Learning and Instruction 58 (2018) 65–79

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

Effectiveness of an early reading intervention in a semi-transparent orthography: A group randomised controlled trial

Oddny Judith Solheim[a,*], Jan C. Frijters[a,b], Kjersti Lundetræ[a], Per Henning Uppstad[a]

[a] Norwegian Centre for Reading Education and Research, University of Stavanger, Norway
[b] Department of Child and Youth Studies, Brock University, Canada

ARTICLE INFO

Keywords:
Early intervention
Reading difficulties
GraphoGame
At risk children
Computer-assisted intervention

ABSTRACT

This study reports on the effectiveness of an early reading intervention, based on current research on early reading acquisition, and aligned to Norwegian orthography. Thirteen schools were randomly assigned to one of two interventions or a control condition. First grade students (n = 744) were screened at school entry, and children at risk of reading difficulties (n = 140) were identified. At-risk students in schools allocated for intervention received comprehensive teacher-led instruction also containing an individually-delivered computer component. The only difference between the two interventions was whether the computer application had adaptive learning features. Both interventions had significant impact on reading and spelling with no significant difference between the two different intervention conditions. Findings indicate that Norwegian children identified to be at-risk at school entry can profit from intensive intervention that combines training in letter knowledge with explicit instruction in phonetic decoding and word recognition, free spelling, connected text reading and shared reading.

1. Introduction

Over the past decades, we have witnessed encouraging advances in how to intervene effectively in early reading development for children at risk of reading difficulties (RD). Yet, for many groups of learners we are basing our instructional decisions on research from disparate populations, including wide variation in language types. Most reported reading interventions, however, have been carried out in English-speaking countries, with English being a highly opaque orthography (Galuschka, Ise, Krick, & Schulte-Körne, 2014; Share, 2008). A body of research from Finland (e.g. Richardson & Lyytinen, 2014; Saine, Lerkkanen, Ahonen, Tolvanen, & Lyytinen, 2011) also informs this work, with Finnish being a highly transparent language.

Problems experienced by children with dyslexia and those struggling to acquire reading skills, have been consistently related to phonological processing and word-level reading (Ehri, Nunes, Stahl, & Willows, 2001; Stanovich & Siegel, 1994). Even so, the ease of becoming phonemically aware and learning to read have been found to relate to the "orthographic depth" of a language, (Katz & Frost, 1992; Seymour, Aro, & Erskine, 2003). Consequently, it is easier to learn to read accurately in orthographies with consistent graphene-phoneme correspondents (i.e. transparent orthographies) than in orthographies containing many inconsistencies and complexities (i.e. opaque orthographies). Further, the phenotype of dyslexia varies with the transparency of the language. While, *reading accuracy* has been observed as a core problem in opaque orthographies, *reading fluency* seems to be the most consistently observed problem across languages. (Landerl, Wimmer, & Frith, 1997; Ziegler, Perry, Ma-Wyatt, Ladner, & Schulte-Körne, 2003). An urgent question is thus whether at-risk students learning to read more transparent orthographies also profit from intervention in early phases of reading development, i.e. when the main goal of the instruction is teaching students to read accurately. In order to answer this question, studies from a wider range of orthographies are needed (Galuschka et al., 2014). The present study was undertaken in a Norwegian context. According to Seymour et al. (2003), Norwegian orthography is considered a semi-transparent orthography (i.e. more transparent than English, but less transparent than Finnish). Norwegian children start school the year they turn six years old, and there is no formal instruction in kindergarten (NMER, 2011). The goal of this study was to identify children at risk for RD at school entry, then develop and evaluate the effect of an early reading intervention informed by previous research.

1.1. *What characterises effective early reading interventions?*

Effective reading interventions (implemented *alongside* formal

* Corresponding author. Norwegian Centre for Reading Education and Research, University of Stavanger, post-box 8600 Forus, 4036 Stavanger, Norway.
E-mail address: oddny.j.solheim@uis.no (O.J. Solheim).

https://doi.org/10.1016/j.learninstruc.2018.05.004
Received 26 July 2017; Received in revised form 14 March 2018; Accepted 13 May 2018
Available online 31 May 2018

reading instruction) typically involve letter-sound correspondences, phonics instruction on how to use these correspondences to read and spell words (Ehri et al., 2001; Galuschka et al., 2014; McArthur et al., 2012) and/or comprehension strategies (Scammacca, Vaughn, Roberts, Wanzek, & Torgesen, 2007; Suggate, 2010). In a meta-analysis of interventions for at-risk and struggling readers in preschool through Grade 7, Suggate (2010) makes a distinction between studies that explicitly target (a) phonemic awareness, (b) phonics, (c) comprehension or (d) a mixture of these, and concludes that phonics interventions were more effective until Grade 1. After Grade 1, comprehension interventions and mixed interventions tended to be most effective. Also, Scammacca et al. (2007) examined interventions that provided 100 sessions or more in Grades K-3, and concluded that all successful interventions included explicit instruction in phonemic awareness and phonetic decoding, along with practice in reading texts and comprehension instruction.

Relatively less evidence is available concerning long-term outcomes following early intervention (Lovett, Barron, & Frijters, 2013; Suggate, 2016). Children already significantly behind their peers in reading in Grade 1 and 2 do not improve relative to their peers in the absence of intervention (Cirino et al., 2002), and some have asserted that relative decline is more the norm (Stanovich, 1986). A few studies have shown that early intervention can result in a persisting benefit over time (e.g., Blachman et al., 2004, 2014; Fälth, Gustafson, Tjus, Heimann, & Svensson, 2013; Lovett et al., 2017; Morris et al., 2012; Saine et al., 2011), but such effects tend to be small, and do not move children's relative skill within their peer group (e.g., a child who is at the 15th percentile at the end of intervention, will remain at the 15th percentile in the years after intervention). Studies that have included long-term follow up, on average have only tracked children for 1-year post-intervention. Of the 71 studies reviewed by Suggate (2016), the mean time to follow up was 11.7 months.

Intervention intensity (e.g. size of the instructional group, how frequently intervention is provided, length of each session, duration of intervention, knowledge and experience of the teacher) has also been found to affect efficacy (Vaughn, Denton, & Fletcher, 2010). Even so, recent meta-analyses (Galuschka et al., 2014; Suggate, 2010, 2016) have failed to document a moderating effect for such intervention features. However, the authors of these meta-analyses suggest that the absence of moderating effects might be a consequence of intervention features typically being confounded, which in turn would reduce the observed association between such moderators and outcomes.

1.2. Computer-assisted interventions

The general rationale for early remediation of children with RD encompasses intensified scaffolding and support directed towards their specific skill needs. As such, computer-assisted interventions (CAI) have attracted major interest in the field because of its capacity to provide highly specialized instruction, with considerations for immediate reinforcing feedback, item difficulty optimized for observed ability, and high levels of engagement (Corbett & Anderson, 1995; Corbett, 2001). Possible advantages have been especially emphasized in early phases of reading instruction and for children at risk for or with RD (Cheung & Slavin, 2012, 2013). In a best-evidence synthesis on the effects on educational technology applications on reading outcomes for struggling readers in Grades 1–6, Cheung and Slavin (2013) report positive, but small, effects on reading skill. Studies reveal higher effect sizes when computer software is provided in small groups (Cheung & Slavin, 2013; Torgesen, Wagner, Rashotte, Herron, & Lindamood, 2010).

There are two main ways in how game progression is implemented in CAIs: a fixed and an adaptive approach. According to research on neuroplasticity, long lasting cognitive benefits can be mediated by two primary design elements: continuous feedback and adaptation to the in-moment performance (Mishra & Gazzaley, 2014). Holding the element of continuous feedback constant, a distinction can be drawn between applications that have a fixed presentation, following a strict sequence of increasing difficulty to be mastered, and adaptive applications that provide the player with content based on the player's actual performance. The preference for an adaptive approach has been based on the increased potential for facilitating engagement by simultaneously providing students with both sufficient challenge and ample opportunity for success (Richardson & Lyytinen, 2014).

1.3. Objectives of the study

Previous contributions to the literature on early intervention for struggling readers have occurred in orthographies at the extreme ends of the transparency-continuum (e.g., English vs Finnish). These studies have only indirectly informed construction of efficient interventions in *semi*-transparent orthographies, such as Norwegian. The present study sought to investigate the efficacy of an early reading intervention delivered alongside formal reading instruction to Norwegian 6-year olds identified to be at risk for RD at school entry through a group randomised controlled trial, with a two-year long term follow up after intervention.

When designing the intervention, we emphasized intensity, structured content and explicit instruction. Through training of grapheme-phoneme correspondences, word reading, text reading, free spelling, and reading comprehension, reading skill was enhanced from different angles. Every session included alternations between the parts (letters or words) and the whole (words or texts), a process that in the present study is considered to be a hermeneutic approach to reading and writing (Lundetræ, Solheim, Schwippert, & Uppstad, 2017; Tønnessen & Uppstad, 2015). Hermeneutics is a general theory about interpretation and meaning, that has been particularly prominent on the European Continent (see Gadamer, 1960). This theory also applies to the learning of literacy: frequent meetings with a letter in words promotes knowledge of the letter and how it is applied and pronounced in different words. Conversely, reading or spelling a word promotes the child's knowledge about the word, but also about the letters it consists of.

The intervention was teacher-delivered, and included guided reading; free spelling; shared reading as well as a CAI component. The computer applications were used for training letter knowledge and word reading, and were custom-programmed for the present study. We implemented a Norwegian version of the learning environment GraphoGame (Richardson & Lyytinen, 2014) as well as a competing game: 'On Track ABC', without individual adaptation, but with 3 levels of increasing difficulty within 3 of 5 mini-games. Although some studies have compared the efficacy of playing different versions of GraphoGame (see e.g. Kamykowska, Haman, Latvala, Richardson, & Lyytinen, 2014; Kyle, Kujala, Richardson, Lyytinen, & Goswami, 2013) no previous study has, to our knowledge, targeted the *adaptive* feature by comparing the efficacy of GraphoGame to other reading software without this adaptive feature.

The research questions were as follows:

1. Is there a short-term effect of the On Track interventions in Grade 1 on word reading, sentence reading and spelling?
2. Is the intervention including adaptive gaming software (GraphoGame) more effective than the intervention including a non-adaptive software (On Track ABC app)?
3. Are the possible effects of the On Track interventions sustained over the long term?

