what Gough and Tunmer (1986) called a trivial prediction: most poor readers may be deficient in both components.

From an educational point of view, Gough and Tunmer's prediction is far from trivial. In most school populations, teachers face a heterogeneous group of learners who represent a wide variety of reader profiles. At the same time, teachers often have limited resources available, and in many school contexts, the implementation of specialized instruction for specific groups may not be achievable. Ultimately, even if it is important to differentiate instruction based on readers' profiles, there are also good reasons for exploring a more comprehensive approach that is applicable for diverse groups of readers. This is the goal of the present study, in which we investigate the effects of an intervention designed to stimulate abilities underpinning both decoding and linguistic comprehension. Compared with interventions that have a specific focus on one of the two components, there is little research on this approach to reading remediation. In this paper, we argue that the approach has both practical and theoretical merits.

Rationale for the Intervention

Even when a clear discrepancy between decoding and linguistic comprehension is assumed, theories of compensatory strategies in reading may arguably form a bridge between the two components. According to this theoretical approach, bottom-up and top-down processes work in parallel during reading, which means that information from different sources (i.e., phonological, orthographic, grammatical and semantic) is synthesized simultaneously (Nation & Snowling, 1998a; Stanovich, 1980). Consequently, deficits in decoding, for example, might be compensated by adequate linguistic comprehension skills. Evidence for this theory comes from experimental studies indicating that exception words are typically read with support from semantics (Nation & Snowling, 1998b) and that decoding is facilitated by reading words in context (Nation & Snowling, 1998a). Interestingly, experimental research has also demonstrated possible contributions from decoding to linguistic comprehension. In a series of studies, Rosenthal and Ehri (2008, 2011) showed that exposure to and pronunciation of the spelling of words enhance students' vocabulary learning. Ehri (2014) proposed that orthographic knowledge diminishes dependence on phonological memory and that superior ability to connect spellings to pronunciations in memory might explain why good readers tend to build larger vocabularies through reading than do poor readers.

The findings of Nation and Snowling (1998a, 1998b) and Rosenthal and Ehri (2008, 2011) suggest that different features of language may interact and reinforce each other, thus leading to higher levels of reading proficiency. These interactions lie at the core of the theoretical position of the Lexical Quality Hypothesis (Perfetti, 2007; Perfetti & Hart, 2002), which is an inherent part of Perfetti and Stafura's (2014) Reading Systems Framework. In line with the Simple View of Reading, this hypothesis depicts two major systems of reading: (a) a word identification system, in which orthography is mapped onto phonology, and (b) a comprehension system, in which language processing mechanisms are engaged to assemble

the identified words into units of meaning (Compton, Miller, Elleman, & Steacy, 2014; Perfetti & Stafura, 2014). Astride the two major reading systems sits the mental lexicon, a subsystem in which representations of words are stored. What is interesting in this regard is the emphasis on the *quality* of these lexical representations. Lexical quality refers to the extent to which a reader's knowledge of a given word represents the features of four constituents of word identity: orthography, phonology, semantics and morpho-syntax (Perfetti, 2007). Together, the quality of these four features and the coherence among them facilitate the rapid, low-resource retrieval of lexical word identities and their integration into a mental model of the text (Perfetti, 2007; Perfetti & Stafura, 2014). In summary, the hypothesis postulates that the lexicon is the midpoint where word identification and comprehension processes meet and that the quality of a reader's word knowledge forms the foundation of reading comprehension (Cain & Parrila, 2014).

The Present Study

An important aim of the present study was to design a theoretically based instructional program that could accommodate the challenges of educational practice by focusing on both decoding and linguistic comprehension. Because the Lexical Quality Hypothesis emphasizes the interface between these two components, we decided to turn some of its theoretical assumptions into principles for teaching. Specifically, we chose to emphasize *word knowledge* as the focal point of our intervention. Moreover, we wanted to enhance lexical knowledge of written word forms and their meanings by integrating instruction in the different features of word identity. Instruction in the semantic and morpho-syntactic properties of words was thus accompanied by work on the spelling and pronunciation of the corresponding word forms. Based on the theoretical rationale above, we hypothesized that children with different reader profiles could benefit from this type of instruction.

A complementary principle guiding the intervention was the idea of *supporting cumulative knowledge* (Nagy, 2005). There has been an ongoing discussion in the field regarding the effectiveness of word knowledge instruction. Instruction that is focused on teaching individual word meanings has in particular been criticized because the sheer number of words needed to make a substantial impact on children's lexicons is so vast that direct instruction is bound to be futile (Nagy & Herman, 1984). However, we argue that word knowledge instruction can matter over time if it provides children with cumulative knowledge. An example of this type of knowledge is knowledge of the morphological features of words. Words are often built by combining different morphemes, which are the smallest units of meaning in a language (e.g., *un – friend – ly*). Because morphemes are meaningful units, knowledge of different morphemic constituents may be used to infer the meaning of unfamiliar words (for example, inferring the meaning “small tree” from “tree + -let”) (Anglin, 1993). Morphological knowledge may therefore lead to an increase in vocabulary over time (Carlisle, McBride-Chang, Nagy, & Nunes, 2010). Furthermore, morphemes have consistent spellings, and the spelling of a specific morpheme may be generalized to a large number of words. Learning the orthographic features of morphemes may thus consolidate bindings between written and spoken units across words and support the development of decoding abilities (Carlisle & Stone, 2005). In sum, understanding the morphological properties of words may arguably stimulate growth in both semantic and orthographic knowledge over time, and morphemes are therefore useful units to focus on when integrating instruction in word forms and meanings.

Comprehensive Word Knowledge Interventions

To the extent that we are aware, few intervention studies have explicitly encompassed the theoretical framework of the present study or included a particular focus on the integration of different features of word identity. There are, however, several studies that incorporate a

similar comprehensive approach to word knowledge instruction. Although focused on a somewhat different population than the present study, an intervention conducted by Carlo et al. (2004) represents a good example. Their quasi-experimental study evaluated the effects of a multi-faceted vocabulary program on the language and literacy skills of 254 monolingual and bilingual fifth graders. The program focused on teaching multiple aspects of word knowledge, including depth of meaning, morphological structure, and the pronunciation and spelling of target words. Following 15 weeks of instruction, the treatment group showed significantly greater gains compared with a “business-as-usual” control group on several researcher-created measures of word knowledge and a measure of reading comprehension. Although effect sizes were modest and there were no transfer effects to a standardized measure of vocabulary, the results are encouraging considering that the intervention was conducted in a naturalistic classroom setting with learners from highly diverse linguistic backgrounds.

A similar study was conducted by Lesaux, Kieffer, Faller, and Kelley (2010). As in the study by Carlo et al. (2004), a comprehensive vocabulary program was introduced in middle school classrooms with high proportions of second language learners. Measured against instruction in a set of comparison classrooms, the 18-week program yielded statistically significant effects on researcher-created measures of taught vocabulary words, knowledge of word meanings in context, and morphological skills. A small effect size, falling just below significance, was also found on a norm-referenced measure of reading comprehension.

The studies of Carlo et al. (2004) and Lesaux et al. (2010) were primarily focused on second language learners. However, the principles underlying both of these interventions may be considered general principles for good word knowledge instruction, regardless of learner status. This is reflected in the results of the studies, where effect sizes were found to be comparable for second language learners and their English-speaking peers. In addition to

promoting word knowledge instruction for heterogeneous groups of learners, the studies of Carlo et al. (2004) and Lesaux et al. (2010) share another important principle with the present study. In discussing the effects of the interventions, both Carlo et al. (2004) and Lesaux et al. (2010) argued that a comprehensive approach with a combined emphasis on word-specific and generative word knowledge (e.g., morphological knowledge) will bolster vocabulary and comprehension skills over time. In other words, knowledge was expected to cumulate.

A study that tests this assumption was performed by Morris et al. (2012). They employed a factorial design that contrasted the effects of four experimental conditions: two different multiple-component conditions, a phonological-only condition and a math control condition. Both of the multiple-component conditions included a simultaneous focus on phonological, orthographic, and morphological word knowledge. However, only one of these programs included an additional focus on semantic and syntactic word knowledge. A sample of 279 second- and third-grade struggling readers with diverse reader profiles was randomly assigned to one of the four conditions. After the 70-hour intervention was completed, all three linguistic conditions proved superior to the math condition on several standardized measures of spelling and reading (decoding accuracy, fluency, and comprehension). Moreover, the two multiple-component programs demonstrated superior outcomes on the majority of measures compared with the phonological-only control group. There was also a tendency for the multiple-component group that had a supplementary focus on semantics to perform better than the group without this instructional component (the difference was statistically significant on a composite measure of reading comprehension and fluency). What is more striking, however, is that these patterns of results were maintained one year after the intervention ended. An additional advantage even emerged for the two multiple-component groups over the phonological-only control group on a measure of spelling and a measure of

word reading efficiency. In other words, the study not only demonstrated longitudinal effects but also showed indications of cumulative learning.

Together with other intervention studies utilizing various approaches to comprehensive word knowledge and reading instruction (e.g., Duff et al., 2008; Vadasy, Sanders, & Peyton, 2006; Wolter & Dilworth, 2014), the studies by Carlo et al. (2004), Lesaux et al. (2010), and Morris et al. (2012) showed promise in promoting language and literacy skills in different learner groups. However, all of the aforementioned studies were conducted in a North American or British educational setting using English as the instructional language. Although progress toward a more universal reading science has been made, the convergence of empirical evidence across educational and linguistic contexts may strengthen theoretical claims regarding reading and reading remediation. The present study thus expands the extant research base by evaluating the effects of comprehensive word knowledge instruction in a Norwegian educational and linguistic context.

Similar to English, Norwegian belongs to the Germanic subgroup of the Indo-European language family. The syntactic and morphological structure of the language is fairly similar to that of English, and the Norwegian language contains many of the affixes that are found in English (e.g., *mislede* [*misguide*], *maler* [*painter*], *snakking* [*talking*]). The pronunciation of morphemes, however, is much more consistent in Norwegian than in English, and compounding is used more extensively. Norwegian orthography is morpho-phonemic, although the letter-sound correspondence is more regular compared with English or French (for more information on the Norwegian language and orthography, see Hagtvet, Helland, & Lyster, 2006). Because the Norwegian orthography is relatively transparent, measures of reading accuracy typically show little variance, and even reading-disabled students often break the alphabetic code rather easily (Hagtvet et al., 2006; Lervåg, Bråten, & Hulme, 2009). This does not mean that poor decoding abilities are not found in samples of

Norwegian children, but their difficulties are first and foremost at the level of reading efficiency.

It has been argued that conclusions drawn from studies conducted in orthographically inconsistent languages may not generalize to consistent orthographies and vice versa (e.g., Seymour, 2005). Because the theory underlying the present study is developed in an English language context, it may therefore not be equally applicable to interventions designed for Norwegian readers. Nevertheless, even if there is some evidence indicating that children learn to read more rapidly in consistent than in less consistent orthographies, the cognitive and linguistic prerequisites for learning to read are highly similar across orthographies (Caravolas, Lervåg, Defior, Málková, & Hulme, 2013). We therefore expected that the guiding principles of the intervention would transfer well to a Norwegian context. However, we did take some of the characteristics of Norwegian readers into account when designing the word knowledge intervention. For example, to accommodate efficiency difficulties rather than difficulties in accuracy, we decided to focus on the phonological and orthographic properties of larger units (e.g., morphemes) as opposed to single phoneme-grapheme correspondences.

Method

Design

We conducted a quasi-experimental study to investigate the effects of a comprehensive word knowledge intervention on the language and literacy skills of children identified as poor readers. The intervention was delivered by teachers in small groups three times per week for a total of 10 weeks. The children's progress during the intervention was compared with a "business-as-usual" control group.

Participants

The participants included 118 monolingual third- and fourth-grade students from a small municipality located in southeastern Norway. Parental consent for participation in the

study was received for all children. Children in the treatment group came from eight elementary schools. They were identified by the teachers who were responsible for their reading instruction and subsequently targeted for intervention. Teachers were instructed to select children who struggled with decoding and/or understanding text, and the teachers' selections were also validated by data from a screening conducted in all of the participating schools. Children from four other schools within the same school district served as controls after being matched with the children in the treatment group (more details on the sampling procedures are provided below). This approach resulted in an equal number of participants ($n = 59$) in the two groups. Descriptive statistics on the demographics of the sample are shown in Table 1. There was no attrition in either group.

Although the teachers in the study varied somewhat in regard to their level of education and years of teaching experience, the teachers' backgrounds were comparable across treatment and control groups. The 11 teachers who delivered the treatment were third- and fourth-grade teachers who worked with reading instruction in the intervention schools. Most of them were class teachers, although 2 worked as special needs teachers. However, all were educators that the children knew well and were in contact with on a regular basis. All of the teachers with the exception of one had more than 3 years of teacher education ($M = 4.28$ years, $SD = 2.19$). Although one of the teachers lacked a formal degree, she had worked as a teacher for 28 years and had taken several courses in special needs education and reading instruction. The other teachers were also highly experienced; only two had less than 10 years of teaching experience (5 and 8 years of experience; $M = 20.00$ years, $SD = 11.05$). All of the teachers consented to participate in the study.

Sampling Procedures

The present study was initiated at the beginning of the school year of 2012/2013, when the participants were second and third graders. At this time, all 13 schools within the

municipality were invited to participate in the study. Although the schools varied in size, they were all public schools run by the same local administration and were subject to equal economic conditions. Twelve schools agreed to participate and were given the option to function as an intervention or control school. Intervention schools were obliged to deliver the treatment according to instructions, while control schools could implement the intervention on their own terms when the study was over. However, no training, instructions or materials would be released to the control schools until the post-test was completed. Eight schools chose to act as intervention schools and four decided to function as control schools and conduct the intervention at a later point in time.

During the spring semester, we conducted screenings in all year 2 and 3 classrooms in the participating schools. A total of 477 students (second grade: $n = 215$, M age = 91 months, $SD = 3.52$; third grade: $n = 262$, M age = 102, $SD = 4.26$) were screened with group-administered measures of reading comprehension and orthographic, morphological and semantic word knowledge (the measures are described later). This screening served three important purposes. First, because there was no random assignment of schools to treatment conditions, the screening was used to explore potential differences between the intervention and control schools. Analyses of the school-level differences are presented in the results section.

Second, the results from the screening were used to conduct an initial matching of children in the treatment and control groups. At this stage, we located children in the control schools who were in the same grade level and had similar reading comprehension scores as the children selected for treatment in the intervention schools. From this pool of children, a proximal matching based on profiles across levels of orthographic and morpho-semantic word knowledge was conducted. However, to reduce selection bias due to regression to the mean, we oversampled the control group by including all children that represented a close match

with someone in the treatment group ($n = 78$). The final one-to-one matching of children was completed based on scores on an individually administered test of reading comprehension (subtest from the Wechsler Individual Achievement Test Second Edition; Wechsler, 2005) that was conducted at the time of the pre-test (fall semester, 2013).

Finally, because we relied on teacher judgment in targeting children for treatment in the intervention schools, the screening allowed us to validate the teachers' selections by obtaining information about the performance of these children compared with their classmates. A comparison of the children in the treatment group and their peers is presented in the results section of this paper.

Intervention Program

The intervention program consisted of 30 sessions delivered by the participating teachers during regular school hours. Children who received the intervention were pulled from their classes and seen in groups of 5-6 participants (with the exception of one group that consisted of 9 children). This arrangement was achieved by organizing the sessions around hours when there was more than one adult available for the class. Five of the eight intervention schools had one instructional group, and the other three schools instructed two groups. Therefore, there was a total of 11 groups across the intervention schools. The treatment group received training for approximately 60 minutes three times per week for ten weeks. Because word knowledge programs have proven most effective when instruction is embedded in meaningful contexts (Marulis & Neuman, 2013), each of these weeks consisted of a lesson cycle that revolved around a theme reflected in a short expository text (between 150-200 words). The texts were selected from subject areas that included Norwegian language arts, social sciences, history and natural sciences, and covered themes such as civil protest, occupations, prehistoric animals and recycling. A brief description of each day in the lesson cycle is presented in Table 2.