2. Method

This study is part of a longitudinal project, On Track, investigating the effect of early intervention for children at risk for RD (Lundetræ et al., 2017). A total of 1199 students from 19 primary schools in the Southwest of Norway were enrolled in the study in 2014. The schools

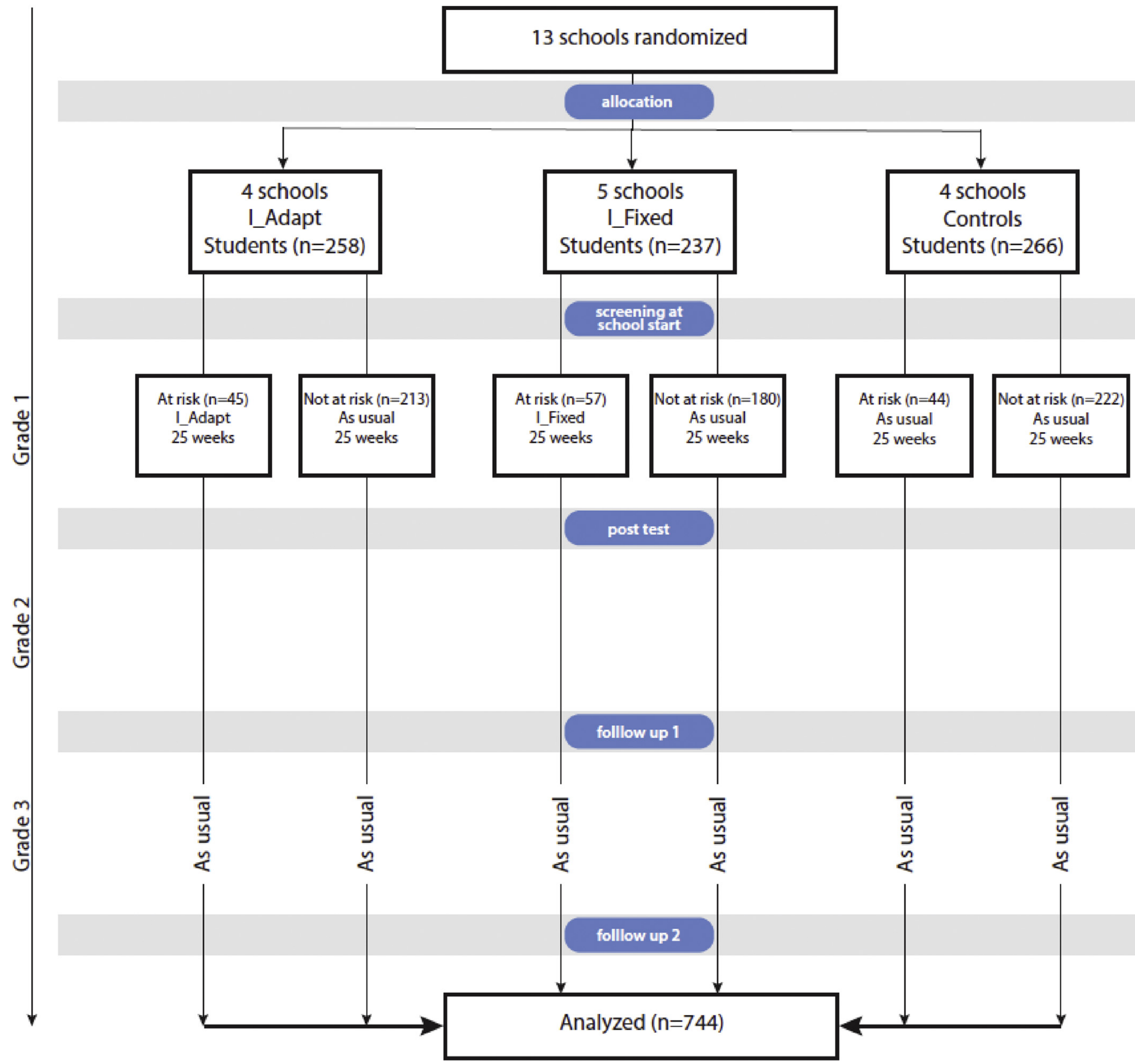


Fig. 1. Diagram showing the flow of participants through the trial.

met the following three criteria: 1) more than 40 children were expected to be enrolled in Grade 1 in 2014, 2) the schools' score on the national reading test in Grade 5 had to be close to the national mean in two of the three previous years, and 3) all letter-sound correspondences had to be introduced by Easter in Grade 1. Of the children enrolled in these schools, parental concent for participation was given for 97.7 percent of the students. The present study includes students from 13 schools that were randomised to one of two different interventions or an enhanced business-as-usual control condition in Grade 1 (see Fig. 1 for study design).

2.1. Participants

First grade students (n = 744) from 13 primary schools (27 classes) participated in the intervention in the present study. Children were on average 6.2 years old when the study began at school entry (range 5.5–6.7), with 49.2% being boys. In Norway, 96.7 percent of primary-school students are enrolled in public (i.e. non-private) schools (NDET, 2015a), and only 0.6 percent of primary-school students are enrolled in special schools (NDET, 2015b). To obtain a sample for the On Track project that mirrors Norwegian classrooms, all students successfully recruited to the study were included in all study components, including intervention and longitudinal tracking. The same inclusion criteria were also applied to L2 children and children with special needs, the only exception being five students with severe cognitive disabilities who were deemed highly unlikely to benefit from participation.

2.2. Procedure

Entire schools were randomly assigned to one of two intervention conditions (I_Adapt or I_Fixed), or an enhanced business-as-usual control condition. Prior to intervention, all teachers in both control and intervention schools received a four hour course in best practice in literacy instruction in order to provide teachers with similar level of

Table 1
Characteristics of at-risk students in the three conditions.

	I_Adapt	I_Fixed	Control
N	43	53	44
Percent at-risk within school	16.9	23.0	16.9
Percent male	60.5	60.4	56.8
Percent minority students	28.6	23.1	26.2
Percent with familial risk for RD	7.0	15.1	13.6

Note: No significant differences in proportions were observed across the three conditions on any of these characteristics.

access to research on best practice.

All students (n = 744) were screened during the first four weeks after school entry in Grade 1. Parents answered a questionnaire relating to demographics, home literacy environment, familial risk of RD, the student's language background, and child health. The questionnaire was administered by the schools, and response rate was 97.5 percent.

At-risk students (n = 140) were identified by combining scores on letter knowledge, RAN, phoneme isolation, and phoneme blending at school start into a student at-risk index (see description of measures under independent variables). Falling below the 30th percentile on any one of these tests yielded one risk point. A student also gained one risk point if they had familial risk for dyslexia i.e. if parents reported that at least two of the student's close relatives (mother, father or sibling) had RD. Students who scored at least three risk points were considered to be at risk of RD (18.8 percent of the total sample). At-risk status was communicated to schools. Table 1 summarizes the characteristics of participants enrolled in the intervention conditions.

From September through April, at-risk children in both control and treatment conditions were treated in groups of 3–7 students. The composition of the groups were decided by the teachers themselves. Intervention groups were taught by the same teachers throughout the intervention period.

The interventions started after seven weeks of formal schooling and consisted of four weekly 45-min sessions during a period of 25 weeks. The at-risk students received the intervention when the class was divided into small groups for literacy centers/station teaching in reading and writing, meaning that no students were taken out of the classroom when the class as a whole received instruction from the teacher. Consequently, at-risk students did not receive extended time of literacy instruction, but rather a specific content delivered in achievement level groups. The On Track intervention programme was developed by the research team and implemented by school teachers who already worked in the participating schools. At-risk students in the enhanced control condition received literacy instruction to the same extent, and organized the same way as the at-risk students in the intervention schools, but the content was after the best of their teachers' ability.

Posttests were administered to the full sample (i.e. including students who were identified not to be at-risk at school entry) within two weeks after the intervention was finished.

Taken together, the comparison group in the present study represents a strong control. The teachers in the control schools participated a four-hour course on best practice in literacy instruction prior to intervention and received information about students' risk status (i.e., based on school entry scores). The at-risk controls also received training in small groups 4 × 45 min a week, for 25 weeks. The control condition thus deviated from regular practice in Norwegian classrooms concerning both identification, organization, and intensity of early training, and therefore could be labelled *enhanced* control. *Business-as-usual* refers to the *content* in the small group instruction, which was developed by the teachers themselves. After the Grade 1 intervention was completed, the control schools were provided with all the material that had been used in the On Track intervention, and were free to use this in the next cohort of students.

2.3. Measures

All students were assessed prior to and after intervention. Follow-up was administered at end of Grade 2 and 3 (i.e. 1 and 2-year follow up).

2.3.1. Pretests

Students were tested individually during school hours in a 20–30 min session in a quiet room in their respective schools. The tests were administered by a team of 18 testers, all trained in psychometric assessment. The team received six hours of training in the test battery prior to data collection, and were issued with a written test manual. The test battery was administered on a Lenovo Yoga Tablet 10, running Android 4.2. For all tests, student responses were scored and recorded on the tablet. Before each test, training items were administered to ensure that the student fully understood the task. Test items were administered without corrective feedback. To prevent scoring errors, the tester had to confirm a student's response on the tablet, before moving on to the next item. The test results from the tablets were generated and stored as Microsoft Excel files.

Letter-sound knowledge was measured using a 15-item multiple-choice test. The stimulus was a pre-recorded letter sound, and the student was to identify the corresponding upper-case letter. The student responded by pressing one of four letters appearing on the screen. Reliability in the overall sample Cronbach's alpha was .85 (at-risk sample reliability = .61). A reason for the lower reliability for the at-risk sample was the inconsistent letter-sound knowledge of at-risk children.

Phonemic awareness was measured by means of eight phoneme-isolation and eight phoneme-blending tasks. Both tasks were ordered by difficulty (easiest first) and was automatically discontinued after two subsequent errors. The phoneme-isolation tasks required the student to isolate and pronounce the initial phoneme in words. Students responded orally, and the tester scored the response on the tablet. Reliability in the present sample (Cronbach's alpha) was 0.92 (for at-risk, $\alpha = 0.79$).

The *phoneme-blending* task required the student to blend a sequence of phonemes into a word. The presentation of stimuli was pre-recorded to ensure that the pronunciation and time interval of the phonemes would be consistent (one phoneme per second) across presentations: "Here you see pictures of /ri/,/rips/,/ris/, and /ring/ [English: 'ride', 'redcurrant', 'rice', 'ring']. Listen carefully and press the picture that goes with:/r//i//s/". The tester pointed at the objects shown in the pictures as they were named. Students responded by pressing one of the four pictures. Reliability in the present sample (Cronbach's alpha) was 0.87 (for at-risk, $\alpha = 0.74$).