To ensure that the texts provided a rich and meaningful context for word knowledge learning, the first day of the lesson cycle was mainly spent on reading and discussing the text. Each child had the opportunity to read a section of the text out loud while receiving feedback from the teacher. Because the texts were relatively short, parts of the texts usually had to be repeated to ensure that all of the children had a chance to read. The goal of discussing the text was twofold: to clarify the content and to expand on the children's knowledge of the theme. The content was clarified by discussing difficult words in the text in addition to literal questions that could be answered by searching the text. A deeper discussion of the theme was organized around open-ended questions that could not be answered by simply searching the text. Typically, these questions required the children to make inferences or to draw on their own experiences to answer. Questions that could be used for discussion were available in the teaching materials.

Each week, three target words were selected from the text and used as a starting point for word knowledge activities. Although the activities varied in type, each lesson cycle included activities that aimed to stimulate the semantic, syntactic and morphological features of word identity. To support the integration of form and meaning, instruction in phonological and orthographic word knowledge was incorporated as part of these activities. Instruction related to a specific morpheme, for example, included phonological and orthographic comparisons of different words containing this morpheme. To enhance the consolidation of orthographic mappings in memory, the children also engaged in repeated readings of the text on days two and three of the lesson cycle. Examples of the word knowledge activities are presented in Table 3.

As noted above, three target words were selected each week from a text. The selected words were targeted because they were determined to be general-purpose academic words that may be found across a variety of domains (Beck, McKeown, & Kucan, 2002). However,

although the total number of target words was relatively low, the target words were used as a starting point for activities that stimulated knowledge of other word meanings as well as generative word knowledge. For example, semantically oriented activities typically involved work with semantic word boundaries (e.g., what constitutes a concept and what does not), discussion of multiple word meanings and work with semantically related words (e.g., synonyms, antonyms, subordinates and superordinates). The goal of these activities was to add a fuller range of meaning dimensions to the target words and enable the children to discriminate among words in the same semantic field (Perfetti, 2007).

Target words were also used in activities designed to increase knowledge of syntactical word features. These activities included the active construction of sentences using words as pieces in a sentence puzzle. Typically, the pieces could be combined in two or three different ways to produce a correct sentence. Various solutions could therefore be discussed in the group. Other activities involved work with sentences that were syntactically incorrect. In these tasks, the children had to discuss alternative ways to repair the sentence to make it correct (e.g., the sentence *It is not a smart to be disadvantage* can be repaired by changing the order of *smart* and *disadvantage*.). The goal of the syntactic activities was to increase the stability of syntactical word representations and to make the children reflect on how words can be combined to convey meaning on a sentence level. Because target words were embedded in the sentences (e.g., *disadvantage*), the activities also provided examples of how these words could be used.

Some of the words that were directly taught in the intervention were specifically chosen because of their morphological properties. These words were used in activities that were designed to stimulate the children's knowledge of morphological word features. Such activities were mainly directed toward derivational morphology, but some work with compounding was also included. The morphological activities were semantically oriented,

meaning that reflection on how different affixes change the meaning of root words was emphasized. These activities were typically word puzzles in which the children constructed different words by combining affixes and roots. To support the learning of affixes, the creation of novel words was also encouraged. In such instances, discussion of the meaning of the new words was central. Morphological activities also included working with word families (i.e., words sharing the same root as the target word) and affix families (i.e., words sharing the same affix as the target word). By including work with word families, we aimed to stimulate the children's lexical networks and increase the generalization of morpheme knowledge by providing high exemplar variability (von Koss Torkildsen, Dailey, Aguilar, Gómez, & Plante, 2013).

Because all of the target words were embedded in the texts, the texts provided direct examples of the words' usage and thus presented a meaningful context for semantically related activities. However, the themes in the texts were also reflected in the morphological activities. For example, when the text of the week was about *recycling* and *renewable resources*, morpheme activities addressed the prefix *re-* and the suffix *-able*; when the text was about *prehistoric animals*, the prefix *pre-* was taught; a text about *civil protest* was followed by work with the prefix *anti-* (these examples are from words and affixes that are directly translatable from Norwegian to English). The texts thus provided a semantic framework not only for the target words but also for the word parts taught in the intervention.

Treatment Fidelity

The teachers attended a two-day intensive course before the onset of the intervention. This course included a total of 12 hours of instruction in the theoretical background of the study and thorough stepwise instruction in the different intervention activities and materials. All of the lessons, including the activities, procedures, materials and resources, were described in a detailed manual that the teachers followed. Before the intervention began, this

manual had been piloted by the first author on a small group of 6 third and fourth graders to ensure that the materials were easy to use and appropriate for the relevant age group (the children who participated in the pilot were not part of the sample described in the present study). When needed, further support for the teachers was provided through telephone and email contact over the course of the intervention.

After each session, the teachers completed logs reporting the time spent on the lesson, attendance and completed activities. All of the logs were retrieved after the intervention, with the exception of one log that was lost due to an electronic malfunction. Information was provided retrospectively from the teacher who wrote this log, but the validity of this information is naturally somewhat reduced. An analysis of the teacher logs revealed that the mean length of the sessions was 53.58 minutes ($SD = 11.17$). However, one teacher reported having administered the sessions somewhat differently from what was intended; during all text-related activities, she included her entire class ($n = 12$, including the children in the treatment group). Additionally, the five children in the treatment group received 30 minutes of instruction outside of the classroom three times a week. This situation made it difficult to estimate the total time the group spent on each session. Of a total of 59 children in the treatment group, 32 were absent from one or more sessions due to illness. Five of these children were absent for four or five sessions, and one child was absent from eight of the 30 sessions.

Following each session, the teachers also completed a short evaluation of the lesson and made suggestions for improvement. An analysis of all of the registered responses across the treatment groups revealed a satisfaction rate of 95.7%. In other words, when asked whether the lesson worked, the teachers answered “no” in only 4.3% of the instances. Subsequent suggestions for improvement were related to the difficulty level of the activities,

texts or words as well as suggestions related to time estimations (e.g., more time needed to complete the activities).

Measures

Screening Measures

As previously mentioned, we conducted screenings in all of the intervention and control schools when the participants were second and third graders. Four group-administered tests assessing morphological knowledge, vocabulary, orthographic knowledge, and reading comprehension were used as screening measures. During the screening, the children were tested in their classrooms, with parental permission, by trained research assistants and the authors of this paper. The four tests were administered in three sessions that each lasted approximately 30-40 minutes.

Morphological knowledge. A researcher-created test named the *Affix Knowledge Test* was administered as a measure of children's knowledge about how affixes change the meaning of root words. In this test, the children had to choose the correct definition of a morphologically transparent novel word from four different response options. The words in the test had been created by adding prefixes and suffixes to high-frequency roots (e.g., lykkeløs [happyless], utørr [undry]). Although none of these words exists in Norwegian, their meaning could be inferred from their constituent morphemes. To reduce confounding of decoding abilities and working memory, all 40 items were presented in writing and read out loud by the test administrator. Answers were registered on individual response sheets. The Affix Knowledge Test was developed for the present study, where it demonstrated adequate psychometric properties. It had a Cronbach's alpha of .85 and showed an expected pattern of correlations with the other screening measures. That is, it was strongly correlated with a measure of vocabulary ($r = .61$) and correlated at a somewhat lower level with a test of orthographic knowledge ($r = .45$).

Vocabulary. A shortened version of the standardized Norwegian adaption of the *British Picture Vocabulary Scale Second Edition* (BPVS II; Dunn, Dunn, Whetton, & Burley, 1997) was used as a measure of vocabulary. For each item in this test, the children were presented with a set of four line drawings and had to choose the drawing that corresponded to a word spoken by the test administrator. All of the drawings were projected on a screen, and responses were registered by the children on individual sheets containing the 50 items. The Norwegian translation of the BPVS II has been frequently used in previous research, where it has demonstrated good reliability and validity (e.g., Lervåg & Aukrust, 2010; Melby-Lervåg et al., 2012). The shortened group-administered version of the test was adapted for use in the present study. The reliability for this version was acceptable ($\alpha = .79$), and the correlations with the other word knowledge measures provided evidence of convergent (Affix Knowledge: $r = .61$) and divergent validity (Orthographic knowledge: $r = .26$).