Word reading was measured with 8 words ranging from easy to difficult presented on the tablet, one at a time. The words represented a variety of letters and letter sequences (CV, VC, CVC, VCC, VCC, CVCV, CVCC, CVCVC). The students were asked to read the word aloud. If they managed to sound out or recognize the word, the tester scored it as correct on the tablet. The test was discontinued after two subsequent errors. Reliability (Cronbach's Alpha) was 0.92 (for at-risk, $\alpha = 0.19$). A reason for the lower reliability for the at-risk sample was that they still hadn't begun to acquire reading skills.

RAN (rapid automatized naming) test required naming familiar objects presented simultaneously on a white background in random order. The stimuli were illustrations of the monosyllabic Norwegian words for 'sun', 'car', 'plane', 'house', 'fish' and 'ball'. Twenty stimuli were presented in a 4 × 5 matrix, with a unique matrix presented for each of two trials. The student was asked to name each stimulus as quickly and accurately as possible, working from left to right and top to bottom. A practice session ensured that the student could name all the objects and understood the task. For each trial, both the completion time (in 1/100ths of a second) and naming errors were recorded.

Vocabulary was measured using an abridged version (20 of the original 40 items) of the Norwegian Vocabulary Test (NVT), which is designed for children aged 5–6 (Størksen, Ellingsen, Tvedt, & Idsøe, 2013). Students were to name pictures that appeared on the screen. Reliability (Cronbach's alpha) for the 20 items in the present sample was .82 (at-risk, $\alpha = 0.80$).

2.3.2. Posttests

All posttests were administered as a paper and pencil test by trained testers in groups (max 16) on two successive days (45-60 minute-sessions). All post-measures were adopted from a mandatory Norwegian screening battery developed to identify children at risk for RD towards the end of Grade 1 (NDET, 2015c).

The *word reading* task consisted of 14 items. Each item consisted of a picture followed by four visually similar words, whereof one corresponded to the picture. The child was asked to read the words as fast as possible and to check the word that matched the picture. For example, a picture of a fish ('fisk' in Norwegian) followed by 'fiske', 'fikse', 'fiks' and 'fisk'. The correct stimuli were presented in a random order. The time limit for the word reading task was 5 min, and maximum score was 14. Reliability in the present sample (Cronbach's alpha) was 0.78 (for at-risk, $\alpha = 0.86$).

The *sentence reading* task consisted of 10 items in which a written sentence was followed by four pictures. The child was asked to choose the picture that best corresponded to the meaning of the sentence. The correct stimuli were presented in a random order. Maximum score was 10, and the time limit was five minutes. The reliability in the present sample (Cronbach's alpha) was 0.86 (for at-risk, $\alpha = 0.86$).

Spelling was measured by 14 items. For all items, the tester read a short sentence containing the target word, for example, "It was a difficult test. Write 'test'". The number of correctly spelled words was measured (maximum = 14). Reliability in the present sample (Cronbach's alpha) was 0.84 (for at-risk, $\alpha = 0.84$).

2.4. Intervention

Each of the 100 On Track sessions consisted of four lessons, each lasting for ten minutes: ABC, Guided Reading, Free spelling, and Shared Reading (for more detailed information on learning objectives and materials, see Table 2). In addition, teachers were given a total of five minutes for switching from one activity to the next. A detailed manual scripted each lesson. The only difference between the two interventions was the type of computer application used in the ABC lessons (10 min per session). Students in the I_Adapt group used the Norwegian version of GraphoGame whereas students in the I_Fixed group used the On Track ABC app.

The Norwegian version of GraphoGame consists of nine different game formats with immediate feedback and a motivational reward system. The content adapts to the individual player according to actual performance in identifying letters, syllables or words matching the auditory stimulus appearing in their headphones. The adaptation algorithm of the game ensures that the student always receives 20% of trials as challenge and 80% as mastery, based on the individual player's previous performance. As such, the game does not include a fixed set of levels, but provides unique levels according to in-the-moment skill. At a certain proficiency level the algorithm provides timed target items and distractors, pushing the player to faster identification. The reward system is managed via a personal avatar, created at the very start of the game. With successful completion of levels, the player collects coins, which can be exchanged to clothes and gear for the avatar.

The On Track ABC app consisted of five mini games: 1) matching upper case and lower-case letters in a memory game; 2) matching one out of five objects with a letter by identifying first sound; 3) matching a letter sound with the correct letter (one out of seven); 4) spelling; and 5) word reading. Memory, spelling and word reading had three difficulty levels, and teachers guided students in choosing the appropriate difficulty level. The students used headphones, and if they tapped an object or a letter in the app, they could hear the word or the letter sound. The On Track ABC app provide immediate feedback in the shape of an enthusiastic child's voice, but involves no performance-based reward system. Features of each CAI are summarized in Table 3.

2.5. Treatment fidelity

Treatment fidelity was monitored in several ways. The school teachers attended two days of training by the research team prior to the intervention delivery and received a detailed manual that scripted each of the 100 lessons. There was no individual supervision of teachers during the intervention, but one of the researchers in the project team observed each teacher during one session in the fall and one session in the spring. After the first observation, all intervention teachers were invited to a meeting in order to discuss challenges and get mutual feedback based on the team's observations. The teachers recorded student attendance across the 100 sessions (mean student attendance was 93.4 session, SD = 4.3). One of the intervention schools reported serious violations to the intervention protocol (e.g. doubling the group size of the intervention group on frequent occasions). The school was removed from the final analysis. Information about the intervention sessions in enhanced control schools were collected through focus group interviews in each control school. The teachers responsible for

Table 2
Learning objectives and materials in the On Track interventions.

Lesson	Learning objectives	Materials
ABC (GraphoGame) 10 min.	Letter knowledge. Linking graphemes and phonemes. Word-reading skills.	Digital tablet and headphones. GraphoGame—a highly adaptive play-like app where the students master about 80 percent of the tasks.
ABC (On Track ABC) 10 min.	Letter knowledge. Linking graphemes and phonemes. Word-reading skills. Phonemic analysis.	Digital tablet and headphones. The On Track ABC—a play-like app with five mini games. Different levels to be chosen by the student/teacher.
Guided Reading 10 min.	Word-reading skills. Decoding. Word recognition. Reading comprehension. Awareness of challenging orthographic patterns.	Easy readers with increasing difficulty (a new book each session). Everyone receives a copy of the same book. Detailed instructions pertaining to each book given in the teacher's manual.
Free spelling 10 min.	Promote reading skills through writing. Enhance phonemic analysis. Promote spelling skills. Letter knowledge.	Headphones and writing software (School Font) on the tablet. The students heard the letter sound when pressing a letter key, and the full word when pressing the spacebar. The writing tasks were described in the teacher's manual; they were often linked to books used in Guided Reading or Shared Reading. Spelling accuracy according to orthographic conventions was not emphasized. The texts were printed on paper after each session.
Shared Reading 10 min.	Experiences with children's literature, written language, text structure, and comprehension strategies. Promote positive attitudes towards reading.	Children's books (picture books, children's poetry and chapter books). Suggestions for how teachers could talk about the texts and words were given in the teacher's manual.

Table 3
Comparison of gaming content in the Norwegian version of GraphoGame and On Track ABC app.

Gaming content	GraphoGame	On Track ABC
Matching upper case and lower case letters		x
Identify first sound in spoken words		x
Letter-sound correspondences	x	x
One syllable words	x	x
Multi-syllable words	x	x
Words with complex graphemes	x	x
Words with diagraphs	x	x
Words with doubling of consonants	x	x
Spelling		x
Timed reading	x	
Immediate feedback	x	x
Reward system	x	

delivering the instruction explained that they mainly had given the group of at-risk students the same tasks as the rest of the class during literacy centers/station teaching (i.e repetition tasks emphasizing letter knowledge, phonics and writing), but with enhanced support (Sunde, 2015).

2.6. Adequacy of randomization

The three intervention groups were tested for equivalence on six measures taken immediately post-randomization: letter-knowledge, first phoneme isolation, phoneme blending, word reading, rapid automatized naming, and vocabulary. A multivariate ANOVA provided evidence that across these measures all groups were equivalent prior to the start of intervention, $F(2, 128) = 1.29$, $p = .23$, $\eta^2 = 0.06$. All MANOVA assumptions and diagnostics indicated that this test was valid. Post-hoc univariate follow-up tests also indicated no differences among groups on any pre-intervention outcome (see the pre-intervention scores, including the scores for the not-at-risk comparison sample, in Appendix A). The overall proportions of at-risk males/females were .595/.405, and these proportions did not differ across intervention group, $\chi^2(2) = 0.29$, $p = .86$. Because the pretest measures were not repeated at the end of Grade 1, the data were analyzed as a post-test only control group design. Since randomization was successful and all groups were equivalent prior to intervention, no covariates were included in the model. However, a sensitivity analysis (reported below) was also conducted, to ensure that the same pattern of results held with pretest covariates included.

Evidence for effective randomization was also observed at the school level as measured by a value-added indicator for Norwegian schools (i.e. schools mean score on National Tests in Reading in Grade 5 controlling for family background (Steffensen, Ekren, Zachrisen, & Kirkebøen, 2017). A one-way ANOVA showed no evidence of differences among schools assigned to the three intervention conditions, or to schools with not-at-risk children, $F(3,20) = 0.42$, $p = .74$). Since the number of schools per condition was small, bias-corrected and accelerated confidence intervals were also examined for each pairwise condition difference, and these also showed no school-to-school differences on the value-added scores.

3. Results

The overall analysis plan was as follows: first, formulating and reporting models for the primary outcome measures; second, conducting sensitivity and supplementary analyses; third, analyzing long-term follow up outcomes.

3.1. Description of outcome models used to evaluate the intervention

As described previously, four primary outcome measures were taken immediately following intervention at the end of Grade 1: word reading, sentence reading, spelling, and letter writing. Significant ceiling effects were observed for the intervention children on all outcomes, with proportion of the sample censored from above as follows: word reading, 21.4%, sentence reading, 35.1%, spelling, 9.4%, and letter writing, 56.5%. As a result, standard linear regression models for evaluating group differences at post-intervention could not be formulated. A form of censored regression, Tobit modeling, was employed to estimate post-intervention outcome models. This technique utilizes the information held in the censored (i.e., reached ceiling or did not reach ceiling) and non-censored (i.e., varying scores on the outcome) cases to estimate a latent value for each outcome measure, which is then contrasted across intervention groups (Tobin, 1958; Wooldridge, 2016). All Tobit models were analyzed with the *tobit* procedure within Stata/MP version 14.2 (StataCorp, 2015).