Orthographic knowledge. A Norwegian standardized orthographic choice test from the STAS battery (Klinkenberg & Skaar, 2003) was used as a measure of orthographic knowledge. During this test, the children had to choose the correctly spelled word from sets of words and non-words that sound alike when they are pronounced (e.g., vem-vemm-hvem [who]; only one of the words is correctly spelled). The test consisted of two subtests with 36 and 77 sets of words respectively. A time limit of 2 minutes was given for each subtest, and performance was measured as the number of correctly marked words minus the number of words incorrectly marked. In the analyses, the two subtests formed a composite score of orthographic knowledge ($r = .80$ between the subtests). Evidence of construct validity for this measure was provided by Lervåg et al. (2009). In their study, the test had estimated correlations with a measure of single-word reading ranging from $r = .87$ -.92 across three time points.

Reading comprehension. A Norwegian translation of the standardized *Gates-MacGinitie Reading Test Fourth Edition* (GMRT-4; MacGinitie & MacGinitie, 2006), was administered as a measure of reading comprehension. In this test, the children were given 35 minutes to read several expository and narrative texts and complete related multiple-choice items. Second- and third-grade versions of the test were administered to children from the corresponding grade levels to ensure that texts and items were age-group appropriate. However, in the current study, the psychometric properties of these two versions of the test were highly comparable. Both versions had a reliability coefficient of $\alpha = .93$; were moderately correlated with the orthographic choice test (second grade: $r = .56$; third grade: $r = .51$); and exhibited similar correlations with the linguistic screening measures (for second grade and third grade, $r = .32$ and $r = .42$, respectively, for the BPVS II, whereas $r = .60$ and $r = .62$, respectively, for the Affix Knowledge Test).

Treatment Measures

A test battery assessing language, literacy and cognitive skills was administered both before the intervention began and after the intervention had finished. The tests were administered in the same order for all of the children in the study, with measures assessing similar constructs balanced across two sessions that each lasted approximately 45 minutes. Testing was conducted by trained research assistants who saw the children individually at their own schools. Reliability estimates of the measures from the pre-test are presented in the results section.

Nonverbal ability. Nonverbal reasoning was measured using the *Matrix Reasoning Subtest* from the Wechsler Intelligence Scale for Children Fourth Edition (WISC-IV; Wechsler, 2003). During this test, the children were presented with a series of incomplete matrices. To complete each matrix, the children had to select the missing part from five different response options. This measure was primarily included to assess potential

differences between the treatment and control groups at pre-test. However, the test was also administered at post-test as a measure of the specificity of the intervention. In other words, because the intervention did not target nonverbal reasoning abilities, a treatment effect on this measure would violate the logic of the intervention and thus warrant methodological concerns. At pre-test, the measure of nonverbal reasoning was significantly correlated with all of the other linguistic and cognitive measures used in this study, with small to moderate Pearson's r coefficients ($r = .19-.40$).

A further assessment of the specificity of the intervention was conducted by including two tests of mathematical skills in the test battery. In the *One Minute Addition Test*, the children had to solve as many addition problems as possible within one minute. In the *One Minute Subtraction Test*, the same procedure was followed for subtraction problems (Westwood, Harris-Huges, Lucas, Nolan, & Scrymgeour, 1974). In the analysis, the two measures formed a combined mathematics score ($r = .77$ between the measures at pre-test).

Morpho-syntactic knowledge. The children's morphological knowledge was assessed using two different tests. In the *Word Analogy Test* (modified from Nunes, Bryant, & Bindman, 1997), the children were orally presented with a set of two words that reflected a morphological change. They had to use the pattern of change from this set to produce the correct morphological change in another word provided by the test administrator (e.g., "I say congratulations and change it to congratulate. We can also change operations to _____ [operate]"). All of the items included derivational morphology, but none of the affixes that appeared in the test were taught in the intervention. Different versions of the Word Analogy Test have previously been used in samples of English (Tong, Deacon, Kirby, Cain, & Parrila, 2011), French (Casalis, Deacon, & Pacton, 2011) and Danish children (Arnbak & Elbro, 2000). The Norwegian version used in the present study was piloted on a sample of 219 third-grade students in a study conducted by the third author of this paper and colleagues (see

Melby-Lervåg et al., 2012, for sample characteristics). This piloting yielded a Cronbach's alpha of .86. Evidence of convergent validity from the current sample includes moderate correlations with the other linguistic measures in the test battery (ranging between $r = .39$ -.49). A weaker correlation of .25 with the mathematics measure provides evidence of divergent validity.

Another researcher-created test was administered as a more proximal measure of intervention effects. The *Affix Knowledge Test* measures children's knowledge of how affixes change the meaning of root words. This test was similar to the Affix Knowledge Test used in the screening, with a few exceptions. As in the screening, morphologically transparent novel words had been created by adding prefixes and suffixes to high-frequency roots. In this version, however, the children were first presented with single words that they were asked to define. If the children were unable to provide a definition, the target word was presented in a sentence, and the children had to select a definition from four response options. The 25 test items had also been adapted from the original version to target affixes taught in the intervention (e.g., *mislesbar* [misreadable], *antimodig* [antibrave]). Because the affixes were taught in different root word contexts, however, the test could function as a measure of generalization of affix knowledge. Test items were scored on a scale from 0-2, with 2 points awarded for plausible definitions of the novel words and 1 point given for correct multiple-choice answers. All of the words and sentences were presented both orally and in writing. As in the screening, this measure was strongly correlated with vocabulary at pre-test (WISC-IV Vocabulary: $r = .62$; Taught vocabulary: $r = .61$), which can be considered further evidence of convergent validity. A non-significant correlation of $r = .07$ with the measure of mathematical skills provides evidence of divergent validity.

The *Formulating Sentences Subtest* from the Norwegian standardized version of the Clinical Evaluation of Language Fundamentals Fourth Edition (CELF-4; Semel, Wiig, &

Secord, 2003) was used as a measure of the ability to produce grammatically and semantically intact sentences. In this test, the children were shown a picture and had to verbally formulate a sentence related to the picture using a word provided by the test administrator. Each sentence's syntactic and grammatical structure and semantic content were scored on a scale from 0-2 according to the test manual. In the present study, the test showed moderate correlations with the other linguistics measures at pre-test (in the range of $r = .36-.46$) and a weaker correlation with mathematical skills ($r = .26$), which constitute evidence of convergent and divergent validity.

Vocabulary. Vocabulary was assessed using the *Vocabulary Subtest* from the Norwegian standardized adaption of the WISC-IV (Wechsler, 2003). The test measured the ability to define a series of words with increasing levels of difficulty. Correct definitions were awarded 2 points, whereas definitions that reflected vague knowledge of the word meanings were given 1 point. The test was discontinued after five consecutive errors.

A researcher-created measure assessing the children's knowledge of *words taught directly* in the intervention was also administered. This test followed the same administration and scoring procedures as the WISC-IV Vocabulary Subtest, except no stop criteria were included.

The two vocabulary measures showed an expected pattern of convergent and divergent correlations at pre-test, with a strong correlation between each other ($r = .62$) and weak correlations with the mathematics score (WISC Vocabulary: $r = .14$; Taught vocabulary: $r = .16$).

Literacy. Word decoding was measured using a Norwegian adaption of the *Sight Word Efficiency Subtest* from the Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999). The test measured the number of real words the children could accurately identify within 45 seconds. Two alternate forms from the subtest were

administered. Before entering the analysis, scores from the two forms were summed into a combined word decoding score.

The *Phonemic Decoding Efficiency Subtest* from the TOWRE was used to assess non-word decoding. This test was equal to the Sight Word Efficiency Subtest, except the children had to decode pronounceable non-words instead of real words. Two alternate forms were also used in this test.

Although the Norwegian adaption of the two subtests from TOWRE has not been as extensively used in research as the English version, some evidence of reliability and validity of the measures can be found in the literature (e.g., Lervåg & Aukrust, 2010). In the present study, the correlations between all four alternate forms of the two subtests ranged between $r = .78$ and $.95$, which indicates good construct validity.