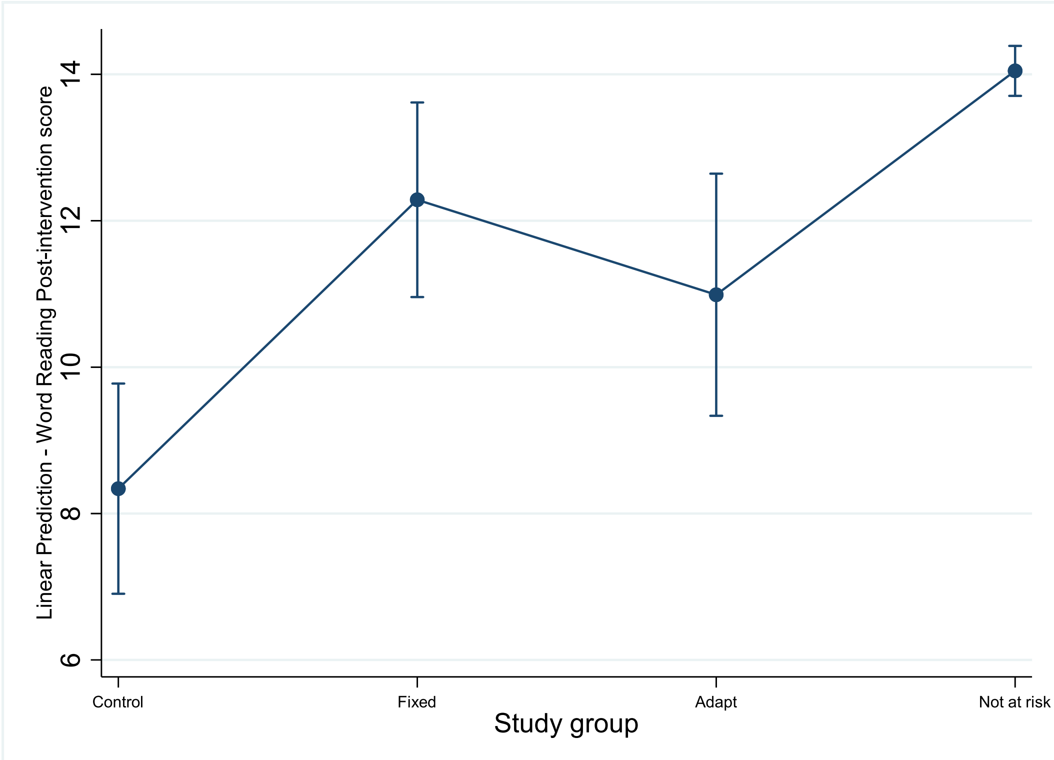
Because schools, not individual participants, were randomly assigned to intervention condition, membership in the 12 schools was included as fixed effects via 11 dummy-coded vectors (0, not a member a particular school; 1, a member of a particular school). A fixed effect for nesting within schools was chosen over a random effect (or a multilevel model) factor because of recent simulation work that suggests this approach leads to greatest power and lowest type-I errors in the context of a small number of groups and/or members per group in nested designs (Bell & Jones, 2015; McNeish & Stapleton, 2016). This strategy allowed for the control of all school-based heterogeneity, but still results in unbiased standard errors. The primary disadvantage of this strategy is that the models cannot be used to make inferences about school-level dynamics, but this is not a goal of the present study. Our implementation of this model followed equation 7 in McNeish and Stapleton (2016).

An a priori, hypothesis-driven approach to group comparisons was taken. Three orthogonal contrasts, representing the three degrees-of-freedom available with the four groups, were formed to represent the study hypotheses. First, post-test scores of the not-at-risk children were compared to those of the at-risk children. Second, scores of the two intervention groups were compared to children in the active control. Finally, the two intervention groups were compared. The proportion of observations censored from above was too high to employ Tobit regression for the letter writing outcome (56.5%). For the remaining three outcomes, Table 4 summarizes the results of each a priori hypothesis. Across these three outcomes, not-at-risk children ended Grade

Table 4
Apriori contrasts and post-hoc followup group comparisons across three post-intervention outcome measures.

Outcome	Contrast	Estimate	SE_{bs}	p_{norm}	95% CI	
					Lower	Upper
Word reading	Not-at-risk vs. At-risk	3.51	0.43	< .001	2.55	4.29
	Intervention vs. Control	3.29	1.08	.002	1.22	5.66
	Adapt vs. Fixed	−1.30	0.17	.227	−3.93	0.24
Sentence reading	Not-at-risk vs. At-risk	3.56	0.45	< .001	2.67	4.44
	Intervention vs. Control	2.64	0.98	.007	0.71	4.56
	Adapt vs. Fixed	−1.69	1.09	.121	−3.81	0.44
Spelling	Not-at-risk vs. At-risk	4.00	0.49	< .001	2.96	4.87
	Intervention vs. Control	3.23	1.09	< .001	1.20	5.86
	Adapt vs. Fixed	0.47	1.10	0.67	−1.69	2.63

Notes: SE_{bs} = boostrap standard error for the contrast estimate; p_{norm} = normal theory p-value testing that the estimate differs from 0; 95% CI = 95% bias corrected and accelerated confidence interval derived from 1000 bootstrap resamples.



a.

Fig. 2. **a**-c. Tobit model point estimates of post-intervention mean performance on word reading by study group, with 95% bias corrected and accelerated confidence intervals derived from 1000 replications. **b**. Tobit model point estimates of post-intervention mean performance on sentence reading by study group, with 95% bias corrected and accelerated confidence intervals derived from 1000 replications. **c**. Tobit model point estimates of post-intervention mean performance on spelling by study group, with 95% bias corrected and accelerated confidence intervals derived from 1000 replications.

1 substantially higher than the at-risk children. For all three outcomes, post-intervention scores at post-test were higher than control (all p-values $< .001$). No differences were observed between the I_Adapt and I_Fixed. Fig. 2 a–c illustrate these dynamics, utilizing model-estimated mean performance for each outcome, along with 95% bias-corrected and accelerated confidence intervals.

3.2. Effect size for the difference between groups

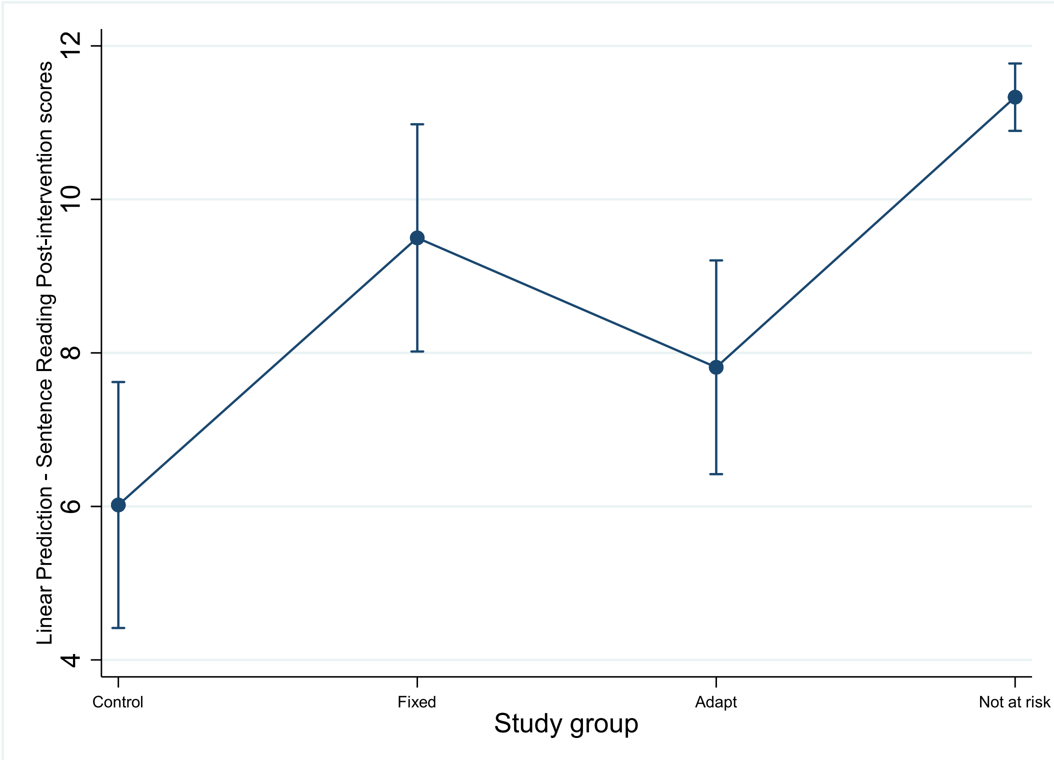
Effect sizes were calculated using the linear prediction estimates of mean outcome scores derived from the Tobit models, along with the pooled pretest standard deviations. Both interventions were associated with moderate to substantial effects relative to control. As displayed in Table 5 the Adapt intervention was associated with effects sizes that ranged from 0.38 to 0.64; the Fixed intervention was associated with effect sizes ranging from 0.58 to 0.75. Because there were fewer participants in the Adapt intervention, the confidence intervals were wider and the effect size estimates more variable. In one case, for I_Adapt, when the outcome was sentence reading, the confidence interval crossed zero, suggesting that this effect might not be reliably different from zero.

For the spelling measure 6 items, designed to reduce ceiling effects and tap into higher spelling ability, were added to the 14 from the National Spelling test (reported above). When considered together, no ceiling effects were observed, and the effect sizes for both Adapt (Hedge's $g = 0.81$) and Fixed (Hedge's $g = 0.68$) were notably larger. The larger effect sizes may be due to the intervention having an effect on spelling only detectable at higher skill levels, with the addition of the more difficult items.