Reading comprehension was measured using a Norwegian translation of the *Reading Comprehension Subtest* of the Wechsler Individual Achievement Test Second Edition (WIAT II; Wechsler, 2005). During this test, the children had to read expository and narrative passages as well as short sentences. They were instructed to read aloud and responded orally to open-ended comprehension questions after each text. Because the participants were poor readers, the test was administered with stop criteria. The test was discontinued when no correct answers were provided on the comprehension questions for two consecutive texts or when the test time exceeded 25 minutes. The English version of this test has been widely used in reading research, where it has demonstrated good psychometric properties (e.g., Berninger & Abbott, 2010). Evidence of criterion validity for the Norwegian translation was found in the present study, where the test showed moderate correlations with the linguistic measures ($r = .30-.49$) and somewhat higher correlations with measures of decoding (word decoding: $r = .58$; non-word decoding: $.54$).

Results

Analysis of Screening Data

An important requirement for causal interpretation of intervention effects is that treatment assignment is uncorrelated with outcome (Morgan & Winship, 2007). This prerequisite is often achieved by random allocation of subjects to treatment conditions. Because we chose to rely on non-random sampling procedures in the present study, we conducted several analyses of the screening data to assess potential sources of selection bias.

First, we explored school-level differences by estimating intra-class correlations (ICC_1 ; level 2: $N = 12$; level 1: $N = 477$). The ICC_1 is an estimate of how strongly units belonging to the same group resemble one another and is often used to assess the impact of school-level factors on individual differences. The analyses yielded coefficients ranging between .008 and .02 for the four screening measures, which means that the proportion of the variance that is accounted for by the school level is very low. In other words, student performance did not systematically vary between the schools in our study.

Second, we used propensity score adjustment to reduce the confounding bias caused by non-random treatment assignment (Rosenbaum & Rubin, 1983). The general idea behind this procedure is to estimate the probability of treatment selection given a set of background variables. Ideally, this probability should be balanced across treatment and control groups for intervention assignment to remain uncorrelated with outcome. We therefore calculated propensity scores for all of the students in the intervention and control schools, indicating the likelihood of being selected for treatment. This was achieved by running a logistic regression analysis using the screening measures (z-scored within grade level) as covariates. A comparison of the propensity scores for children in the treatment and control groups revealed that the means, variances, and ranges of scores were almost perfectly balanced across the two groups (treatment group: $M = .66$, $SD = .06$, variance = .003, range = .24; control group: $M =$

.66, $SD = .06$, variance = .003, range = .21). To further improve our estimates of the treatment effects, we included the propensity scores as a covariate in analyzing the intervention data.

Finally, because the children were selected to take part in the intervention based on teacher evaluations, we wanted to obtain a better understanding of what characterized the children in the treatment group in comparison with the children in the intervention schools who were not selected for participation. We therefore compared the scores of the treatment group on the screening measures with those of their classroom peers ($n = 216$). As expected, the children in the treatment group scored significantly lower than their peers on all of the screening measures, with effect sizes ranging from medium to large (Affix knowledge: $d = -.86$, $t = -5.89$, $p = .000$; BPVS II: $d = -.56$, $t = -3.71$, $p = .000$; Orthographic knowledge: $d = .99$, $t = -5.82$, $p = .000$; GMRT-4: $d = -1.02$, $t = -6.41$, $p = .000$). In other words, the results of the analysis show that the teachers were able to identify children who were substantially behind their classmates as indicated by the screening measures.

Analysis of Intervention Effects

The means and standard deviations for the treatment and control groups on all of the measures at pre-test and post-test are shown in Table 4 together with Cronbach's alpha as a measure of test reliability and Cohen's d as a measure of treatment effect. As the table shows, the treatment and control groups were approximately equal on most measures at pre-test. However, although the differences in means in general were low, there was a trend toward an advantage for the control group on the pre-test measures. On two of these measures, word decoding and mathematics, the means for the treatment group were significantly lower than those for the control group (word decoding: $t = -2.18$, $p = .032$; mathematics: $t = -2.71$, $p = .008$).

To evaluate the effects of the intervention, a series of analyses of covariance (ANCOVAs) was conducted. Group differences on the outcome measures at post-test were

assessed using dummy coding of groups in a regression model, controlling for the corresponding pre-test measures and propensity scores as covariates. However, the participants in the present study were clustered within groups, and the groups were clustered within schools. Because single-level regression relies on the assumption of independent residuals, the clustering of data may cause biased estimates of standard errors followed by inflated type I error rates (Snijders & Bosker, 1999). Although the intra-class correlations of the screening measures revealed very low school-level variance, this result does not guarantee the absence of bias caused by clustering in the analyses of intervention effects. To estimate the impact of clustering, we therefore calculated design effects (*deff*) for all of the outcome measures based on the ICC_1 and the average cluster size. Two separate design effects were calculated for each measure using group and school as cluster variables. Four of the measures yielded design effects above the threshold for single-level analysis, as recommended by Satorra and Muthén (1995) (*deff* > 2, ranging from 2.1-4.8: Affix knowledge, WISC-IV Vocabulary, sentence formulation and taught vocabulary). All of the analyses that included these measures were thus conducted using the Huber-White correction in Mplus version 7 (Muthén & Muthén, 1998-2012) to provide robust standard errors and *p*-values. One measure (taught vocabulary) had a design effect above the recommended threshold at both the group level (*deff* = 2.7) and the school level (*deff* = 4.8). Because the Huber-White correction only controls for one level, the level with the highest *deff* was used as cluster variable in the analyses.

The results of the analyses show that there was no statistically significant difference at post-test between the treatment and control groups on measures of nonverbal ($d = -.05$, $t = -1.72$, $p = .088$) and mathematical abilities ($d = .09$, $t = -.18$, $p = .856$). The treatment group, however, made statistically significant gains relative to the control group on measures of affix knowledge, taught vocabulary, sentence formulation and reading comprehension. Thus,

aligned with the rationale of the intervention, the effects were specific to language and reading abilities. As expected, the effect of the intervention was largest for taught vocabulary ($d = 1.77, t = 8.34, p = .000$), but the effect on a measure of sentence formulation was also substantial ($d = .76, t = 2.14, p = .032$). The effect on affix knowledge was moderate in size ($d = .55, t = 3.12, p = .001$), and the effect on a measure of reading comprehension was somewhat smaller ($d = .30, t = 2.09, p = .039$). We were not able to establish statistically significant differences between the groups on a standardized measure of vocabulary ($d = .33, t = .76, p = .447$), a researcher-created measure of morphological knowledge ($d = .08, t = -.13, p = .895$) or measures of decoding (word decoding: $d = .14, t = 1.25, p = .214$; non-word decoding: $d = .09, t = .72, p = .471$).

The use of ANCOVA relies on the assumption of homogeneity of regression slopes between groups. To test whether this assumption was fulfilled, a second series of regression analyses was conducted. In these analyses, Group and Covariate were entered as predictors together with a Group x Covariate interaction term. All of these interaction terms, with the exception of affix knowledge, failed to reach significance ($ts = -.76-1.60, ps = .112-.851$). Thus, for these variables, the assumption of homogeneity of regression slopes holds. In the case of affix knowledge, however, the Group x Covariate interaction term was statistically significant ($t = -2.51, p = .012$), which suggests that group differences at post-test depend on the level of the covariate at pre-test. An analysis in which pre-test scores were centered at the student-level grand mean was thus conducted. This analysis revealed a highly statistically significant group difference in marginal means of 3.12 at the mean level of the covariate in favor of the treatment group ($t = 3.21, p = .001$). This difference increased progressively at lower levels of the covariate, which indicates larger effects for children who started out with low scores on affix knowledge. With initial scores above the mean, however, the effect of intervention gradually decreased to non-significant levels.

Discussion

The primary goal of the present study was to evaluate the effects of a comprehensive word knowledge intervention that aimed to improve children's decoding abilities and linguistic comprehension. The results show promise, especially with regard to enhancing linguistic abilities, which are often characterized by developmental stability and continuity (Melby-Lervåg et al., 2012; Verhoeven & van Leeuwe, 2008). Although the largest statistically significant effect size was found for a researcher-created measure of words taught in the intervention, the effects on a standardized measure of sentence formulation were substantial. For children who started with low levels of morpheme knowledge, there was also a highly significant effect on a morphological measure of near transfer.

Considering the expressive nature of the linguistic measures, caution must be exercised in interpreting these results as an indication of increased linguistic comprehension. However, some argue that linguistic measures that require verbal responses are useful for detecting subtle factors that influence language proficiency (Tilstra, McMaster, Van den Broek, Kendeou, & Rapp, 2009). An expressive measure of vocabulary may, for example, reveal children's understanding of the nuances of word meanings, thus representing a better assessment of the quality of lexical knowledge compared with a vocabulary recognition measure. The effects of intervention on the linguistic measures included in the present study may therefore signal increased knowledge of semantic and morpho-syntactic word features in the treatment group. Thus, this study confirms the utility of comprehensive word knowledge instruction in improving linguistic abilities, as previous studies have demonstrated (Carlo et al., 2004; Lesaux et al., 2010). In other words, the study provides important convergent evidence of the instructional approach across language contexts and learner groups.

The results were less satisfactory for two measures of decoding, for which there was no substantial difference between the treatment group and the control group at post-test.

However, decoding was partly targeted by focusing on morphemic spelling patterns that recur in multiple words. According to the rationale of the study, learning such patterns may consolidate bindings between written and spoken units across words that contain these morphological structures. An experimental measure of reading words with affixes that were included in the intervention may have provided a better test of this hypothesis. Furthermore, as mentioned in the introduction, we decided to focus on the phonological and orthographic properties of larger linguistic units, such as words and morphemes, rather than single phoneme-grapheme correspondences. Even if poor decoders in Norway primarily struggle with reading efficiency, it is possible that explicit instruction in phoneme-grapheme correspondences should have been included as part of the intervention. Either way, the differential effects on the measures of linguistic abilities and decoding suggest that more studies are needed to determine how best to choose, integrate and balance instructional components.

Nevertheless, bearing in mind the lack of intervention effects on decoding, the effect on a transfer measure of reading comprehension is a particularly encouraging finding. Standardized tests of reading comprehension are often insensitive with regard to detecting intervention effects, and intervention studies rarely produce transfer effects on such measures. In a meta-analysis of vocabulary interventions, Elleman, Lindo, Morphy, and Compton (2009) found an overall non-significant effect size of $d = .10$ on standardized measures of reading comprehension. Although relatively small in size ($d = .30$), the obtained effect on reading comprehension in the present study is thus worthy of notice. Several explanations for this promising finding may be offered. Each week, the word knowledge activities in the intervention were contextualized using a short expository text. It is possible that the in-depth reading and discussion of the text may have supported a meaning-oriented reading strategy in the treatment group. However, several other word knowledge interventions have included

similar text-based activities without obtaining statistically significant transfer effects on reading comprehension (e.g., Lesaux et al., 2010). Few studies, though, appear to have focused on the integration of the different features of word identity as intensely as the present study did. The effect of intervention on reading comprehension may therefore be a modest indication of how the quality of different word features and the coherence among them facilitate the retrieval of lexical word identities and their integration into a mental model of the text. Nonetheless, more evidence is needed to confirm this hypothesized explanation.

Several limitations must be considered when drawing conclusions from the present study. Out of consideration for the participating schools, we chose a quasi-experimental research design. Even if the study was strengthened by the fact that there was no attrition, there is a need for an evaluation that uses random assignment to confirm causal inferences concerning intervention effects. Another threat to the internal validity of the study is the use of indirect measures of treatment fidelity. Although the study was small in scale, which allowed for open communication between the teachers and the first author regarding implementation, independent observations of the instruction would have strengthened the conclusions regarding intervention effects.

Moreover, because the study was small in scale, it had limited power to yield statistical significance when the effect sizes were small. It is worth noting that the effects we found on a standardized measure of vocabulary ($d = .33$), and even on a measure of word decoding ($d = .14$), are of a magnitude that is often considered to be of educational value. However, in the present study, we were not able to exclude the possibility that the two effects were due to random chance. A replication of these effects in a study with a larger sample size would provide a more reliable and generalizable estimate of instructional impact.

Another way to evaluate instructional impact is to explore the cumulative effects of the intervention. Considering that the support of cumulative knowledge was an important

principle guiding this intervention, it is unfortunate that we were unable to include measures of long-term effects. It is our belief that further attempts to investigate the present approach to intervention must include an examination of such longitudinal effects.

Notwithstanding these limitations, we argue that this study is of practical and theoretical importance. If interventions are to be adopted by teachers and influence educational practice, then they must be conducted in ecologically valid ways (Lesaux et al., 2010). Considering the diversity of reader profiles that are represented in classrooms and the limited resources available to teachers, instruction that can be applied to heterogeneous groups of readers may often be necessary. The present intervention was designed to accommodate this educational challenge by employing a comprehensive approach to reading remediation – an approach that has previously received relatively little attention in reading research. The effects of intervention on several outcome measures lend support to the utility of this approach. Furthermore, the high satisfaction rate reported in the teacher logs indicates that the intervention was implemented with relative ease. It must be noted that this does not necessarily validate the fidelity of the implementation, but it does suggest that the program was feasible. Although not reported here, an evaluation of the program at the end of the intervention generated overall positive feedback from the participating teachers and the teachers confirmed that the instructions and materials had been easy to use. Feedback from the teachers also suggested that the children enjoyed the program and found the lessons to be motivating. Thus, this study represents an important contribution toward the further development of educationally realistic reading instruction.

However, being educationally realistic is not sufficient for an intervention to be considered “well founded”; it must also be theoretically motivated (Snowling & Hulme, 2011). This study was based on the theoretical position of the Lexical Quality Hypothesis (Perfetti, 2007; Perfetti & Hart, 2002), which places word knowledge as a midpoint where

word identification and comprehension processes meet. Although the Lexical Quality Hypothesis has been influential in reading research, it has not explicitly formed the theoretical basis of many intervention studies. Evidence for the theory primarily comes from correlational studies and small experiments (summarized in Perfetti, 2007). Although the present study was not designed to confirm or refute the Lexical Quality Hypothesis, it represents a novel attempt to turn this theory into practice by formulating recommendations for teaching – which in our opinion should be the ultimate goal of any theory of learning. These recommendations may be further refined and tested, and this study thus provides a springboard for future investigations of this theoretical approach to reading remediation.

References

- Anglin, J. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development*, 58(10), 1-166. doi: 10.2307/1166112.
- Arnbak, E., & Elbro, C. (2000). The effects of morphological awareness training on the reading and spelling skills of young dyslexics. *Scandinavian Journal of Educational Research*, 44(3), 229-251. doi: 10.1080/00313830050154485.
- Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. New York: Guilford.
- Berninger, V. W., & Abbott, R. D. (2010). Listening comprehension, oral expression, reading comprehension, and written expression: Related yet unique language systems in grades 1, 3, 5, and 7. *Journal of Educational Psychology*, 102(3), 635-651. doi: 10.1037/a0019319.
- Botting, N., Simkin, Z., & Conti-Ramsden, G. (2006). Associated reading skills in children with a history of Specific Language Impairment (SLI). *Reading and Writing*, 19(1), 77-98. doi: 10.1007/s11145-005-4322-4.
- Cain, K., & Parrila, R. (2014). Introduction to the special issue. Theories of reading: What we have learned from two decades of scientific research. *Scientific Studies of Reading*, 18(1), 1-4. doi: 10.1080/10888438.2013.836525.
- Caravolas, M., Lervåg, A., Defior, S., Málková, G. S., & Hulme, C. (2013). Different patterns, but equivalent predictors, of growth in reading in consistent and inconsistent orthographies. *Psychological Science*, 24(8), 1398-1407. doi: 10.1177/0956797612473122.
- Carlisle, J. F., McBride-Chang, C., Nagy, W., & Nunes, T. (2010). Effects of instruction in morphological awareness on literacy achievement: An integrative review. *Reading Research Quarterly*, 45(4), 464-487. doi: 10.1598/RRQ.45.4.5.

Carlisle, J. F., & Stone, C. A. (2005). Exploring the role of morphemes in word reading.

Reading Research Quarterly, 40(4), 428-449. doi: 10.1598/RRQ.40.4.3.