3.3. Sensitivity analyses to support the findings

One implication of excluding the non-complying school may have been to introduce an upward bias to the effect sizes for intervention relative to control. This school was excluded due to being non-compliant with the intervention protocol. A sensitivity analysis was thus conducted with this school included, to estimate conservative standardized effect sizes for both interventions together relative to the control condition. Overall effect sizes were very similar to the main results without this school: word reading, $d = 0.73$ (95% CI 0.37 to 1.10); sentence reading, $d = 0.56$ (95% CI 0.20 to 0.93); spelling, $d = 0.60$ (95% CI 0.23 to 0.96).

A second sensitivity analysis was conducted to evaluate the effects of covariates on the estimation of intervention effects. While we have strong evidence that randomization was effective, there is still the possibility that post-randomization imbalances in the three treatment groups may have influenced our conclusions. To the end, the three outcome models were re-evaluated with the following pre-intervention covariates included: familial risk for reading problems, phoneme isolation, phoneme blending, rapid automatized naming, letter knowledge and vocabulary. The results of this analysis is reported in Appendix C, showing that the primary intervention versus control contrast was upheld. The at-risk versus not-at-risk contrast was found to be non-significant, as expected, since the pretest covariates were used to select out these groups. Overall, this sensitivity analysis demonstrated that our results were robust to the inclusion of six critical early literacy covariates. (Please see Appendix B for detailed correlations between and within pretest measures and post-intervention outcomes).



b.

Fig. 2. (*continued*)

3.4. Post-intervention follow up

All children were also measured on reading, spelling, and sentence reading outcomes at two follow up points, being the end of Grade 2 and 3. Since reading skill changes rapidly in this period, and the instruments needed to be adapted according to the age of the children, item response theory (IRT) models were implemented to create scores that allow measuring change over time for each child. For this scaling, tests were designed to have overlapping items between the measuring points in order to serve as anchor items. Students who took part in different measurement occasions was added as pseudo-students for the scaling. For the IRT scaling model, the one parameter logistic model (1PL – also known as "Rasch-Model") was used to estimate item difficulties and student abilities simultaneously for word reading, sentence reading and spelling separately. As selection criteria for the items, a weighted Fit-statistic between 0.7 = < MNSQ = < 1.3 was accepted. The overall EAP-reliability was as follows: word reading, 0,970, sentence reading, 0.919, and spelling, 0.916. Each of the scales have been linear transformed to a distribution with a mean of 200 and standard deviation of 20. As person-ability estimate, the WLE (warms weighted maximum likelihood estimate) was used, which is the best person-point estimate.

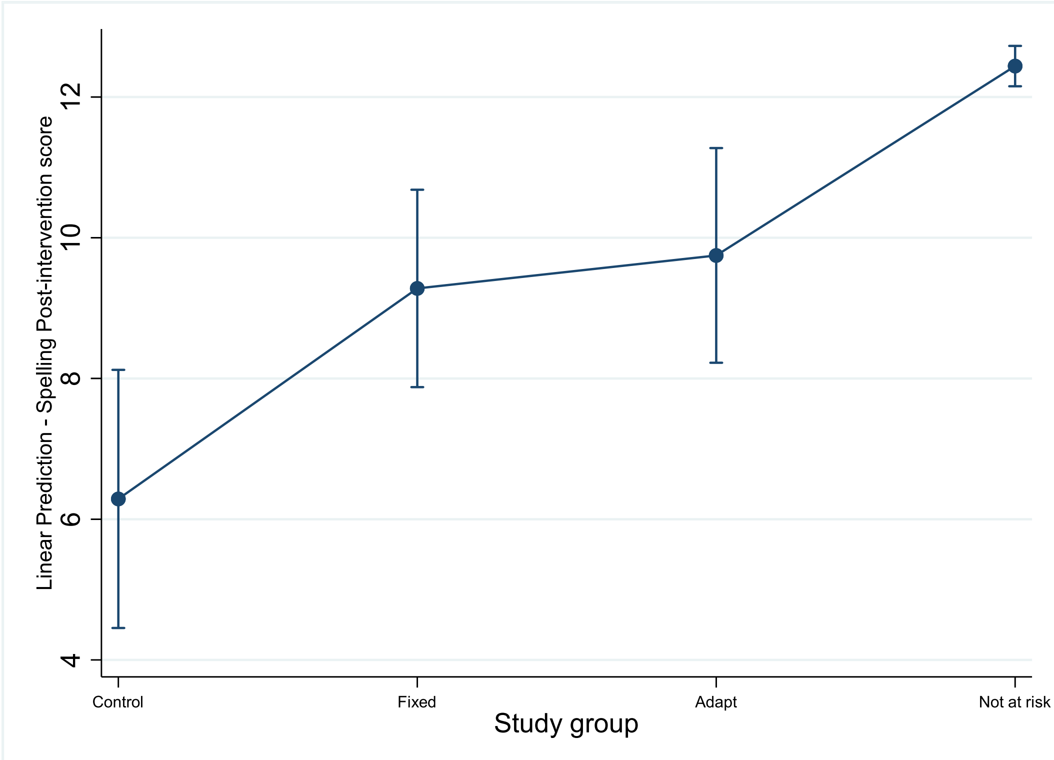
The three waves of IRT-derived WLE scores were entered into a multilevel growth model with controls for school, and with three a priori group contrasts as predictors: 1) not-at-risk children were contrasted with those who had been assigned to I_Adapt or I_Fixed; 2) the two active treatments were contrasted with each other; 3) control children were contrasted with all others. The best model of reading growth over the follow up period converged and included both fixed and random slopes and intercepts, with time centered at the middle follow-up occasion, the end of Grade 2. In word reading, significant child-to-child variability in Grade 2 level ($\sigma_0^2 = 156.05$, $SE = 10.21$, $p < .001$) and rate of growth ($\sigma_2^2 = 37.22$, $SE = 13.42$, $p < .001$) was observed over the two years. Fig. 3 illustrates the model-estimated fixed effects per group for the reading outcome. At the end of Grade 2, not-at-risk children were 10.43 (SE = 1.67, $p < .001$) points higher than those who had been in one of the active intervention conditions. Children in one or the other of the active interventions did not differ from each other. Children in the control condition were 9.18 (SE = 2.54, $p < .001$) points lower than all other children combined. The overall growth rate over the three waves was 18.21 points per wave (SE = 0.74, $p < .001$), and no groups differed from any other groups on this growth rate, for the reading outcome.

Similar models were formed for the spelling and sentence reading outcomes. On these outcomes, the pattern of results was like that of the reading outcome. One notable difference emerged. The rate of growth for the control group *exceeded* the rate of growth for the other children (not-at-risk or intervention children), on both spelling (control slope advantage = 8.21, $SE = 1.59$, $p < .001$) and sentence reading (control slope advantage = 6.30, $SE = 1.66$, $p < .001$).

4. Discussion

4.1. Efficacy of the On Track intervention

In this study we found that an early multi-component reading intervention, involving a combination of teacher-led instruction and CAI, enhanced word reading, sentence reading and spelling in a sample of Norwegian first-graders with low pre-reading skills and at risk for RD. The results confirm that intensive intervention, mainly informed by research from English-speaking countries, can be effective for first-graders learning to read a semi-transparent orthography. The intervention included training of graphene-phoneme correspondences and phonics, a common approach in orthographies with regular phoneme-graphene correspondences (Goswami, Gombert, & de Barrera, 1998; Share, 1995). Further, the intervention took an approach to reading instruction, operationalized according to general principles of hermeneutics, in which the fundamental idea is that the whole and the part must always be understood in relation to each other (Gadamer, 1960).



c.

Fig. 2. (*continued*)

As such, students who received the On Track intervention engaged in meaningful situations around text from the very beginning of formal instruction. The students would read, write, listen, discuss, verbalize, suggest, practice and reflect; shifting back and forth between a focus on parts and wholes in each session. Tønnessen and Uppstad (2015) argue that the focus in the hermeneutic approach is on fine-tuning the equilibrium of letters and words versus text.

Children who learn to read transparent orthographies show faster progress than children who learn to read opaque orthographies (Seymour et al., 2003). Further, English-speaking children with RD have more severe problems as reading accuracy *continues* to be a core problem, whereas reading fluency is the most persitant problem across languages (Ziegler et al., 2003). Landerl et al. (2012) have suggested that a simple and transparent representation of the phonological structure may help children to overcome early deficits, even more so if formal reading instruction is strongly phonics-based. According to this reasoning, phonics-based classroom instruction could be considered sufficient for learning how to read accurately for children who display early deficits in reading in transparent orthographies. However, the efficacy of the intervention in the present study is comparable to previously reported early multi-component reading intervention in English-speaking countries (see e.g. Al Otaiba et al., 2014; Blachman et al., 2004; Mathes et al., 2005). The results are also is in line with a previous intervention study where children learnt to read the transparent Finnish orthography (Saine et al., 2011), and suggests that also at-risk children learning to read in more transparent orthographies can profit from intensive intervention in early phases of reading development - even if the classroom instruction for most children is already phonics-based.

Table 5
Intervention effect sizes relative to control.

	Effect	Hedge's g	95% CI	
			Lower	Upper
Word reading	Adapt vs. Control	0.50	0.05	0.95
	Fixed vs. Control	0.75	0.34	1.17
	Intervention vs. Control	0.75	0.37	1.12
Sentence reading	Adapt vs. Control	0.38	−0.07	0.82
	Fixed vs. Control	0.66	0.25	1.08
	Intervention vs. Control	0.57	0.20	0.94
Spelling	Adapt vs. Control	0.64	0.18	1.09
	Fixed vs. Control	0.58	0.17	0.99
	Intervention vs. Control	0.61	0.24	0.98

Notes: SE_{bs} = boostrap standard error for the contrast estimate; 95% CI = 95% bias corrected and accelerated confidence interval derived from 1000 bootstrap resamples.

4.2. *Computer-assisted reading instruction: adaptivity versus difficulty levels*

The only difference between Adapt and Fixed intervention was whether students used the adaptive GraphoGame, or the non-adaptive computer app, On Track ABC, in the ABC lesson. There were no significant differences on any outcome measures between children who received either of the two interventions. As this was a multi-component reading intervention it is not possible to determine whether all components were necessary for success, this also holds for the ABC-lesson. However, in the following, we will assume that the ABC-lesson contributed at least somewhat to students' documented progress and suggest three different interpretations for the lack of significant differences between the Adapt and Fixed conditions: 1) the additional content included in the On Track ABC app compensates for any gain triggered by the adaptive features in GraphoGame, 2) there is no benefit in including adaptive features in gaming software developed to support letter-sound knowledge and word reading for this age group, or 3) the adaptive features in GraphoGame do not function at an optimal level due to students' opting out of the learning process by guessing answers at random.

The first interpretation suggests that the additional orthographic

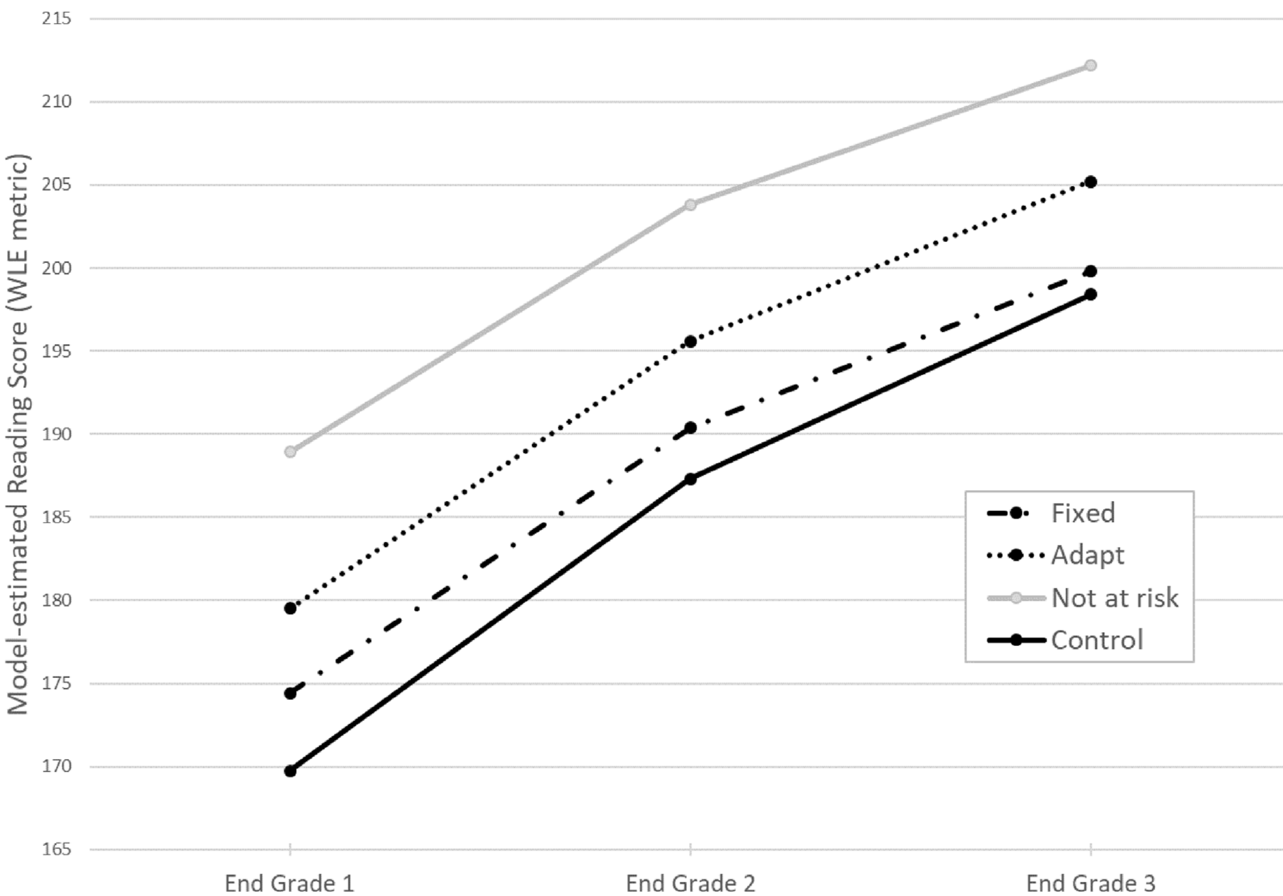


Fig. 3. Multilevel growth model-estimated reading outcome scores over the long term follow up testing points.

and phonological content included in the On Track ABC app (i.e., matching lower case and upper-case letters, identifying the initial sound in words and spelling) compensated for any gain triggered by the adaptive features in GraphoGame. However, as the additional content also was addressed in other components of the intervention, this might not be a plausible explanation. The second interpretation would suggest that there is no additional gain achieved via the adaptive features (over simply increasing difficulty levels) for training letter-sound correspondences and word reading in at-risk students in Grade 1. Consistent with this explanation, Ronimus, Kujala, Tolvanen, and Lyytinen (2014) found neither a reward- nor a level-of-challenge-effect in a study where students played two different versions of GraphoGame (a fixed versus an adaptive). However, we cannot rule out the possibility that the school-based setting with a restricted time slot for playing, might have restricted students engagement or the possibility to act based on their engagement (e.g. through several short sessions a day). The final possibility is that the adaptive function was pre-empted by non-engagement in the CAI tasks. For the adaptive features to function at an optimal level, students' performance in vivo should reflect their actual skill level. If students guess or select responses at random (even if they *can* identify the correct response) the algorithm will take their random performance as baseline and default to easier tasks, compromising the optimal level of challenge. As a game, GraphoGame is developed to maintain engagement in the learning tasks. However, both Ronimus et al. (2014) and Ecochard (2015) have reported that even if the reward system initially seemed to trigger children's interest, the effect vanished after a few sessions. Thus, there is a possibility that students this young, and who also struggle to learn, don't have sufficient internal motivation to stay devoted to these tasks, and that the gaming features of adaptation and feedback, intended to keep them in the learning task, are insufficient.

4.3. Long term effect of the On Track intervention

The controlled-trial portion of the present study resulted in skill gains for the at-risk children that exceeded at-risk children in the enhanced control condition. This pattern of results is consistent with studies of children with RD who remain untreated (e.g., Cirino et al., 2002), and with the consistent positive effect of early intervention that has previously been demonstrated (e.g., Lovett et al., 2017). Within the two-year follow up period, the at-risk children who had received the intervention had growth curves that were parallel to the not-at-risk children. This is also consistent with past research that has shown continued skill gain post-intervention. With parallel growth curves, the present study provided evidence that the intervention effect was sustained for two-years post-intervention. Comparable to the effect sizes reported immediately post intervention, the persistence of the effect are similar to previous intervention studies in English orthography (see e.g. Blachman et al., 2004, 2014).

The controlled-trial portion of the intervention design ended with the end of Grade 1. After this point, schools were free to implement a full range of supports for all at-risk children. Casual conversations with teachers and administrators in the control schools indicated unease with knowing who the at-risk children were, but without tools to intervene fully. In several cases, control schools informally indicated their determination to 'catch-up' the children we had identified as at-risk. The follow up analyses indicated that this process may indeed have been at play, since two out of three outcomes (e.g., spelling and sentence reading) showed significantly faster growth rates for control children over the two-year follow up. On word reading, control growth was parallel to that of the not-at-risk and/or intervention groups. Since control children started the follow up period lower than the at-risk intervention children, such 'compensatory rivalry' cannot conclusively explain the observed effects, which may also be due to regression to the mean.

4.4. Limitations

Although we believe that our study is an important contribution to early reading intervention, we recognize several limitations. At-risk identification was based on pre-reading skills that previous research has found to be highly predictive of later reading achievement. However, since children in the present study were identified before they received *any* formal instruction, it is challenging to differentiate children with poor pre-reading skills and at risk of dyslexia, from children whose low pre-reading skills at school entry mainly reflected limited literacy-related experiences. As such, our sample of at-risk students are likely to have included false positives. For example, the at-risk sample may have included children who would learn to read once exposed to formal instruction. As well, since all students received a multi-component intervention consisting of phonics instruction, guided reading, free spelling and shared reading, it is not possible to determine whether all components of intervention were necessary for success.

There were also measurement limitations. Two of the measures used pre-intervention had low reliability for the at-risk students; letter-sound knowledge and word reading. Low letter-sound skills of the at-risk children resulted in low reliability coefficients for the measure of letter knowledge, since there, as expected, were individual differences in which letters they knew. This would primarily impact identification of at-risk children for random assignment to the intervention, possibly allowing inclusion of children who were not at risk based on letter knowledge alone. The at-risk determination was multi-dimensional, minimizing this concern. In addition, no significant post-randomization differences amongst groups were observed on this measure. Concerning words reading, this variable was only used for investigating equivalence between conditions before intervention, and the results confirmed that at-risk children had not begun to acquire reading skills at school entry. Due to the distribution, word reading was not included among the covariates in the sensitivity analysis.

Intervention outcomes were evaluated by measures that were not equivalent pre and post intervention. This is in part inevitable due to the large variation in skills at school entry, and the fast progress students make when they start to receive formal instruction. With outcomes commensurate at pre and post, we could have evaluated intervention effects using repeated measures models. Given the evidence for adequacy of randomization, the post-test only outcome models still have strong validity. Finally, the ceiling effects in the outcome measures might have underestimated the full effect of the intervention. In one case where more difficult items were added to the outcome (i.e., spelling), intervention effect sizes were larger, suggesting that the intervention had additional effects that were not being captured at the upper end of the skill distribution. Finally, we choose to remove a non-complying school from the final analysis, which might have disrupted the randomization. However, we performed a sensitivity analysis that show significant effect of intervention even if the non-complying school is included.

5. Conclusions

The results of the present study indicate that at-risk children learning to read in a semi-transparent orthography (e.g. Norwegian) can profit from similar intervention content as children learning to read in more opaque orthographies. The magnitude of the effect size indicate that these findings should have consequences for practice in schools (Cooper, 2008).

Funding

This study was supported by a grant from the Norwegian Research Council (grant number 237861).

Acknowledgements

The authors are grateful to Professor Ulla Richardson, University of Jyväskylä, for her support in making the Norwegian version of GraphoGame and Professor Knut Schwippert, University of Hamburg, for implementing the IRT models that made it possible to measure change over time.

We also thank research assistants, teachers, parents, and students involved in the project.

Appendix A. Pre-intervention scores across study groups

Table A1
Pre-intervention scores for the three randomly-assigned intervention groups, plus the not-at-risk comparison group.