Carlo, M. S., August, D., McLaughlin, B., Snow, C. E., Dressler, C., Lippman, D. N., Lively,

T.J., White, C. E. (2004). Closing the gap: Addressing the vocabulary needs of

English-language learners in bilingual and mainstream classrooms. *Reading Research*

Quarterly, 39(2), 188-215. doi: 10.1598/RRQ.39.2.3.

Casalis, S., Deacon, S. H., & Pacton, S. (2011). How specific is the connection between

morphological awareness and spelling? A study of French children. *Applied*

Psycholinguistics, 32(3), 499-511. doi: 10.1017/S014271641100018X.

Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). Language deficits in poor

comprehenders: A case for the simple view of reading. *Journal of Speech, Language,*

and Hearing Research, 49(2), 278-293. doi: 10.1044/1092-4388(2006/023).

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (1999). Language basis of reading and

reading disabilities: Evidence from a longitudinal investigation. *Scientific Studies of*

Reading, 3(4), 331-361. doi: 10.1207/s1532799xssr0304_2.

Catts, H. W., Hogan, T. P., & Fey, M. E. (2003). Subgrouping poor readers on the basis of

individual differences in reading-related abilities. *Journal of Learning Disabilities*,

36(2), 151-164. doi: 10.1177/002221940303600208.

Compton, D. L., Miller, A. C., Elleman, A. M., & Steacy, L. M. (2014). Have we forsaken

reading theory in the name of “quick fix” interventions for children with reading

disability? *Scientific Studies of Reading*, 18(1), 55-73. doi:

10.1080/10888438.2013.836200.

Duff, F. J., Fieldsend, E., Bowyer-Crane, C., Hulme, C., Smith, G., Gibbs, S., & Snowling,

M. J. (2008). Reading with vocabulary intervention: Evaluation of an instruction for

- children with poor response to reading intervention. *Journal of Research in Reading*, 31(3), 319-336. doi: 10.1111/j.1467-9817.2008.00376.x.
- Dunn, L. M., Dunn, D. M., Whetton, C., & Burley, J. (1997). *The British Picture Vocabulary Scale* (2nd ed.). London: GL Assessment.
- Ehri, L. C. (2014). Orthographic mapping in the acquisition of sight word reading, spelling memory, and vocabulary learning. *Scientific Studies of Reading*, 18(1), 5-21. doi: 10.1080/10888438.2013.819356.
- Elleman, A. M., Lindo, E. J., Morphy, P., & Compton, D. L. (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. *Journal of Research on Educational Effectiveness*, 2(1), 1-44. doi: 10.1080/19345740802539200.
- Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, 7(1), 6-10. doi: 10.1177/074193258600700104.
- Hagtvet, B., Helland, B., & Lyster, S.-A. H. (2006). Literacy acquisition in Norwegian. In R. M. Joshi, & P. G. Aaron (Eds.), *Handbook of orthography and literacy* (pp. 15-29). Mahwah, N.J.: Lawrence Erlbaum Associates.
- Kirby, J. R., & Savage, R. S. (2008). Can the simple view deal with the complexities of reading? *Literacy*, 42(2), 75-82. doi: 10.1111/j.1741-4369.2008.00487.x.
- Klinkenberg, J. E., & Skaar, E. (2003). *STAS. Standardisert test i ordavkoding og staving* [Standardized test for word decoding and spelling]. Jaren, Norway: PP-Tjenestens materiellservice.
- Lervåg, A., & Aukrust, V. G. (2010). Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and second language learners. *Journal of Child Psychology and Psychiatry*, 51(5), 612-620. doi: 10.1111/j.1469-7610.2009.02185.x.

- Lervåg, A., Bråten, I., & Hulme, C. (2009). The Cognitive and linguistic foundations of early reading development: A Norwegian latent variable longitudinal study. *Developmental Psychology*, 45(3), 764-781. doi: 10.1037/a0014132.
- 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. doi: 10.1598/RRQ.45.2.3.
- MacGinitie, W., & MacGinitie, R. (2006). *Gates-MacGinitie Reading Tests (4th ed.)*. Iowa City, IA: Houghton Mifflin.
- Marulis, L. M., & Neuman, S. B. (2013). How vocabulary interventions affect young children at risk: A meta-analytic review. *Journal of Research on Educational Effectiveness*, 6(3), 223-262. doi: 10.1080/19345747.2012.755591.
- McArthur, G. M., Hogben, J. H., Edwards, V. T., Heath, S. M., & Mengler, E. D. (2000). On the “specifics” of specific reading disability and specific language impairment. *Journal of Child Psychology and Psychiatry*, 41(7), 869-874. doi: 10.1111/1469-7610.00674.
- Melby-Lervåg, M., Lervåg, A., Lyster, S.-A.H, Klem, M., Hagtvet, B., & Hulme, C. (2012). Nonword-repetition ability does not appear to be a causal influence on children’s vocabulary development. *Psychological Science*, 23, 1092-1098. doi: 10.1177/0956797612443833.
- Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge: Cambridge University Press.
- Morris, R. D., Lovett, M. W., Wolf, M., Sevcik, R. A., Steinbach, K. A., Frijters, J. C., & Shapiro, M. B. (2012). Multiple-component remediation for developmental reading

- disabilities: IQ, Socioeconomic status, and race as factors in remedial outcome. *Journal of Learning Disabilities*, 45(2), 99-127. doi: 10.1177/0022219409355472.
- Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- Nagy, W. (2005). Why vocabulary instruction needs to be long-term and comprehensive. In E. H. Hiebert, & M. L. Kamil (Eds.), *Teaching and learning vocabulary: Bringing research to practice* (pp. 27-44). Mahwah, N.J.: Lawrence Erlbaum Associates.
- Nagy, W. E., & Herman, P. A. (1984). Limitations of vocabulary instruction (Tech. Rep. No. 326). Urbana: University of Illinois, Center for the Study of Reading.
- Nation, K., Cocksey, J., Taylor, J. S., & Bishop, D. V. (2010). A longitudinal investigation of early reading and language skills in children with poor reading comprehension. *Journal of Child Psychology and Psychiatry*, 51(9), 1031-1039. doi: 10.1111/j.1469-7610.2010.02254.x.
- Nation, K., & Snowling, M. J. (1998a). Individual differences in contextual facilitation: Evidence from dyslexia and poor reading Comprehension. *Child Development*, 69(4), 996-1011. doi: 10.1111/j.1467-8624.1998.tb06157.x.
- Nation, K., & Snowling, M. J. (1998b). Semantic processing and the development of word-recognition skills: Evidence from children with reading comprehension difficulties. *Journal of Memory and Language*, 39(1), 85-101. doi: 10.1006/jmla.1998.2564.
- Nunes, T., Bryant, P., & Bindman, M. (1997). Backward readers' awareness of language: Strengths and weaknesses. *European Journal of Psychology of Education*, 12(4), 357-372. doi: 10.1007/BF03172798.
- Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, 11(4), 357-383. doi: 10.1080/10888430701530730.

- Perfetti, C., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy* (pp. 67-86). Amsterdam: John Benjamins Publishing.
- Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, 18(1), 22-37. doi: 10.1080/10888438.2013.827687
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55. doi: 10.1093/biomet/70.1.41.
- Rosenthal, J., & Ehri, L. C. (2008). The mnemonic value of orthography for vocabulary learning. *Journal of Educational Psychology*, 100(1), 175-191. doi: 10.1037/0022-0663.100.1.175.
- Rosenthal, J., & Ehri, L. C. (2011). Pronouncing new words aloud during the silent reading of text enhances fifth graders' memory for vocabulary words and their spellings. *Reading and Writing*, 24(8), 921-950. doi: 10.1007/s11145-010-9239-x.
- Satorra, A., & Muthén, B. O. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, 25, 267-316. doi: 10.2307/271070.
- Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical evaluation of language fundamentals (CELF - 4th edition)*. San Antonio, TX: Harcourt Assessment.
- Seymour, P. H. K. (2005). Early Reading Development in European Orthographies. In M. J. Snowling & C. Hulme (Eds.), *The Science of Reading: A Handbook* (pp. 296-315). Malden: Blackwell Publishing. doi: 10.1002/9780470757642.ch16
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage Publishing House.