Measure	Group	M	SD	95% CI	
				Lower	Upper
Letter-knowledge	Control	8.16	2.74	7.40	8.97
	Fixed	7.92	2.64	7.15	8.64
	Adapt	8.26	2.86	7.19	9.22
	Not-at-risk	12.72	2.84	12.46	12.98
First phoneme isolation	Control	1.43	1.69	1.00	1.89
	Fixed	1.54	1.32	1.19	1.92
	Adapt	1.03	1.34	0.63	1.48
	Not-at-risk	5.87	2.61	5.65	6.08
Word reading	Control	0.77	0.77	0.54	1.00
	Fixed	0.52	0.61	0.37	0.68
	Adapt	0.66	0.59	0.46	0.86
	Not-at-risk	3.54	2.89	3.32	3.77

(*continued on next page*)

Table A1 (*continued*)

Measure	Group	M	SD	95% CI	
				Lower	Upper
Rapid automatized naming	Control	40.04	12.91	36.80	43.46
	Fixed	37.99	10.18	35.49	40.89
	Adapt	41.14	11.34	37.39	45.03
	Not-at-risk	30.65	7.66	29.99	31.35
Vocabulary	Control	9.05	4.06	7.90	10.09
	Fixed	9.62	3.74	8.59	10.64
	Adapt	9.83	3.53	8.56	11.00
	Not-at-risk	13.01	3.72	12.69	13.33
Phoneme blending	Control	0.82	1.24	0.51	1.18
	Fixed	0.96	1.08	0.67	1.28
	Adapt	1.29	1.30	0.91	1.71
	Not-at-risk	3.89	2.48	3.70	4.09

Notes: Control n = 44, Fixed n = 52, Adapt n = 35, Not-at-risk n = 550; 95% CI represent Bias-corrected and accelerated bootstrap confidence intervals derived from 1000 resamples.

Appendix B. Correlations among pretest variables and post-intervention outcomes

Table B1
Correlations within Pretest measures, over the full sample and by group.

	Pretest	1.	2.	3.	4.
Full sample	1. Letter knowledge				
	2. Phoneme isolation	.48**			
	3. Phoneme blending	.33**	.55**		
	4. RAN	-.21**	-.24**	-.23**	
	5. Vocabulary	.30**	.41**	.34**	-.27**
Not at-risk	1. Letter knowledge				
	2. Phoneme isolation	.26**			
	3. Phoneme blending	.16**	.43**		
	4. RAN	-.03	-.05	-.11**	
	5. Vocabulary	.15**	.31**	.27**	-.17**
Control	1. Letter knowledge				
	2. Phoneme isolation	.11			
	3. Phoneme blending	-.11	-.19		
	4. RAN	.15	-.04	.12	
	5. Vocabulary	.27	.11	-.22	-.05
Fixed	1. Letter knowledge				
	2. Phoneme isolation	.01			
	3. Phoneme blending	-.17	.06		
	4. RAN	.05	.27	.13	
	5. Vocabulary	.08	.14	-.28*	-.15
Adapt	1. Letter knowledge				
	2. Phoneme isolation	-.01			
	3. Phoneme blending	-.35*	-.12		
	4. RAN	-.05	.27	.26	
	5. Vocabulary	-.08	-.18	.22	-.26

Notes: RAN = Rapid automatized naming; $^{*}p < .05$; $^{**}p < .01$.

Table B2
Correlations within Posttest measures, over the full sample and by group.

	Pretest	1.	2.
Full sample	1. Spelling		
	2. Word reading	.69**	
	3. Sentence reading	.70**	.73**
Not at-risk	1. Spelling		

(*continued on next page*)

Table B2 (*continued*)

	Pretest	1.	2.
	2. Word reading	.59**	
	3. Sentence reading	.57**	.60**
Control	1. Spelling		
	2. Word reading	.79**	
	3. Sentence reading	.85**	.88**
Fixed	1. Spelling		
	2. Word reading	.64**	
	3. Sentence reading	.75**	.74**
Adapt	1. Spelling		
	2. Word reading	.65**	
	3. Sentence reading	.77**	.69**

Notes: *$p < .05$; **$p < .01$.

Table B3
Correlations between Pretest measures and Posttest measures, over the full sample and by group.

Pretests		Posttest Outcomes		
		Word reading	Sentence reading	Spelling
Full sample	Letter knowledge	.36**	.35**	.42**
	Phoneme isolation	.36**	.33**	.47**
	Phoneme blending	.32**	.24**	.39**
	RAN	-.32**	-.28**	-.32**
	Vocabulary	.32**	.23**	.35**
Not at-risk	Letter knowledge	.20**	.20**	.28**
	Phoneme isolation	.25**	.20**	.37**
	Phoneme blending	.25**	.13**	.31**
	RAN	-.22**	-.14**	-.18**
	Vocabulary	.21**	.11*	.26**
Control	Letter knowledge	.37*	.34*	.26
	Phoneme isolation	.25	.18	.30*
	Phoneme blending	.07	.16	.09
	RAN	-.16	-.33*	-.35*
	Vocabulary	.22	.15	.29
Fixed	Letter knowledge	.22	.18	.03
	Phoneme isolation	-.04	.19	.06
	Phoneme blending	-.16	-.04	-.09
	RAN	-.20	-.13	-.01
	Vocabulary	.33*	.23	.22
Adapt	Letter knowledge	.20	.38*	.32
	Phoneme isolation	-.09	.11	.26
	Phoneme blending	.05	-.03	-.06
	RAN	-.23	-.20	-.31
	Vocabulary	.17	.11	.06

Notes: RAN = Rapid automatized naming; *$p < .05$; **$p < .01$.

Appendix C. Results of sensitivity analysis

Table C1
Apriori contrasts and post-hoc followup group comparisons across three post-intervention outcome measures, with pretest covariates: presence of familial risk for reading problems, letter knowledge, phoneme isolation, phoneme blending, vocabulary, and rapid automatized naming (analogue to Table 4 in the main manuscript).

Outcome	Contrast	Estimate	SE_{bs}	p_{norm}	95% CI	
					Lower	Upper
Word reading	Not-at-risk vs. At-risk	−0.04	0.53	.937	−1.08	0.99
	Intervention vs. Control	2.85	0.84	.001	1.19	4.50
	Adapt vs. Fixed	−1.09	0.93	.238	−2.92	0.72

(*continued on next page*)

Table C1 (*continued*)

Outcome	Contrast	Estimate	SE_{bs}	p_{norm}	95% CI	
					Lower	Upper
Sentence reading	Not-at-risk vs. At-risk	0.05	0.53	.991	−1.03	1.05
	Intervention vs. Control	2.17	0.79	.006	0.62	3.71
	Adapt vs. Fixed	−1.38	1.12	.218	−3.57	0.82
Spelling	Not-at-risk vs. At-risk	0.61	0.55	.269	−0.47	1.69
	Intervention vs. Control	2.96	0.85	< .001	1.29	4.63
	Adapt vs. Fixed	0.73	1.05	0.70	−1.32	2.79

Notes: SE_{bs} = boostrap standard error for the contrast estimate; p_{norm} = normal theory p-value testing that the estimate differs from 0; 95% CI = 95% bias corrected and accelerated confidence interval derived from 1000 bootstrap resamples.

References

Al Otaiba, S., Connor, C. M., Folsom, J. S., Wanzek, J., Greulich, L., Schatschneider, C., et al. (2014). To wait in tier 1 or intervene immediately: A randomized experiment examining first grade response to intervention (RTI) in reading. *Exceptional Children, 81*(1), 11–27. http://dx.doi.org/10.1177/0014402914532234.

Bell, A., & Jones, K. (2015). Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods, 3*, 133–153. http://dx.doi.org/10.1017/psrm.2014.7.

Blachman, B. A., Schatschneider, C., Fletcher, J. M., Francis, D. J., Clonan, S. M., Shaywitz, B. A., et al. (2004). Effects of intensive reading remediation for second and third graders and a 1-year follow-up. *Journal of Educational Psychology, 96*(3), 444–461,. https://doi.org/10-1037/0022-0663.96.3.444.

Blachman, B. A., Schatschneider, C., Fletcher, J. M., Murray, M. S., Munger, K. A., & Vaughn, M. G. (2014). Intensive reading remediation in grade 2 or 3: Are there effects a decade later? *Journal of Educational Psychology, 106*(1), 46–57. http://dx.doi.org/10.1037/a0033663.

Cheung, A. C. K., & Slavin, R. E. (2012). How features of educational technology applications affect student reading outcomes: A meta-analysis. *Educational Research Review, 7*(3), 198–215. http://dx.doi.org/10.1016/j.edurev.2012.05.002.

Cheung, A. C. K., & Slavin, R. o. E. (2013). Effects of educational technology applications on reading outcomes for struggling readers: A best-evidence synthesis. *Reading Research Quarterly, 48*(3), 277–299.

Cirino, P. T., Rashid, F. L., Sevcik, R. A., Lovett, M. W., Frijters, J. C., Wolf, M., et al. (2002). Properties of measures of beginning decoding and related abilities in poor readers. *Journal of Learning Disabilities, 35*(6), 526–539.

Cooper, H. (2008). The search for meaningful ways to express the effect of interventions. *Child Development Perspectives, 2*(3), 181–186. http://dx.doi.org/10.1111/j.1750-8606.2008.00063.x.

Corbett, A. (2001). Cognitive computer Tutors: Solving the two-sigma problem. In M. Bauer, P. J. Gmytrasiewicz, & J. Vassileva (Vol. Eds.), *User modeling 2001. UM 2001. Lecture notes in computer science: Vol. 2109*Berlin, Heidelberg: Springer. http://dx.doi.org/10.1007/3-540-44566-8_14.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-adapted Interaction, 4*(4), 253–278. http://dx.doi.org/10.1007/BF01099821.

Ecochard, S. (2015). *Learning to read with GraphoGame, an ethnography in a Peruvian rural school. (Pro-Gradu)*Unpublished doctoral dissertation. University of Jyväskylä.

Ehri, L., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research, 71*(3), 393–447. http://dx.doi.org/10.3102/00346543071003393.

Fälth, L., Gustafson, S., Tjus, T., Heimann, M., & Svensson, I. (2013). Computer-assisted interventions targeting reading skills of children with reading disabilities – a longitudinal study. *Dyslexia, 19*(1), 37–53. http://dx.doi.org/10.1002/dys.1450.

Gadamer, H.-G. (1960). *Wahrheit und Methode, Grundzüge einer philosophischen Hereneutik.* Tübingen: Suhrkamp.

Galuschka, K., Ise, E., Krick, K., & Schulte-Körne, G. (2014). Effectiveness of treatment approaches for children and adolescents with reading disabilities: A meta-analysis of randomized controlled trials. *PLoS One, 9*(8), e89900. http://dx.doi.org/10.1371/journal.pone.0089900.

Goswami, U., Gombert, J. E., & de Barrera, L. F. (1998). Children's orthographic representations and linguistic transparency: Nonsense word reading in English, French, and Spanish. *Applied Psycholinguistics, 19*(01), 19–52. http://dx.doi.org/10.1017/S0142716400010560.

Kamykowska, J., Haman, E., Latvala, J.-M., Richardson, U., & Lyytinen, H. (2014). Developmental changes of early reading skills in six-year-old polish children and GraphoGame as a computer-based intervention to support them. *L1: Educational Studies in Language and Literature, 13*, 1–17. http://dx.doi.org/10.17239/L1ESLL-2013.01.05.

Katz, L., & Frost, R. (1992). The reading process is different for different orthographies: The orthographic depth hypothesis. *Advances in Psychology, 94*, 67–84. http://dx.doi.org/10.1016/S0166-4115(08)62789-2.

Kyle, F., Kujala, J., Richardson, U., Lyytinen, H., & Goswami, U. (2013). Assessing the effectiveness of two theoretically motivated computer-assisted reading interventions in the United Kingdom: GG rime and GG phoneme. *Reading Research Quarterly, 48*(1), 61–76. http://dx.doi.org/10.1002/rrq.038.

Landerl, K., Ramus, F., Moll, K., Lyytinen, H., Leppänen, P. H. T., Lohvansuu, K., et al. (2012). Predictors of developmental dyslexia in European orthographies with varying complexity. *Journal of Child Psychology and Psychiatry, 54*(6), 686–694. http://dx.doi.org/10.1111/jcpp.12029.

Landerl, K., Wimmer, H., & Frith, U. (1997). The impact of orthographic consistency on dyslexia: A German-English comparison. *Cognition, 63*(3), 315–334. http://dx.doi.org/10.1016/S0010-0277(97)00005-X.

Lovett, M. W., Barron, R. W., & Frijters, J. C. (2013). Word identification difficulties in children and adolescents with reading disabilities: Intervention research findings. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.). *Handbook of learning disabilities* (pp. 329–360). New York: Guilford Publications.

Lovett, M. W., Frijters, J. C., Wolf, M., Steinbach, K. A., Sevcik, R. A., & Morris, R. D. (2017). Early intervention for children at risk for reading disabilities: The impact of grade at intervention and individual differences on intervention outcomes. *Journal of Educational Psychology, 109*(7), 889–914. http://dx.doi.org/10.1037/edu0000181.

Lundetræ, K., Solheim, O. J., Schwippert, K., & Uppstad, P. H. (2017). Protocol: 'On Track', a group-randomized controlled trial of an early reading intervention. *International Journal of Educational Research, 86*, 87–95. http://dx.doi.org/10.1016/j.ijer.2017.08.011.

Mathes, P. G., Denton, C. A., Fletcher, J. M., Anthony, J. L., Francis, D. J., & Schatschneider, C. (2005). The effects of theoretically different instruction and student characteristics of struggling readers. *Reading Research Quarterly, 40*(2), 148–182. http://dx.doi.org/10.1598/RRQ.40.2.2.

McArthur, G., Eve, P. M., Jones, K., Banales, E., Kohnen, S., Anandakumar, T., et al. (2012). Phonics training for English-speaking poor readers. *The Cochrane Library*. http://dx.doi.org/10.1002/14651858.CD009115.

McNeish, D., & Stapleton, L. M. (2016). Modeling clustered data with very few clusters. *Multivariate Behavioral Research, 51*(4), 495–518. http://dx.doi.org/10.1080/00273171.2016.1167008.

Mishra, J., & Gazzaley, A. (2014). Harnessing the neuroplastic potential of the human brain & the future of cognitive rehabilitation. *Frontiers in Human Neuroscience, 8*, 218. http://dx.doi.org/10.3389/fnhum.2014.00218.

Morris, R. D., Lovett, M. W., Wolf, M., Sevcik, R. A., Steinbach, K. A., Frijters, J. C., et al. (2012). Multiple-component remediation for developmental reading disabilities: IQ, SES, and RACE as factors on remedial outcome. *Journal of Learning Disabilities, 45*(2), 99–127. http://dx.doi.org/10.1177/0022219409355472.

Norwegian Directorate for Education and Training (NDET) (2015a). *Utdanningsspeilet 2015: Tall og analyse av barnehager og grunnopplæringen i Norge [Education Mirror 2015: Facts and analysis of kindergartens and primary and secondary education in Norway].* Oslo: Utdanningsdirektoratet.

Norwegian Directorate for Education and Training (NDET) (2015b). *Tall fra Grunnskolens informasjonssystem (GSI) 2014/15 [Figures from the information system for primary and secondary school, 2014/15].* Retrieved from http://www.udir.no/globalassets/upload/statistikk/gsi/grunnskole-gsi-notat-2014-15.pdf.

Norwegian Directorate for Education and Training (NDET) (2015c). *Kartleggingsprøve i lesing 1. trinn. Veiledning til lærere 2015 [Assessment test in reading, Grade 1. Guidelines for teachers].* Oslo: Utdanningsdirektoratet.

Norwegian Ministry for Education and Research (NMER) (2011). *Framework plan for the content and tasks of kindergarten. Amended by regulation 10th of january 2011.* No. 51. Oslo, NO. Retrieved from https://www.udir.no/in-english/framwork-plan-for-the-content-and-tasks-of-kindergartens/.

Richardson, U., & Lyytinen, H. (2014). The GraphoGame method: The theoretical and methodological background of the technology-enhanced learning environment for learning to read. *Human Technology: An Interdisciplinary Journal on Humans in ICT Environments, 10*(1), 39–60. http://dx.doi.org/10.17011/ht/urn.201405281859</a.

Ronimus, M., Kujala, J., Tolvanen, A., & Lyytinen, H. (2014). Children's engagement during digital game-based learning of reading: The effects of time, reward and challange. *Computers & Education, 71*, 237–246. http://dx.doi.org/10.1016/j.compedu.2013.10.008.

Saine, N. L., Lerkkanen, M.-K., Ahonen, T., Tolvanen, A., & Lyytinen, H. (2011). Computer-assisted remedial reading intervention for school beginners at risk for reading disability. *Child Development, 82*(3), 1013–1028. http://dx.doi.org/10.1111/

j.1467-8624.2011.01580.x.

Scammacca, N., Vaughn, S., Roberts, G., Wanzek, J., & Torgesen, J. K. (2007). *Extensive reading interventions in grades K-3: From research to practice.* Porthsmouth, NH: RMC Research Corporation, Center on Instruction Retrieved from: http://buildingrti.utexas.org/sites/default/files/booklets/Extensive_Reading_Interventions_K-3.pdf.

Seymour, P. H., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94*(2), 143–174.

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition, 55*(2), 151–218. http://dx.doi.org/10.1348/000712603321661859.

Share, D. L. (2008). On the Anglocentricities of current reading research and practice: The perils of overreliance on an" outlier" orthography. *Psychological Bulletin, 134*(4), 584–615. http://dx.doi.org/10.1037/0033-2909.134.4.584.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*(4), 360–407.

Stanovich, K. E., & Siegel, L. S. (1994). Phenotypic performance profile of children with reading disabilities: A regression-based test of the phonological-core variable-difference model. *Journal of Educational Psychology, 86*(1), 24. http://dx.doi.org/10.1037/0022-0663.86.1.24.

StataCorp (2015). *Stata statistical software.* College Station, TX: StataCorp LP.

Steffensen, K., Ekren, R., Zachrisen, O. O., & Kirkebøen, L. J. (2017). *Er det forskjeller i skolers og kommuners bidrag til elevenes læring i grunnskolen? En kvantitative studie. [Are there differences in the degree to which schools and municipalities contribute to students learning? A quantitative study]*Report 2017/2. Oslo: Statistics Norway.

Størksen, I., Ellingsen, I. T., Tvedt, M. S., & Idsøe, E. M. (2013). Norsk vokabulartest (NVT) for barn i overgangen mellom barnehage og skole: Psykometrisk vurdering av en nettbrettbasert test [Norwegian Vocabulary Test (NVT) for children in transition between day care center and school. Reviewing psychometric properties of a computer tablet based test]. *Spesialpedagogikks forskningsdel, 4*(13), 40–54.

Suggate, S. P. (2010). Why what we teach depends on when: Grade and reading intervention modality moderate effect size. *Developmental Psychology, 46*(6), 1556. http://dx.doi.org/10.1037/a0020612.

Suggate, S. P. (2016). A meta-analysis of the long-term effects of phonemic awareness, phonics, fluency, and reading comprehension interventions. *Journal of Learning Disabilities, 49*(1), 77–96. http://dx.doi.org/10.1177/0022219414528540.

Sunde, K. (2015). *Tidlig identifisering - er det nok? Læreres erfaringer knyttet til å få identifisert elever som kan være i faresonen for å utvikle lesevansker. [Early identification - is it enough? How teachers experience early identification of students at risk for reading difficulties]*(Masters thesis). University of Stavanger Retrieved from: https://brage.bibsys.no/xmlui/bitstream/handle/11250/299372/Sunde_Kristin%20.pdf?sequence=1&isAllowed=y. .

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica, 26*(1), 24–36. http://dx.doi.org/10.2307/1907382.

Tønnessen, F. E., & Uppstad, P. H. (2015). *Can we read letters? Reflections on fundemental issues in reading and dyslexia research.* Rotterdam: Sense Publishers.

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Herron, J., & Lindamood, P. (2010). Computer-assisted instruction to prevent early reading difficulties in students at risk for dyslexia: Outcomes from two instructional approaches. *Annals of Dyslexia, 60*(1), 40–56. http://dx.doi.org/10.1007/s11881-009-0032-y.

Vaughn, S., Denton, C. A., & Fletcher, J. M. (2010). Why intensive interventions are necessary for students with severe reading difficulties. *Psychology in the Schools, 47*(5), 432–444. http://dx.doi.org/10.1002/pits.20481.

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach* (6th ed.). Boston: Cengage.

Ziegler, J. C., Perry, C., Ma-Wyatt, A., Ladner, D., & Schulte-Körne, G. (2003). Developmental dyslexia in different languages: Language-specific or universal? *Journal of Experimental Child Psychology, 86*(3), 169–193. http://dx.doi.org/10.1016/S0022-0965(03)00139-5.