- Snowling, M. J., & Hulme, C. (2011). Evidence-based interventions for reading and language difficulties: Creating a virtuous circle. *British Journal of Educational Psychology*, 81(1), 1-23. doi: 10.1111/j.2044-8279.2010.02014.x.
- Snowling, M. J., & Hulme, C. (2012). Interventions for children's language and literacy difficulties. *International Journal of Language & Communication Disorders*, 47(1), 27-34. doi: 10.1111/j.1460-6984.2011.00081.x.
- Snowling, M. J., Muter, V., & Carroll, J. (2007). Children at family risk of dyslexia: A follow-up in early adolescence. *Journal of Child Psychology and Psychiatry*, 48(6), 609-618. doi: 10.1111/j.1469-7610.2006.01725.x.
- Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly*, 16(1), 32-71. doi: 10.2307/747348.
- Tilstra, J., McMaster, K., Van den Broek, P., Kendeou, P., & Rapp, D. (2009). Simple but complex: Components of the simple view of reading across grade levels. *Journal of Research in Reading*, 32(4), 383-401. doi: 10.1111/j.1467-9817.2009.01401.x.
- Tong, X., Deacon, S. H., Kirby, J. R., Cain, K., & Parrila, R. (2011). Morphological awareness: A key to understanding poor reading comprehension in English. *Journal of Educational Psychology*, 103(3), 523-534. doi: 10.1037/a0023495.
- Torgesen, J. K., Wagner, R., & Rashotte, C. (1999). *Test of word reading efficiency*. Austin, TX: Pro-Ed.
- Vadasy, P. F., Sanders, E. A., & Peyton, J. A. (2006). Paraeducator-supplemented instruction in structural analysis with text reading practice for second and third graders at risk for reading problems. *Remedial and Special Education*, 27(6), 365-378. doi: 10.1177/07419325060270060601.

- Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D. M. (2004). Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of Child Psychology and Psychiatry*, 45(1), 2-40. doi: 10.1046/j.0021-9630.2003.00305.x.
- Verhoeven, L., & van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology*, 22(3), 407-423. doi: 10.1002/acp.1414.
- von Koss Torkildsen, J., Dailey, N. S., Aguilar, J. M., Gómez, R., & Plante, E. (2013). Exemplar variability facilitates rapid learning of an otherwise unlearnable grammar by individuals with language-based learning disability. *Journal of Speech, Language, and Hearing Research*, 56(2), 618-629. doi: 10.1044/1092-4388.
- Wechsler, D. (2003). *Wechsler intelligence scale for children* (4th ed.). San Antonio, TX: The Psychological Corporation.
- Wechsler, D. (2005). *Wechsler individual achievement test* (2nd ed.). San Antonio, TX: The Psychological Corporation.
- Westwood, P., Harris-Huges, M., Lucas, G., Nolan, J., & Scrymgeour, K. (1974). One minute addition test-one minute subtraction test. *Remedial Education*, 9, 70-72.
- Wolter, J. A., & Dilworth, V. (2014). The effects of a multilinguistic morphological awareness approach for improving language and literacy. *Journal of Learning Disabilities*, 47(1), 76-85. doi: 10.1177/0022219413509972.

Table 1

Sample Demographics Assessed at Pretest

Group	Gender (M:F)	Age (SD) (months)	Age range (months)	Grade level (third:fourth)
Control ($n = 59$)	35:24	107 (6.71)	94-118	22:37
Treatment ($n = 59$)	35:24	106 (6.70)	91-118	22:37

Table 2

Description of Lesson Cycle

	Lesson content	Time on task*
Day 1	1. Introduce theme of the week and target words	1. 10 min
	2. Guided reading of text	2. 15 min
	3. Discussion of text	4. 20 min
	4. Word knowledge activities	5. 15 min
Day 2	1. Repetition of theme of the week and target words	1. 5 min
	2. Word knowledge activities	2. 40 min
	3. Repeated reading of text	3. 15 min
Day 3	1. Repetition of theme of the week and target words	1. 5 min
	2. Word knowledge activities	2. 25 min
	3. Repeated reading of text	3. 15 min
	4. Game adapted to fit the theme of the week (e.g., Alias, Memory and Word Bingo)	4. 15 min

* time is approximate

Table 3

Examples of word knowledge activities

Semantic activities	Syntactic activities	Morphological activities
<ul style="list-style-type: none"> • Discussion of pictures illustrating target word and concepts with semantic boundaries to target word. Which pictures illustrate target word, which do not and why? (e.g., does a picture of a boxing match depict a <i>conflict</i>*) • Odd man out (e.g., which word is not an <i>object</i>* and why?: [microphone, <u>music</u>, guitar, flute]) • Work with semantically related words (synonyms, antonyms, subordinates and superordinates) 	<ul style="list-style-type: none"> • Sentence repair (discussion of ways to repair a sentence that is syntactically wrong) • Sentence puzzles (construction of sentences using words as pieces in a puzzle) • Missing words (different morphological forms of target word have been removed from various sentences. Which word form belongs to which sentence, and where in the sentence does it fit?) 	<ul style="list-style-type: none"> • Word sorts (find words belonging to the same word family as target word/find words with same affix) • Word puzzle (create different words by combining prefix, root, suffix/create compound words with target word) • Morphological problem solving (identify the meaning of novel words by using word parts) • Cloze tasks with different derivations of target word

* target words

Table 4

Means (Standard Deviations) and Reliability of Outcome Measures at Pre-Test and Post-Test Together with Cohen's d as a Measure of Treatment Effect

Outcome measure (max score)	Reliability	Pre-test		Post-test		Cohen's d
		Control	Treatment	Control	Treatment	
Nonverbal reasoning (35)	.84	16.59 (4.82)	15.44 (3.81)	18.92 (4.48)	17.54 (3.80)	-.05
Mathematics (60)	.89	33.80 (12.06)	28.37 (9.56)	31.20 (11.42)	26.81 (9.62)	.09
Word analogy (30)	.82	16.39 (5.18)	15.95 (5.60)	20.00 (5.14)	19.98 (4.64)	.08
Affix knowledge (50)	.81	31.34 (6.90)	32.58 (7.26)	35.81 (7.61)	40.95 (6.17)	.55**
Sentence formulation (44)	.80	25.20 (7.27)	22.56 (7.71)	25.71 (7.17)	28.76 (6.74)	.76*
Vocabulary (70)	.67	17.36 (3.74)	17.31 (4.16)	17.20 (4.27)	18.47 (4.47)	.33
Taught vocabulary (50)	.71	9.95 (5.68)	10.56 (5.62)	9.68 (6.01)	20.29 (7.65)	1.77***
Word decoding (208)	.95 ^a	83.68 (28.67)	73.54 (21.38)	88.59 (29.50)	81.97 (24.79)	.14
Non-word decoding (126)	.78 ^a	46.41 (17.86)	41.20 (12.76)	48.85 (20.91)	45.03 (14.32)	.09
Reading comprehension (126)	.89	51.54 (14.07)	47.66 (10.65)	59.44 (13.40)	59.27 (14.17)	.30*

Note. All reliability coefficients are Cronbach's alpha at pre-test, except when otherwise indicated. Cohen's d was calculated by dividing

the difference in raw gain scores between the treatment group and control group by the pooled pre-test standard deviation. Group

differences tested with analysis of covariance: * significant at $p < .05$, ** significant at $p < .01$, *** significant at $p < .001$.

^a = correlation between alternate forms at pre-test.

Author statement

Special thanks to the Department for Children and Youth at the Oslo Public Library for help finding texts to use in the intervention. We also gratefully acknowledge the children, teachers and schools for their participation in the study.

Ellen Irén Brinchmann (corresponding author) is a PhD student in the Department of Special Needs Education at the University of Oslo, Norway; e-mail: e.i.brinchmann@isp.uio.no

Hanne Næss Hjetland is a PhD student in the Department of Special Needs Education at the University of Oslo, Norway; e-mail: h.n.hjetland@isp.uio.no

Solveig-Alma Halaas Lyster is a professor in the Department of Special Needs Education at the University of Oslo, Norway; e-mail: s.a.h.lyster@isp.uio.no